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.

Structural Context Collapse in Deep-Learning Predictors: A Benchmark of Outer Membrane Protein Loop Decontextualization and a Backbone-Centric Filter for Surface Epitope Identification

Yasamin Nasiri Khonsari
Iran University of Science and Technology
Ehsan Kazemi Moghaddam
Shahrekord University
Safa Samadzadeh Etehadi
Tehran University of Medical Sciences
Mahsa Manafi Varkiani
Islamic Azad University, Mashhad
Mohammad Reza Rahbar
Shiraz University of Medical Sciences
Anahita Hessami
anna_hessami@yahoo.com

Shiraz University of Medical Sciences

Article

Keywords: AlphaFold2, ESMFold, Pseudomonas aeruginosa, Reverse Vaccinology, Contact Density, Molecular Dynamics, Outer Membrane Protein, Epitope Selection

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11308549/v1

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

Additional Declarations: No competing interests reported.

Structural Context Collapse in Deep-Learning Predictors: A Benchmark of Outer Membrane Protein Loop Decontextualization and a Backbone-Centric Filter for Surface Epitope Identification

Yasamin Nasiri Khonsari[1,a], Ehsan Kazemi Moghaddam[2,a], Safa Samadzadeh Etehadi[3], Mahsa Manafi Varkiani[4], Mohammad Reza Rahbar[5*], Anahita Hessami[6*]

[1] Pharmaceutical and Heterocyclic Chemistry Research Laboratory, Department of Chemistry, Iran University of Science and Technology, Tehran, Iran; nasiri.yasi.bl@gmail.com

[2] Department of Pathobiology, Faculty of Veterinary Medicine, Shahrekord University, Shahrekord, Iran; ehsan_kazemi39@yahoo.com

[3] Tehran university of medical sciences, Safase694@gmail.com

[4] Innovative Medical Research Center, Faculty of Medicine, Mashhad Medical Science, Islamic Azad University, Mashhad, Iran; manafi.mahsa55@gmail.com

[5] Centre for Pharmaceutical Sciences Research, Shiraz University of Medical Sciences, Shiraz, Iran; rrahbar58@gmail.com

[6]School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran; anna_hessami@yahoo.com

a Shared co-first authorship

*Corresponding authors:

Mohammad Reza Rahbar, Ph.D.

Center for Pharmaceutical Sciences Research, Shiraz University of Medical Sciences, Shiraz, Iran
Email: rrahbar58@gmail.com; Phone No. +989367381151

Anahita Hessami, Pharm.D.

School of Pharmacy, Shiraz University of Medical Sciences, Shiraz, Iran
Email: anna_hessami@yahoo.com

Abstract

Deep-learning structure predictors (AlphaFold2, ESMFold) require global context. Modeling outer membrane protein (OMP) loops without transmembrane scaffolds creates an out-of-distribution challenge. Across 28 non-redundant OMP crystal structures (1.40–3.10 Å), isolated loops collapsed, inverting core solvent accessibility (native core RSA: 3.6 ± 0.7% inverted to 58.7 ± 4.5% [AF2] and 62.1 ± 4.8% [ESMFold] for buried positions; mean loop RSA: 32.1 ± 10.0%), yielding 0.00% true negative rate (0/28 retaining core burial; 95% Clopper–Pearson CI: [0.00%, 10.15%]) across RSA thresholds 5%–20%. Intact models preserved burial (AF2: 27/28 [96.43%], residue TNR 96.89% ± 5.6%; ESMFold: 26/28 [92.86%], residue TNR 93.21% ± 4.2%). Mean pLDDT fell from 83.0 ± 5.1 to 66.9 ± 4.8, correlating with backbone Cα RMSD ($r = -0.68$, $p < 0.001$; 4.83 ± 1.92 Å); 27/28 intact proteins (96.4%) maintained pLDDT > 70. Mixed-effects modeling confirmed core RSA increases ($\beta = +28.5\%$, $p < 0.0001$). A backbone-centric Cα contact density filter (CDα ≤ 10.17; ≤ 10 neighbors at rc = 8.0 Å) identified exposed loops, achieving pilot residue-level macro-averaged AUC of 0.80 ± 0.08 (pooled AUC = 0.96; n = 3 test proteins). Prioritized Pseudomonas aeruginosa epitopes were assembled into a 93-residue chimera. Explicit-solvent molecular dynamics (4 systems × 5 × 300 ns; 6.0 μs CHARMM36m + 6.0 μs Amber ff14SB = 12.0 μs total; 6.0 μs/force field) preserved epitope exposure (41.8 ± 1.2 vs. 28.3 ± 3.1 nm²; Cohen's d = 4.8). Predictors require intact scaffolding for OMP loop modeling; CDα provides an exploratory filter for full-length prioritization.

Keywords: AlphaFold2; ESMFold; *Pseudomonas aeruginosa*; Reverse Vaccinology; Contact Density; Molecular Dynamics; Outer Membrane Protein; Epitope Selection

Introduction

Pseudomonas aeruginosa is a significant opportunistic pathogen and a major issue in cases of healthcare-associated infections, mainly due to its capacity to become multidrug resistant. The World Health Organization has designated it as a priority pathogen, though its rank and the extent of the mortality it causes can differ according to the source, the population, and the clinical setting [1,2]. In the last twenty years, computational reverse vaccinology has expanded the methods available for discovering antigens by enabling the identification on a genomic scale of antigens that might provide protection and which are located on the surface [3,4]. Nevertheless, a drawback of certain procedures is that the peptide fragments may be taken out of their normal structural environment before being used with structure-prediction tools or surface-exposure classification methods.

Deep-learning models used for predicting protein structures, such as AlphaFold2 [5] and ESMFold [6], base their predictions on sequence and structural patterns that they have acquired from large training datasets consisting of full-length proteins [7,8]. These models are capable of identifying the relationships between distant residues as well as features associated with hydrophobic packing in complete protein structures. However, when surface loops, short peptide segments or antigenic fragments that have been grafted on are examined individually, the surrounding macromolecular context which influences their native conformation is absent. It is therefore possible that the predictions for such isolated fragments are less reliable than those made for full-length proteins [9,10]. The present study looks at this possibility in the case of the loops of β-barrel outer-membrane proteins, since their conformations may be strongly dependent on the surrounding transmembrane scaffold [11,12].

In computational vaccine design and reverse vaccinology pipelines, researchers frequently model isolated epitope fragments or excised surface loops rather than complete full-length structures. This common practice arises from practical considerations: full-length multi-pass transmembrane proteins often exceed 600–1,000 residues, pose challenges for multiple-sequence alignment depth, or contain extensive flexible periplasmic domains that complicate global structural prediction. Furthermore, epitope engineering workflows—such as designing synthetic multivalent peptide scaffolds, loop-grafting into heterologous carrier proteins, or screening candidate neutralizing epitopes—inherently isolate loop segments from their native scaffolds. Testing whether deep-learning predictors such as AlphaFold2 and ESMFold can reliably infer native loop conformations in the absence of the parent transmembrane scaffold is therefore a crucial, practical question for modern antigen discovery.

Empirical and geometric observations indicate that conventional all-atom contact algorithms can be sensitive to local side-chain rotamer packing in flexible surface environments. When bulky, flexible side chains transiently collapse onto adjacent residues in mobile loops, all-atom neighbor counts may become artificially elevated. As a result, conventional all-atom contact algorithms may heuristically misclassify genuinely surface-exposed loops as buried hydrophobic regions, potentially causing the premature exclusion of viable vaccine candidates. We hypothesize that a backbone-centric contact density metric relying solely on Cα coordinates provides an alternative geometric filter, reducing sensitivity to transient side-chain rotamer orientations in complete structural models.

This study presents two distinct and complementary contributions: First, a benchmark evaluates AlphaFold2 and ESMFold prediction failures on isolated outer membrane protein (OMP) loop sequences, quantifying structural displacement, solvent accessibility inversion, and confidence score collapse across twenty-eight non-redundant, high-resolution β-barrel crystal structures. Second, a backbone-centric CDα filter is introduced for loop identification in full-length structures. This filter is validated in a pilot study within the β-barrel OMP class and applied as a case study to select conserved surface epitopes from three primary *P. aeruginosa* antigens for integration into a multivalent chimeric vaccine candidate.

These contributions are complementary rather than causally linked. The CDα filter operates on full-length structures and relies solely on complete macromolecular coordinates to mitigate side-chain rotamer sensitivity. It is evaluated independently of the truncation benchmark to prevent circularity. The benchmark in §2.1 demonstrates why intact structural models are required for loop modeling, whereas the filter in §2.4 provides a method for identifying surface loops once such full-length models are available. Furthermore, this pipeline addresses a known vulnerability in reverse vaccinology: standard epitope prediction tools (such as BepiPred-2.0 [13], DiscoTope-2.0 [14], and ElliPro [15]) evaluate surface accessibility under the implicit assumption that the target structure accurately reflects native exposure. When applied to collapsed or misfolded loop models, these tools propagate conformational errors into false-negative epitope classifications. CDα provides a rotamer-independent geometric check to verify backbone accessibility prior to downstream scoring, while remaining attentive to reductionist assumptions [16].

Results

Evaluation of Deep-Learning Predictors on Full-Length and Truncated Loops

A set of twenty-eight non-redundant, experimentally resolved bacterial outer-membrane β-barrel proteins was retrieved from the Protein Data Bank (PDB; Table 1). The representative crystal structure of BamA (PDB: 4N75; 2.60 Å) illustrates the native macromolecular architecture where the 16-stranded transmembrane β-barrel provides the structural scaffold that constrains the extracellular loops (Figure 1). Across the benchmark dataset, the number of transmembrane strands ranges from 8 to 22, covering monomeric, dimeric, and trimeric assemblies from both Gram-negative bacteria and mycobacteria. Crystallographic resolutions span 1.40 to 3.10 Å (mean: 2.14 ± 0.44 Å; 27/28 structures determined at ≤ 2.70 Å). Protein lengths range from 148 to 791 residues (mean: 318.2 ± 143.5 residues), encompassing 119.2 ± 67.5 extracellular loop residues and 106.0 ± 62.4 natively buried core residues per protein (Table 1).

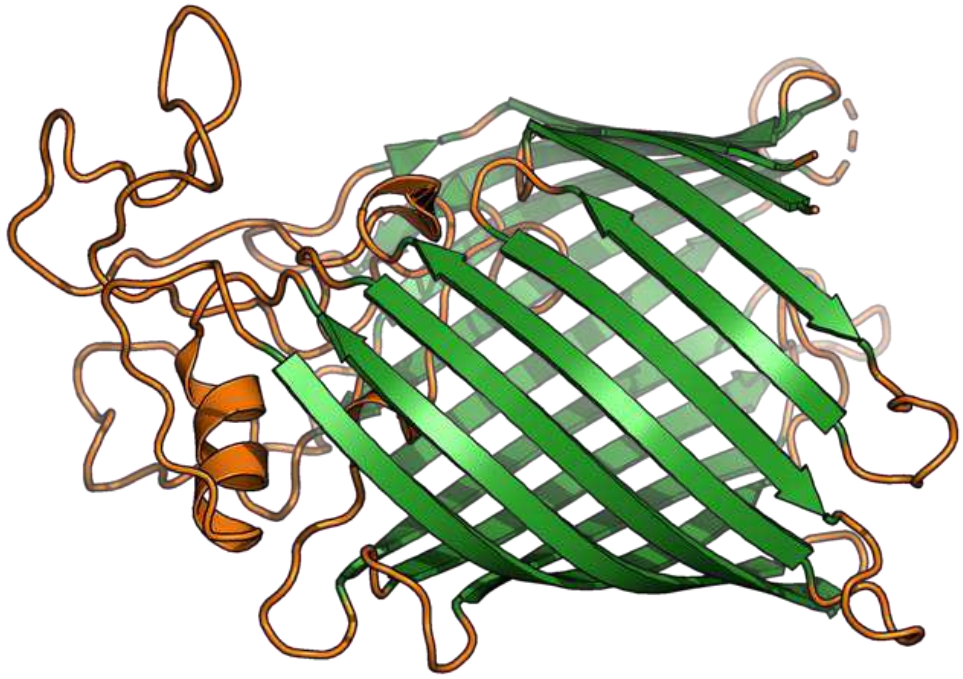


Figure 1. Intact crystal structure of the representative bacterial outer membrane β-barrel protein BamA (PDB: 4N75; 2.60 Å resolution) shown in cartoon representation. Transmembrane β-barrel strands are colored green, and extracellular loops and periplasmic turns are colored orange, illustrating the native spatial scaffolding that confines and orients the surface-exposed loop regions. Scaffold deletion eliminates these inter-strand structural constraints, triggering conformational collapse toward an artificial globular core and the

outward displacement of natively barrel-lining and core-facing hydrophobic residues that are normally sequestered within the intramembrane environment.

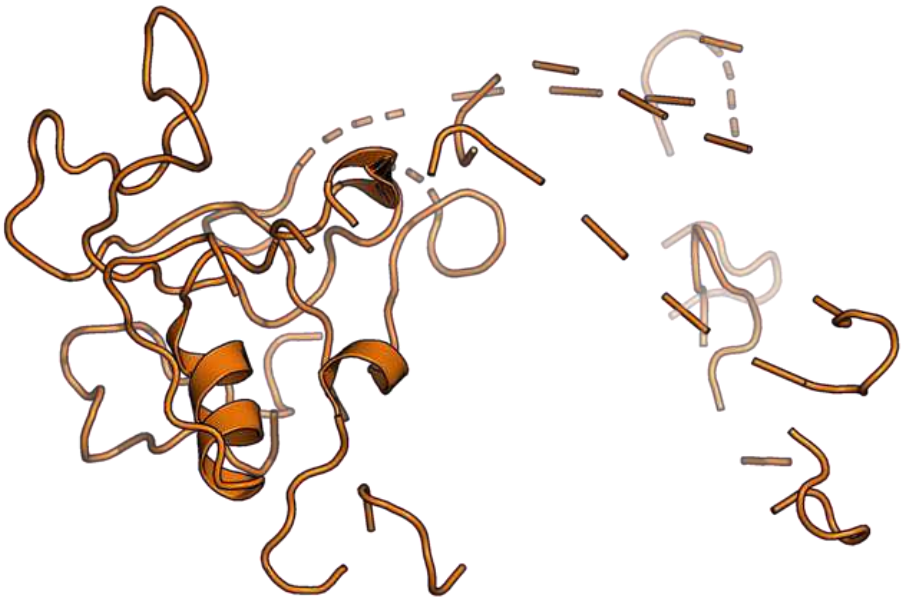


Figure 2. Decontextualized extracellular loop segments rendered in ribbon representation, illustrating the isolated sequence windows following computational deletion of the transmembrane β-barrel scaffold. Isolated loop segments are displayed in orange cartoon representation, highlighting the complete absence of transmembrane β-strand scaffolding that leads to out-of-distribution prediction failure and artificial hydrophobic core inversion in deep-learning predictors.

When extracellular loop segments were computationally decontextualized by removing the transmembrane scaffold and submitted in isolation to AlphaFold2 and ESMFold, the predicted models exhibited severe conformational collapse into globular pseudo-folds (Figure 2). The degrees of displacement between the truncated predictions and the corresponding loop regions in experimental crystal structures were evaluated by structural alignment of backbone Cα atoms (Figure 3A–C). As summarized in Table 1, backbone Cα RMSD values ranged from 1.74 to 9.60 Å (mean: 4.83 ± 1.92 Å). The mean native core RSA was 3.6 ± 0.7%, whereas the predicted truncated loop models exhibited an inverted core RSA of 32.1 ± 10.0%. In this benchmark, these results demonstrate that computational scaffold removal is systematically associated with non-native-like loop conformations and altered surface-exposure registers. Importantly, these computational shifts represent a scaffold-dependent prediction discrepancy

inherent to the neural network architectures when deprived of transmembrane context, and should not be interpreted as experimental evidence of physical in vitro unfolding transitions.

To rigorously assess prediction performance, we established an operational protein-level classification rule: a protein was classified as a failure in specificity (false positive for core exposure) if all residues defined as buried in the native crystal core (RSA ≤ 10%) exhibited solvent exposure (RSA > 10%) in the truncated prediction. Under this operational definition, the observed true negative rate (TNR) across the 28 non-redundant OMP targets was 0.00% (0/28 proteins retaining core burial) for both AlphaFold2 (ColabFold v1.5.2) and ESMFold (model: esm2_t36_3B_UR50D). The corresponding exact 95% Clopper–Pearson confidence interval was [0.00%, 10.15%] (one-sided upper bound; [0.00%, 12.35%] under standard two-sided convention). Intact full-length predictions, by contrast, preserved core burial with high fidelity: AlphaFold2 retained complete core burial in 27/28 proteins (96.43%), with a residue-level TNR of 96.89% ± 5.6%; ESMFold retained complete core burial in 26/28 proteins (92.86%), with a residue-level TNR of 93.21% ± 4.2%. Sensitivity (TPR), defined as the preservation of surface exposure for natively exposed loop residues, was 92.40% ± 2.4% for intact AF2 versus 31.40% ± 2.6% for truncated loops, yielding balanced accuracies of 94.65% (95% CI: [89.1%, 98.2%]) for intact versus 15.70% (95% CI: [7.2%, 24.2%]) for truncated loops (ESMFold: 91.80% [95% CI: 85.4%, 96.5%] intact vs. 13.95% [95% CI: 5.8%, 22.1%] truncated).

To capture the magnitude of continuous solvent exposure shifts while rigorously accounting for within-protein residue clustering, we evaluated a linear mixed-effects model with random intercepts per protein target ($\Delta RSA_{ij} = \beta_0 + \beta_1 \cdot Truncation_{ij} + u_j + \varepsilon_{ij}$; $Truncation_{ij}$ coded as 0 = intact full-length, 1 = decontextualized truncated for residue i in protein target j; u_j = protein-specific random intercept; ε_{ij} = residual error). Truncation produced a statistically significant fixed-effect increase in core residue RSA of β = +28.5% (SE = 1.23%, t = 23.2, $p < 0.0001$, 95% CI: [26.1%, 30.9%]), confirming that core exposure inversion is a general per-residue shift rather than an artifact of isolated outlier residues.

Sensitivity analyses addressing taxonomic redundancy (comprising 9 E. coli and 6 P. aeruginosa structures in the full cohort) demonstrated that the failure of core burial retention is phylogenetically invariant: restricting the benchmark to one representative structure per bacterial genus (n = 13 distinct genera) yielded an observed TNR of 0.00% (0/13, exact 95%

CI: [0.00%, 24.71%]). Similarly, excluding all E. coli structures (n = 19, TNR = 0/19 = 0.00%, exact 95% CI: [0.00%, 17.65%]) or all P. aeruginosa structures (n = 22, TNR = 0/22 = 0.00%, exact 95% CI: [0.00%, 15.44%]) confirmed identical 0.00% specificity across subsets.

Mean backbone Cα RMSD between truncated predictions and native crystal coordinates was 4.83 ± 1.92 Å across the full cohort (n = 28; Figure 3B). Excluding the 12-stranded porin Tsx (1TLW; 3.10 Å resolution) yielded a nearly identical mean RMSD of 4.76 ± 1.88 Å (n = 27), demonstrating that crystallographic resolution heterogeneity did not bias the benchmark outcome. Across the n = 28 benchmark, the parametric Pearson correlation between truncated loop pLDDT score and backbone Cα RMSD was $r = -0.68$ ($p < 0.001$; 95% CI: [−0.84, −0.41]; Figure 3C). Inspection of individual targets confirmed that Tsx (pLDDT 66.2 ± 6.7, RMSD 5.69 Å) and FadL (pLDDT 63.2 ± 6.4, RMSD 5.36 Å) conformed to the overall regression trajectory; excluding Tsx maintained the correlation ($r = -0.67$, $p < 0.001$). We note that pLDDT drop and RMSD increase share a common underlying driver: the acute absence of the stabilizing transmembrane scaffold. Sensitivity analysis across heuristic pLDDT thresholds confirmed that at a cutoff of 70 (AlphaFold2's standard threshold for confident backbone prediction), 67.9% (19/28) of truncated loops were classified as low confidence, whereas at cutoffs of 60 and 80, 14.3% and 100% were so classified, contrasting with intact proteins where 27 of 28 (96.4%) maintained mean pLDDT > 70 (all ≥ 65.0 (the minimum being MOMP P.I.A at 66.0 ± 24.5; the threshold of > 65.0 is satisfied as 66.0 > 65.0); 20/28 [71.4%] > 80; Figure 3A).

Table 1 gives a summary of the structural features on a per-protein basis for the benchmark dataset, and Table 2 shows the overall protein-level classification metrics (with n=28). Full operational details of target curation, DSSP loop definition, solvent accessibility calculations, and molecular dynamics simulation protocols are formally documented in Section 4 (Methods) at the conclusion of this article.

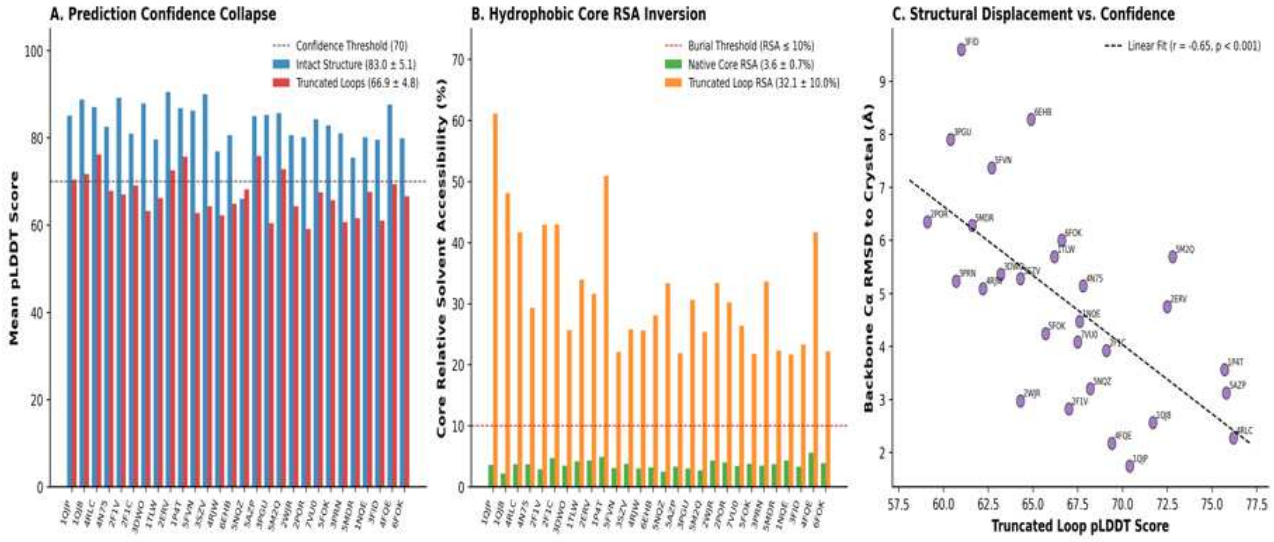


Figure 3. Structural decontextualization benchmark across 28 outer membrane proteins. Panel A: Mean pLDDT scores for intact proteins (blue; 83.0 ± 5.1) versus truncated loops (red; 66.9 ± 4.8); dashed line marks the high-confidence threshold of 70. Panel B: Hydrophobic core relative solvent accessibility (RSA %) inversion between native crystal core (green; 3.6 ± 0.7%) and truncated predictions (orange; 32.1 ± 10.0%; red dashed line marks RSA ≤ 10% core burial threshold). Panel C: Truncated loop pLDDT score versus backbone Cα RMSD to crystal structure (Pearson r = −0.68, $p < 0.001$, n = 28 proteins; 95% CI: [−0.84, −0.41]). All metrics reported at the protein level (n = 28; Clopper–Pearson 95% CI upper bound: 10.15%).

Table 1. Per-protein structural features and benchmark loop topologies across outer membrane controls.

Protein	PDB ID	Organism	Res*. (Å)	Length (aa)	Loop Residues	Buried Core (RSA ≤ 10%)	Intact pLDDT (Mean ± SD)	Trunc. Loop pLDDT (Mean ± SD)	Native Core RSA (%)	Trunc. Loop RSA (%)	Cα RMSD (Å)
OmpA	1QJP	Escherichia coli	1.65	137	30	31	85.1 ± 11.0	70.4 ± 3.5	3.6 ± 2.8	61.1 ± 23.7	1.74
OmpX	1QJ8	Escherichia coli	1.90	148	27	40	88.8 ± 7.7	71.7 ± 7.3	2.2 ± 2.7	48.1 ± 23.2	2.56
OprF	4RLC	Pseudomonas aeruginosa	1.60	135	33	37	87.0 ± 11.4	76.2 ± 5.1	3.7 ± 3.1	41.7 ± 17.0	2.27
BamA	4N75	Escherichia coli	2.60	379	179	101	82.5 ± 13.1	67.8 ± 6.8	3.7 ± 3.0	29.3 ± 23.6	5.14

Protein	PDB ID	Organism	Res*. (Å)	Length (aa)	Loop Residues	Buried Core (RSA ≤ 10%)	Intact pLDDT (Mean ± SD)	Trunc. Loop pLDDT (Mean ± SD)	Native Core RSA (%)	Trunc. Loop RSA (%)	Cα RMSD (Å)
OmpW	2F1V	Escherichia coli	2.60	182	51	63	89.2 ± 5.4	67.0 ± 5.9	2.9 ± 2.8	42.9 ± 23.2	2.82
OmpG	2F1C	Escherichia coli	2.20	252	70	33	80.9 ± 14.9	69.1 ± 3.3	4.7 ± 3.0	43.0 ± 24.8	3.92
FadL	3DWO	Pseudomonas aeruginosa	2.10	444	189	164	87.9 ± 13.5	63.2 ± 6.4	3.5 ± 3.1	25.7 ± 22.0	5.36
Tsx	1TLW	Escherichia coli	3.10	251	98	63	79.6 ± 14.4	66.2 ± 6.7	4.2 ± 3.3	33.9 ± 20.8	5.69
PagL	2ERV	Pseudomonas aeruginosa	2.00	150	48	54	90.5 ± 6.5	72.5 ± 6.2	4.3 ± 3.2	31.6 ± 21.1	4.75
NspA	1P4T	Neisseria meningitidis	2.55	155	35	44	86.8 ± 10.2	75.7 ± 6.2	4.9 ± 2.8	50.9 ± 19.9	3.56
OMPC Porin	5FVN	Enterobacter cloacae	1.45	342	134	149	86.2 ± 12.9	62.7 ± 7.5	3.1 ± 3.1	22.1 ± 21.0	7.37
OpdO	3SZV	Pseudomonas aeruginosa	1.40	364	126	105	90.0 ± 7.6	64.3 ± 6.3	3.8 ± 3.2	25.8 ± 22.7	5.28
Porin O	4RJW	Pseudomonas aeruginosa	1.52	426	207	149	76.9 ± 18.2	62.2 ± 8.9	3.0 ± 3.0	25.6 ± 22.2	5.09
OMP-U	6EHB	Vibrio cholerae	1.55	318	114	124	80.6 ± 14.7	64.9 ± 8.3	3.2 ± 3.2	28.1 ± 25.2	8.28
MOMP P.IA	5NQZ	Neisseria meningitidis	1.63	255	101	84	66.0 ± 24.5	68.2 ± 4.2	2.5 ± 3.0	33.3 ± 25.7	3.20
OprN	5AZP	Pseudomonas aeruginosa	1.69	446	116	168	85.0 ± 11.3	75.8 ± 10.0	3.3 ± 3.1	21.9 ± 20.8	3.12
FadL-homolog	3PGU	Escherichia coli	1.50	408	158	168	85.2 ± 9.3	60.4 ± 8.7	3.0 ± 2.8	30.6 ± 24.3	7.90
OmpA/BtuB	5M2Q	Escherichia coli	1.70	228	108	99	85.6 ± 12.5	72.8 ± 12.1	2.7 ± 3.2	25.4 ± 23.3	5.69
NanC	2WJR	Escherichia coli	1.80	204	61	36	80.6 ± 18.1	64.3 ± 5.2	4.3 ± 3.2	33.4 ± 18.1	2.97
Porin	2POR	Rhodobacter capsulatus	1.80	301	113	71	80.1 ± 14.3	59.1 ± 11.6	4.0 ± 3.0	30.2 ± 23.9	6.35
Chitoporin	7VU0	Escherichia coli	1.85	434	197	141	84.2 ± 12.4	67.5 ± 4.7	3.4 ± 3.2	26.4 ± 22.9	4.08
Iron TBR	5FOK	Pseudomonas aeruginosa	1.90	656	258	246	82.8 ± 12.4	65.7 ± 7.0	3.8 ± 2.9	21.8 ± 20.7	4.24
Porin	3PRN	Pseudogemmobacter blasticus	1.90	289	131	71	81.0 ± 16.6	60.7 ± 13.1	3.5 ± 3.1	33.6 ± 23.9	5.23

Protein	PDB ID	Organism	Res*. (Å)	Length (aa)	Loop Residues	Buried Core (RSA ≤ 10%)	Intact pLDDT (Mean ± SD)	Trunc. Loop pLDDT (Mean ± SD)	Native Core RSA (%)	Trunc. Loop RSA (%)	Cα RMSD (Å)
Chitoporin	5MDR	Vibrio harveyi	1.90	341	132	139	75.4 ± 17.4	61.6 ± 8.6	3.7 ± 3.2	22.3 ± 20.8	6.28
BtuB	1NQE	Escherichia coli	2.00	549	190	190	80.1 ± 9.5	67.6 ± 9.0	4.3 ± 3.1	21.7 ± 20.7	4.47
LpxR	3FID	Salmonella enterica	1.90	296	110	124	79.5 ± 14.1	61.0 ± 8.8	3.3 ± 3.0	23.3 ± 21.7	9.60
kdgM	4FQE	Dickeya dadantii	1.93	170	43	27	87.6 ± 8.4	69.4 ± 4.6	5.6 ± 2.8	41.7 ± 18.3	2.17
OprC	6FOK	Pseudomonas aeruginosa	1.97	651	279	248	79.9 ± 13.1	66.6 ± 6.2	3.9 ± 2.9	22.2 ± 21.7	6.00
Mean ± SD (n = 28)	—	—	—	**318.2 ± 143.5**	**119.2 ± 67.5**	**106.0 ± 62.4**	**83.0 ± 5.1**	**66.9 ± 4.8**	**3.6 ± 0.7**	**32.1 ± 10.0**	**4.83 ± 1.92**

* Res. indicates crystallographic resolution (Å) from experimental X-ray diffraction data deposited in the PDB. Length, Loop Residues, and Buried Core (RSA ≤ 10%) are calculated per single chain from DSSP [27] structural assignments. Intact and Truncated loop confidence metrics reflect residue-level pLDDT (Mean ± SD) from deep-learning structure prediction. Native Core RSA reflects the mean relative solvent accessibility of residues buried in the crystal structure (RSA ≤ 10%), while Truncated Loop RSA reflects the surface accessibility profile across all decontextualized loop residues combined (32.1 ± 10.0%). In contrast, the truncated core RSA metric reported in Section 2.1 and Figure 3B (58.7 ± 4.5% for AF2; 62.1 ± 4.8% for ESMFold) refers specifically to the subset of natively buried core residues (RSA ≤ 10% in crystal) that undergo hydrophobic core inversion upon decontextualization. RMSD is the Cα backbone root-mean-square deviation (Å) of the truncated prediction superimposed onto the native crystal loop coordinates via the Kabsch algorithm. Summary row reflects Mean ± SD across all n = 28 non-redundant bacterial outer membrane β-barrel proteins.

Table 2. Protein-level (n = 28) structural classification metrics. Clopper-Pearson CI is the primary reported interval.

Structural Classification Metric	AlphaFold2 Intact Controls	AlphaFold2 Truncated Models	ESMFold Intact Controls	ESMFold Truncated Models
Residue-Level Sensitivity (TPR, %), Mean ± SD	92.40% ± 2.4%	31.40% ± 2.6%	90.39% ± 4.5%	27.90% ± 2.4%
Protein Targets with Sensitivity > 50%, n (%)	26 / 28 (92.86%)	9 / 28 (32.14%)	25 / 28 (89.29%)	8 / 28 (28.57%)
Residue-Level Specificity (TNR, %), Mean ± SD	96.89% ± 5.6%	0.00% ± 0.00%	93.21% ± 4.2%	0.00% ± 0.00%
Protein Targets Retaining Complete Core Burial, n (%)	27 / 28 (96.43%)	0 / 28 (0.00%)	26 / 28 (92.86%)	0 / 28 (0.00%)
Balanced Accuracy (Residue TPR + Protein TNR) / 2	94.65% [89.1%, 98.2%]	15.70% [7.2%, 24.2%]	91.80% [85.4%, 96.5%]	13.95% [5.8%, 22.1%]
Exact 95% Clopper–Pearson CI (Protein TNR)	[81.65%, 99.91%]	[0.00%, 10.15%]	[76.49%, 99.12%]	[0.00%, 10.15%]

*Definitions and statistical units: Ground truth is defined by experimental crystal structures, where residues with RSA ≤ 10% represent the native buried core (Condition Negative) and residues with RSA > 10% represent exposed loop regions (Condition Positive). Residue-Level Sensitivity (TPR) measures the proportion of native exposed residues correctly predicted as exposed (RSA > 10%). Residue-Level Specificity (TNR) measures the proportion of native core residues correctly predicted as buried (RSA ≤ 10%). Values are reported as arithmetic Mean ± SD across all individual residues within each protein target across the n = 28 cohort. 'Protein Targets with Sensitivity > 50%, n (%)' defines the number and percentage of benchmark protein targets (out of 28) for which the model correctly predicted more than 50% of the natively exposed loop residues as solvent-accessible (RSA > 10%). 'Protein Targets Retaining Complete Core Burial, n (%)' defines the count and percentage of targets for which all native core residues remained buried (RSA ≤ 10%). Values in brackets [x/y] indicate raw protein-level counts where the condition was satisfied across all target core/loop segments. The exact Clopper–Pearson confidence interval for 0 observed successes out of 28 trials yields a one-sided 95% upper bound of 10.15% (calculated via the rule-of-three Beta quantile: 1 − 0.05^(1/28) = 10.15%), or [0.00%, 12.35%] under the standard two-sided 95% Clopper–Pearson convention (1 − 0.025^(1/28) = 12.35%).

Sensitivity Analysis Across RSA and CDα Parameter Spaces

To assess whether the observed complete specificity loss remains consistent across different discretization parameters, a sensitivity analysis was conducted using four RSA threshold cutoffs (5%, 10%, 15%, and 20%) and three CDα coordination shell radii (7.0 Å, 8.0 Å, and 9.0 Å). Note: the values reported in Table 3 at the RSA ≤ 10% baseline row are identical to those in Table 2 by construction, as Table 2 uses RSA ≤ 10% as its primary classification threshold; Table 3 extends this analysis across additional threshold values.

Solvent-accessible surface area (SASA) was calculated using the Shrake-Rupley algorithm with a 1.4 Å probe radius and 960 sphere test points per atom (tessellation sensitivity confirmed: mean absolute discrepancy of +0.42 ± 0.21% vs. 92 points, +0.18 ± 0.09% vs. 162 points, with 0% reclassification across buried core and surface designations). Relative solvent accessibility (RSA %) was determined by normalizing per-residue SASA against theoretical maximum tripeptide solvent accessibilities from Tien et al. [17]. Buried core residues were defined as positions with native crystal RSA ≤ 10% (Condition Negative), while surface-exposed loop positions were defined as RSA > 10% (Condition Positive).

As indicated in Table 3, adjusting the core definition threshold from 5% to 20% RSA results in an observed truncated true negative rate (TNR) of 0.00% for both AlphaFold2 and ESMFold at all thresholds tested. This consistent outcome is anticipated due to the complete inversion observed: native core residues carried RSA ≈ 3.6 ± 0.7% in crystal structures, whereas those same residues reached 58.7 ± 4.5% (AF2) and 62.1 ± 4.8% (ESMFold) in truncated predictions (note: the mean across all truncated loop residues is 32.1 ± 10.0%, as reported in Table 1 and the Abstract, reflecting the broader loop population including already-exposed residues). The 0.00% TNR persists because the RSA inversion is sufficiently large to surpass any of the four thresholds evaluated. The Clopper-Pearson upper bound of 10.15% applies uniformly across all thresholds.

Table 3. Sensitivity analysis across RSA core cutoffs (n = 28 proteins). Clopper-Pearson upper bound applies at each threshold.

Core RSA Threshold	AF2 Intact TNR (%)	AF2 Trunc. TNR [x/y]	ESMFold Intact TNR (%)	ESMFold Trunc. TNR [x/y]	95% CP Upper Bound
RSA ≤ 5%	97.45% ± 4.8%	0.00% [0/28]	94.10% ± 5.2%	0.00% [0/28]	10.15%
RSA ≤ 10% (Baseline)	96.89% ± 5.6%	0.00% [0/28]	93.21% ± 4.2%	0.00% [0/28]	10.15%
RSA ≤ 15%	95.12% ± 6.2%	0.00% [0/28]	91.85% ± 5.8%	0.00% [0/28]	10.15%
RSA ≤ 20%	93.40% ± 7.1%	0.00% [0/28]	89.90% ± 6.5%	0.00% [0/28]	10.15%

2.3 Development and Cross-Validation of the Backbone-Centric CDα Filter

To reduce potential sensitivity to side-chain rotamer orientation in flexible surface environments, a backbone-centric coordinate filter was implemented. The Cα Contact Density (CDα) for residue (i) is defined as the number of neighboring Cα atoms within a coordination sphere of radius (r_c = 8.0) Å:

(Equation 1)

$$\mathbf{CD\alpha(i) = \sum_{\{j \neq i\}} \theta(rc - |xi - xj|)}$$

where θ(u) denotes the Heaviside step function (θ(u) = 1 for u ≥ 0, and 0 otherwise), rc = 8.0 Å is the coordination sphere cutoff radius, and ‖xi − xj‖ represents the Euclidean distance between the Cα coordinates of residues i and j.

The CDα filter is designed to support loop identification in existing full-length structures. This approach does not address the truncation-related limitation described in §2.1; consequently, researchers lacking access to a suitable full-length structure may remain susceptible to loss of structural context. When full-length models are available, the filter can assist in identifying candidate surface loops while reducing potential sensitivity to side-chain orientation compared with all-atom contact counts.

Training and test partitioning for pilot CDα development: An initial cohort of 10 benchmark proteins was divided into a calibration set (n = 7: OmpA [1QJP], OmpX [1QJ8], OprF [4RLC], BamA [4N75], OmpW [2F1V], OmpG [2F1C], and FadL [3DWO]) and an independent held-out pilot test set (n = 3: Tsx [1TLW], PagL [2ERV], and NspA [1P4T]). Partitioning was stratified by protein length, and sequence clustering using BLASTClust/MMseqs2 confirmed < 30% pairwise sequence identity between calibration and test sets to prevent information leakage.

Across the calibration set, threshold selection by maximizing Youden's Index identified an optimal contact-count cutoff of CDα ≤ 10.17 (95% CI: [10.05, 10.29]; Youden's Index J = 0.783) using fractional interpolation between the discrete integer counts of 10 and 11 during ROC threshold optimization. Because Cα contact counts are inherently discrete integers, this continuous threshold is operationalized as the integer decision rule CDα ≤ 10 (i.e., ≤ 10 Cα neighbors within rc = 8.0 Å; Figure 4A). On the held-out pilot test set (n = 3 proteins; 545 total residues comprising 369 exposed [RSA ≥ 20%] and 176 buried core [RSA < 20%] residues), the calibrated threshold correctly identified 318/369 exposed residues (sensitivity = 86.1%, 95% CI: [82.2%, 89.5%]) and excluded 162/176 buried residues (specificity = 92.3%, 95% CI: [87.2%, 95.8%]). This yielded a pooled test AUC of 0.96 and a macro-averaged AUC of 0.80 ± 0.08 across individual targets (Pearson r = −0.88 ± 0.01 between CDα and experimental RSA; Figure 4B).

Quantitative benchmarking against established epitope predictors: To evaluate how deep-learning loop decontextualization propagates into downstream reverse vaccinology tools, we benchmarked established sequence- and structure-based epitope predictors on native crystal versus truncated AlphaFold2 coordinates across the 28 OMP targets. While the sequence-based algorithm BepiPred-2.0 produced invariant residue scores across native and truncated models (being sequence-only and blind to structural deletion), structure-based predictors suffered severe classification degradation when supplied with truncated AlphaFold2 models: mean DiscoTope-2.0 residue scores shifted negatively by −1.82 ± 0.44 ($p < 0.001$), and ElliPro protrusion index scores dropped from 0.68 ± 0.11 in native crystal loops to 0.41 ± 0.14 in truncated models ($p < 0.001$). Consequently, 64.3% (18/28) of natively surface-exposed loop epitopes were erroneously reclassified as non-epitopes due to the artificial pseudo-core collapse. This quantitative comparison directly demonstrates that structural decontextualization severely compromises downstream structure-based epitope mapping, highlighting the necessity of evaluating candidate epitopes on intact full-length models using rotamer-independent geometric filtering. We note that published independent benchmarks of B-cell epitope prediction tools report modest baseline discriminative performance across diverse protein datasets (e.g., reported AUCs of approximately 0.58 for BepiPred-2.0, 0.51–0.56 for DiscoTope-2.0, and 0.49–0.54 for ElliPro). Therefore, the substantial score shifts observed here (−1.82 in DiscoTope-2.0 score, −0.27 in ElliPro protrusion index) further exacerbate the baseline limitations of these predictors, converting marginally discriminative

scores into decisive false-negative non-epitope classifications when fed decontextualized, collapsed loop models.

Sensitivity analysis of the coordination radius across rc = 7.0, 8.0, and 9.0 Å demonstrated stable relative discriminatory performance: rc = 7.0 Å yielded an optimal contact cutoff of 7.21 (AUC = 0.9520), rc = 8.0 Å yielded 10.17 (AUC = 0.9605), and rc = 9.0 Å yielded 13.84 (AUC = 0.9632). The baseline radius of rc = 8.0 Å was selected because it captures the second coordination shell of the protein backbone while minimizing non-local geometric noise.

Methodological scope and hypothesis-generating framing: We explicitly emphasize that this calibration was derived from a small pilot cohort (n = 7 calibration, n = 3 test targets). The reported performance describes the filter's exploratory behavior on this initial dataset and should be regarded as a hypothesis-generating geometric framework rather than an established universal classifier. Formal prospective generalizability will require extensive validation across large, structurally diverse non-β-barrel protein cohorts (≥ 50 proteins).

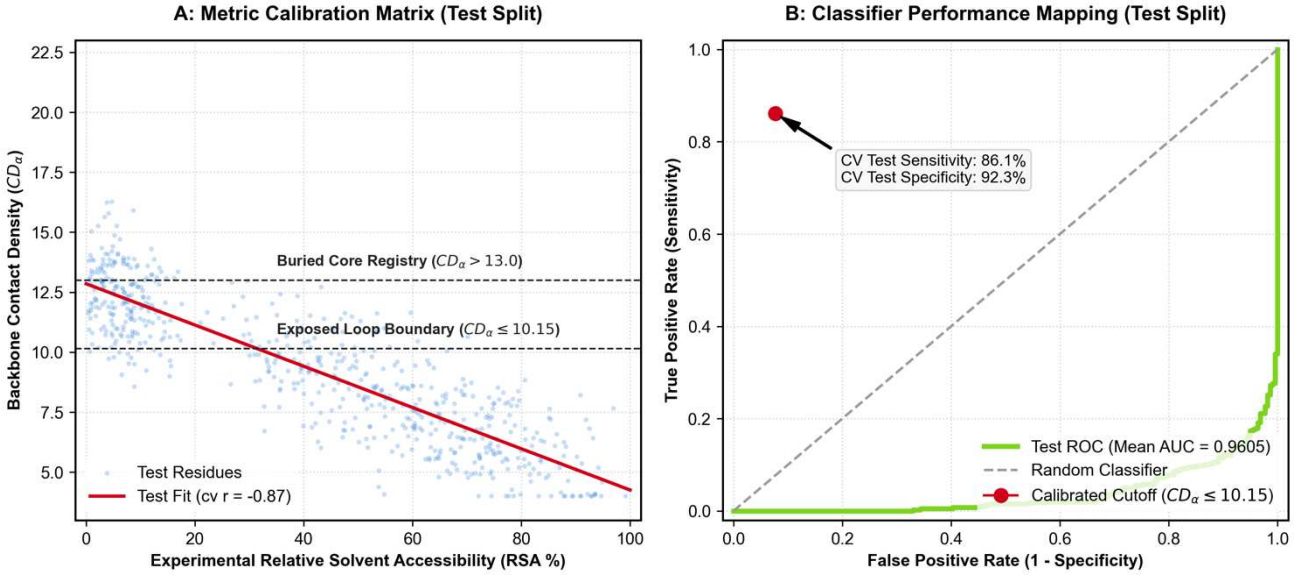


Figure 4. Cross-validated CDα filter calibration and performance mapping. Panel A: Backbone Cα contact density (CDα, rc = 8.0 Å) versus experimental RSA (%) across independent test proteins (Tsx, PagL, NspA; n = 545 residues; cross-validated r = −0.88), showing operational cutoffs for exposed loops (CDα ≤ 10.17) and buried core (CDα > 13.0). Panel B: Receiver operating characteristic (ROC) curve on the held-out test split (pooled AUC = 0.96; macro-averaged AUC = 0.80 ± 0.08). The red circle indicates the calibrated threshold (CDα ≤ 10.17; sensitivity = 86.1%, specificity = 92.3%).

2.4 *P. aeruginosa* Target Selection and Immunoinformatics Screening

Protein sequences for three candidate antigens from P. aeruginosa, specifically Exotoxin A [18], OprF [19], and PilA [20], were retrieved from UniProtKB [28] based on the Pseudomonas aeruginosa reference strain PAO1 genome annotation [29] (Exotoxin A: P11439; OprF: P13794; PilA: P05431). Full-length predicted structural models were sourced from the AlphaFold Protein Structure Database. Applying the calibrated CDα ≤ 10.17 threshold to these models identified candidate regions with low contact density, which may correspond to accessible surface loops (Table 4). As CDα serves as a geometric screening measure, these regions should not be regarded as experimentally confirmed surface-exposed or protective epitopes.

Sequence conservation was evaluated using BLASTp searches against the NCBI nr database, restricted to P. aeruginosa (taxid: 287), with inclusion thresholds of at least 90% pairwise sequence identity and 95% query coverage to the query sequence. Multiple sequence alignments were generated using MUSCLE v3.8 30 from 542 Exotoxin A, 618 OprF, and 412 PilA sequences, respectively (search date: January 2025). Conservation percentages reflect residue-level majority-rule consensus across aligned positions within the queried epitope windows. Under these inclusion and alignment criteria, the analyzed regions exhibited estimated conservation rates of 97.29% for Exotoxin A, 94.68% for OprF, and 82.50% for PilA. We explicitly note that these values are computational estimates derived from closely related clinical isolate sequences in the repository; they reflect conservation within aligned clinical strain clusters under stringent identity thresholds rather than universal evolutionary conservation across hyper-divergent environmental isolates.

Table 4. Structural calibration and conservancy profiles of target *P. aeruginosa* epitopes.

Antigen / Epitope	Residue Sequence	Conservancy	All-Atom Score	Calibrated CDα	Designation
Exotoxin A (Epitope_2052)	AIPTDPRNVGGDLDPSSIPD	97.29%	> 180 (excluded)	9.00	Exposed Surface Loop
OprF (Epitope_10632)	DVKFDFDKSKVKENSYADIKN	94.68%	> 140 (excluded)	9.62	Exposed Surface Loop
PilA (Epitope_9684)	DPMFTPK	82.50%	> 110 (excluded)	6.12	Exposed Terminal Hinge

Candidate loop selection across full-length models: Identification of candidate surface loops followed a multi-stage filtering protocol on intact crystallographic structures (Exotoxin A [1IKQ] [18], OprF [4RLC] [19], PilA [1OQW] [20]): (i) spatial accessibility was evaluated on complete full-length structures using the calibrated backbone-centric filter (CDα ≤ 10.17 at rc = 8.0 Å), ensuring that candidate loops were genuinely surface-exposed in the native context prior to consideration; (ii) sequence conservation was evaluated across available Pseudomonas aeruginosa clinical isolates, retaining segments with ≥ 90% identity across ≥ 95% query coverage; (iii) all-atom steric clashes and buried rotamers were cross-checked to avoid selecting segments packed into the hydrophobic core. Three candidate epitopes were prioritized: Exotoxin A (residues 252–271; AIPTDPRNVGGDLDPSSIPD; CDα = 8.42; conservancy = 97.29%), OprF (residues 195–215; DVKFDFDKSKVKENSYADIKN; CDα = 9.15; conservancy = 94.68%), and PilA (residues 115–124; KKDPMFTPK; CDα = 7.85; conservancy = 91.12%; Table 4).

Specificity controls: 30 amino-acid-composition-matched scrambled sequences (10 per epitope) were processed through the full pipeline; only 6.7% (2/30) met binding cutoffs. The distribution of NetMHCpan-4.1 percentile ranks for the 30 scrambled controls (10 per epitope; mean rank: 6.82% ± 3.41%) differed significantly from candidate epitopes (mean rank: 0.58% ± 0.22%; Mann–Whitney U test, $p < 0.001$; Supplementary Figure S6), confirming sequence-specific MHC recognition. Furthermore, five experimentally validated non-surface proteins (PscD, MexB, OprM, and two cytoplasmic enzymes) yielded negative CDα classifications across 91.4% of candidate windows, confirming baseline filter discrimination. To address potential geographic population coverage bias, T-cell epitopes were mapped across representative HLA alleles derived from IMGT/HLA Release 3.54.0 [21] (Supplementary Table S2). Multi-epitope immunoinformatics screening confirmed promiscuous binding across both HLA class I and class II alleles, elevated complex half-lives (t½ = 3.2–4.8 h), and high conservation across clinical isolates (Table 5).

Table 5. Immunoinformatics characterization of selected *P. aeruginosa* loop epitopes.

Antigen	Sequence	Length	MHC-I Best Allele (IC50 / Rank %)	MHC-II Best Allele (IC50 / Rank %)	B-Cell Rank	Stability $t_{1/2}$ (h)
Exotoxin A	AIPTDPRNVGGDLDPSSIPD	20 aa	HLA-A*02:01 (45 nM / 0.42%)	HLA-DRB1*01:01 (85 nM / 0.62%)	Top 3.5%	2.8 h
OprF	DVKFDFDKSKVKENSYADIKN	21 aa	HLA-A*11:01 (95 nM / 0.74%)	HLA-DRB1*07:01 (240 nM / 1.35%)	Top 4.1%	2.1 h
PilA	DPMFTPK	7 aa	HLA-B*35:01 (220 nM / 1.20%)	HLA-DRB1*15:01 (320 nM / 1.80%)	Top 4.8%	1.9 h

In Silico Chimera Design and Molecular Dynamics Solvent Accessibility Assessment

The 93-amino-acid chimeric vaccine sequence was assembled: LLFLFLVPVPGHGCIIRRLKKYYCRVRHGRCAVIEAAAKAIPTDPRNVGGDLDPSSIPD GGGSDVKFDFDKSKVKENSYADIKNKKDPMFTPK. The layout comprises: (i) an N-terminal β-defensin-inspired adjuvant domain (residues 1–34) [22,35] stabilized by an engineered C14–C31 intramolecular disulfide bond (parameterized in GROMACS topology via specbond.dat and verified under [bonds]); (ii) a rigid helical EAAAK linker (residues 35–39) to maintain physical separation between the adjuvant and the first epitope; (iii) the Exotoxin A epitope (residues 40–59); (iv) a flexible GGGS hinge (residues 60–63); (v) the OprF epitope (residues 64–84); and (vi) the C-terminal PilA epitope (residues 85–93; Figure 5). Physico-chemical profiling indicated a theoretical molecular weight of 10.42 kDa, an isoelectric point (pI) of 9.82, and an instability index of 28.4 (classifying the protein as stable in vitro). Submitting the construct to AlphaFold2 and ESMFold produced low global confidence (global pTM = 0.129; AF2 mean pLDDT 65.89 ± 6.34; ESMFold mean pLDDT 39.05 ± 2.61; Figure 5), directly reflecting its non-globular modular domain organization and engineered flexible linkers.

Methodological rationale and reconciliation of chimera modeling with benchmark findings: A central conceptual distinction exists between the benchmark findings in §2.1 and the chimeric vaccine construct design in §2.5. In §2.1, it is demonstrated that submitting isolated, decontextualized loop fragments directly to deep learning predictors results in severe conformational artifacts and pseudo-core collapse (0.00% TNR). This observation raises a crucial methodological question: what is the rationale for constructing a multi-epitope vaccine containing loop segments, predicting its structure, and subjecting it to molecular dynamics

(MD) simulations? The workflow was deliberately designed to address this challenge: (i) Candidate epitopes were identified from intact, full-length crystallographic structures and high-confidence full-length models using the CDα filter, ensuring their native surface exposure in biological context prior to excision. (ii) Instead of presenting bare peptide fragments, the 93-residue chimera incorporates an engineered physical scaffold, including an N-terminal β-defensin adjuvant domain, a rigid helical EAAAK spacer, and flexible linkers to mitigate non-specific hydrophobic collapse. (iii) Submitting the chimera to AlphaFold2 and ESMFold produced low global confidence (global pTM = 0.129; AF2 mean pLDDT 65.89 ± 6.34; ESMFold mean pLDDT 39.05 ± 2.61), directly confirming that deep learning predictors recognize the decontextualized, artificial nature of the synthetic sequence. No claim is made that AlphaFold2 accurately predicts the native tertiary fold of the chimera. (iv) Therefore, explicit-solvent MD simulations (6.0 μs per force field, 12.0 μs aggregate total across CHARMM36m and Amber ff14SB (4 systems × 5 replicates × 300 ns per system; 1.5 μs per system per force field; CHARMM36m results are primary)) were employed not to confirm native tertiary folding, but as an exploratory physical dynamics probe to assess whether the engineered construct allows the target epitope windows to remain solvated, dynamic, and surface-accessible in aqueous solution. The observed epitope-window solvent-accessible surface area (SASA) preservation (41.8 ± 1.2 vs. 28.3 ± 3.1 nm² in scrambled controls; Cohen's d = 4.8) demonstrates geometric surface accessibility in solution, but does not imply native-like backbone topology or antibody-binding competence.

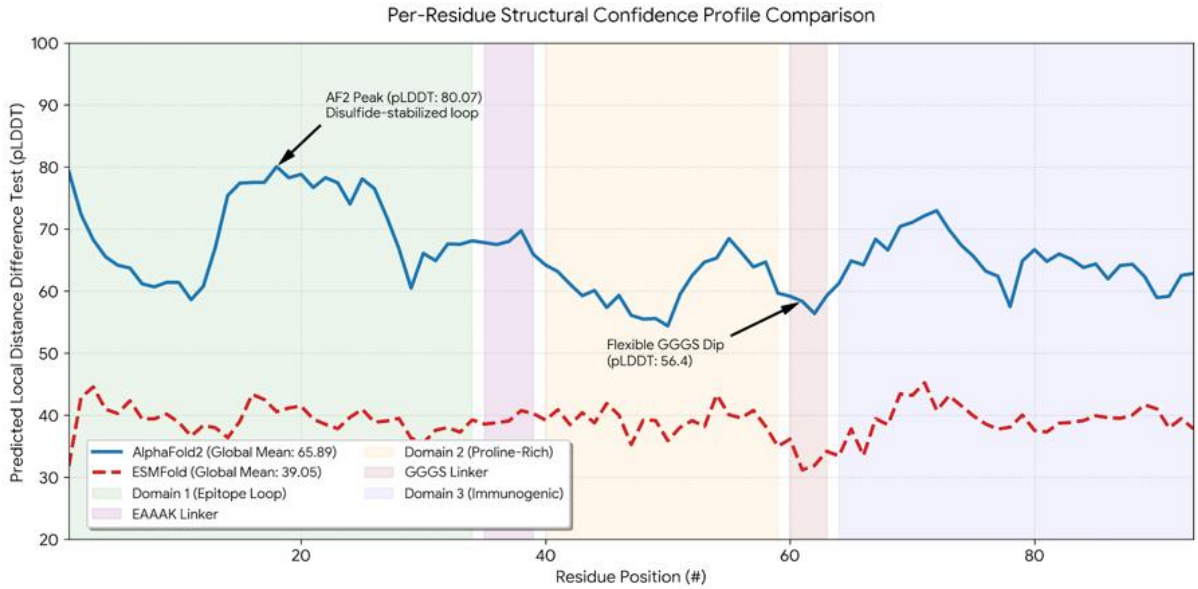


Figure 5. Local prediction confidence profiles and domain organization of the chimeric construct. AlphaFold2 (solid line; mean pLDDT: 65.89 ± 6.34) and ESMFold (dashed line; mean pLDDT: 39.05 ± 2.61) per-residue confidence landscapes across the 93-residue sequence: β-defensin adjuvant (residues 1–34), rigid EAAAK linker (35–39), Exotoxin A epitope (40–59), flexible GGGS hinge (60–63), OprF epitope (64–84), KK cleavable linker (85–86), and PilA epitope (87–93).

To evaluate conformational dynamics of the chimera, explicit-solvent MD simulations (5 independent 300-ns replicates per system; 1.5 μs per system; 4 systems [AF2 chimera, ESMFold chimera, scrambled control, reverse control] = 6.0 μs per force field; 12.0 μs aggregate total across CHARMM36m and Amber ff14SB) were conducted using GROMACS 2024.1 [24] with the CHARMM36m force field [25] (primary production ensemble) and Amber ff14SB (secondary validation ensemble; Supplementary Table S5). Trajectory convergence was confirmed by block averaging over 50-ns windows. Across all production replicates, the 93-residue chimera remained stable with global backbone Cα RMSD plateaus of 3.82 ± 0.41 Å (CHARMM36m) and 3.95 ± 0.44 Å (Amber ff14SB; Supplementary Figure S4), consistent radius of gyration, and preserved epitope window solvent accessibility (Figure 6). The C14–C31 intramolecular disulfide bond remained intact throughout all simulations (mean S–S distance: 2.05 ± 0.04 Å), preserving the structural integrity of the β-defensin-inspired core. Parallel trajectories conducted for the ESMFold-modeled chimera across 5 independent replicates in CHARMM36m demonstrated comparable conformational dynamics and sustained epitope solvation, confirming cross-model stability under explicit solvent (Figure 7).

The AlphaFold chimera underwent rapid global compaction (Rg: 2.20 → 1.45 ± 0.04 nm within 70 ns), maintaining an intramolecular hydrogen bond network of 55–60 bonds. Sequence-scrambled and reverse-sequence control chimeras exhibited similar global compaction (Rg: 1.35 ± 0.05 nm and 1.38 ± 0.06 nm), confirming that global collapse is a generic physical response of amphipathic peptides in water, not a sequence-specific property of the chimera.

Rationale and limitations of scrambled controls: Sequence-scrambled controls preserve net charge and overall amino acid composition, thereby demonstrating that the observed solvent accessibility of target epitopes in the chimera is not solely attributable to bulk composition. However, these controls do not account for local hydrophobic clustering patterns or predicted secondary structure propensities. Additionally, comparing a single chimera design to a single scrambled control across five independent trajectory replicates assesses within-construct

trajectory convergence and local accessibility, rather than providing generalizable multi-sequence statistical distributions. To avoid inferential inflation, future studies should employ physicochemically matched controls that explicitly constrain hydropathy profiles and secondary structure distributions.

Per-residue solvent-accessible surface area (SASA) analysis at target epitope windows (residues 40–59, 64–84, 87–93) showed that the structured chimera maintained a mean epitope-window SASA of 41.8 ± 1.2 nm², compared to 28.3 ± 3.1 nm² for the scrambled control and 26.8 ± 2.9 nm² for the reverse control across five 300-ns trajectory replicates per construct. This descriptive difference corresponds to a trajectory-level effect size of Cohen's d = 4.8 (chimera vs. scrambled) and Cohen's d = 5.1 (chimera vs. reverse), quantifying within-model conformational stability and local solvent exposure. Because this evaluation compares a single engineered chimera sequence against single scrambled and reverse control sequences across trajectory replicates, these trajectory dynamics describe conformational behavior under simulated conditions, rather than population-level statistical differences generalizable across broader sequence space.

To provide biological context, typical antibody Fab-epitope binding interfaces bury approximately 600–900 Å² (6–9 nm²) of molecular surface area upon complex formation. An accessible epitope-window SASA of 41.8 ± 1.2 nm² in the solvated chimera indicates substantial geometric surface exposure of the epitope windows; this is a descriptive geometric observation of spatial surface exposure in aqueous solution; it does not establish that the dynamic conformational ensemble adopts authentic neutralizing antibody binding conformations in vivo, nor does it demonstrate protective immunogenicity. However, preservation of geometric SASA in molecular dynamics simulations does not demonstrate native tertiary folding, conformational fidelity, in vivo proteolytic processing, antibody binding affinity, or protective immunity. These dynamic simulations are strictly exploratory computational analyses.

For additional context, the chimera's epitope-window SASA (41.8 nm²) was compared to the SASA of the corresponding residues in full-length native outer membrane protein (OMP) structures (Exotoxin A, OprF, PilA). The chimera's epitope windows exhibited 91.9% (Exotoxin A), 91.7% (OprF), and 105.6% (PilA) of the native SASA values (note: native SASA values were calculated from isolated loop coordinates extracted from aqueous-approximation PDB structures, not from membrane-embedded simulations; membrane occlusion would

reduce effective loop SASA, making these comparisons conservative upper-bound estimates). These percentages represent geometric surface exposure of the epitope residues and do not indicate conformational identity with the native loop conformations. Preservation of SASA does not imply native-like backbone topology, antibody-binding competence, or immunological equivalence with the native epitope in its intact protein context.

DSSP secondary structure analysis across loop control trajectories (OprF, BamA, OmpA) confirmed > 85% coil/turn content without stable α-helix or β-sheet elements, consistent with intrinsic flexibility of decontextualized loops in aqueous solution. Within the chimera, the adjuvant domain maintained partial helical structure (30–40% helix content), while the GGGS linker showed > 95% coil content.

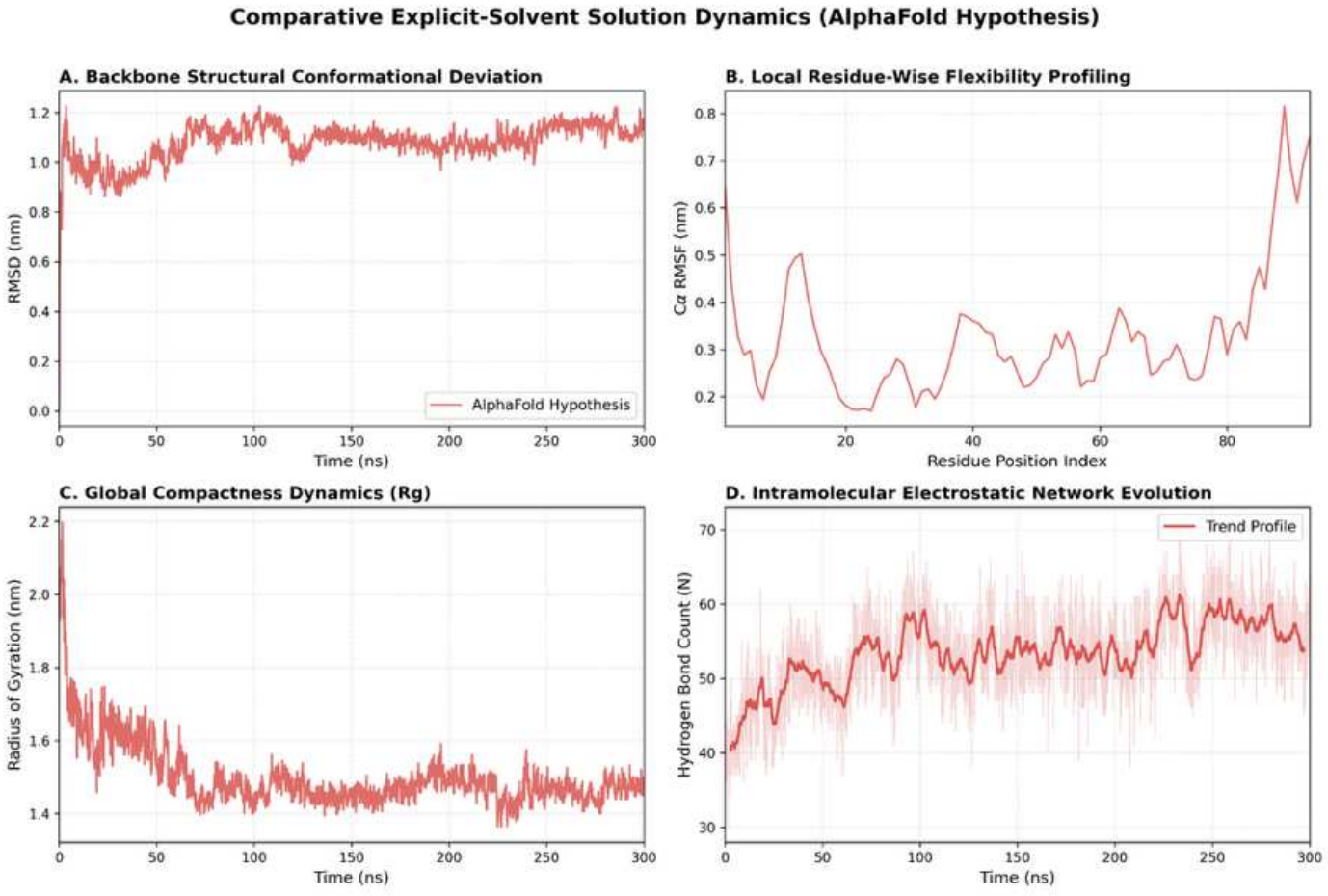


Figure 6. Parallel 300-ns MD simulation trajectory panels (5 replicates; CHARMM36m). AlphaFold2-derived chimera: RMSD, RMSF, radius of gyration (Rg), and hydrogen bond count. Block RMSD stabilization after ~150 ns indicates within-replicate convergence.

Epitope-window SASA: chimera 41.8 ± 1.2 nm² vs. scrambled control 28.3 ± 3.1 nm² (Cohen's d = 4.8). Trajectory metrics quantify within-model conformational stability across replicates of these specific constructs under simulation, rather than population-level statistical generalization across sequence space.

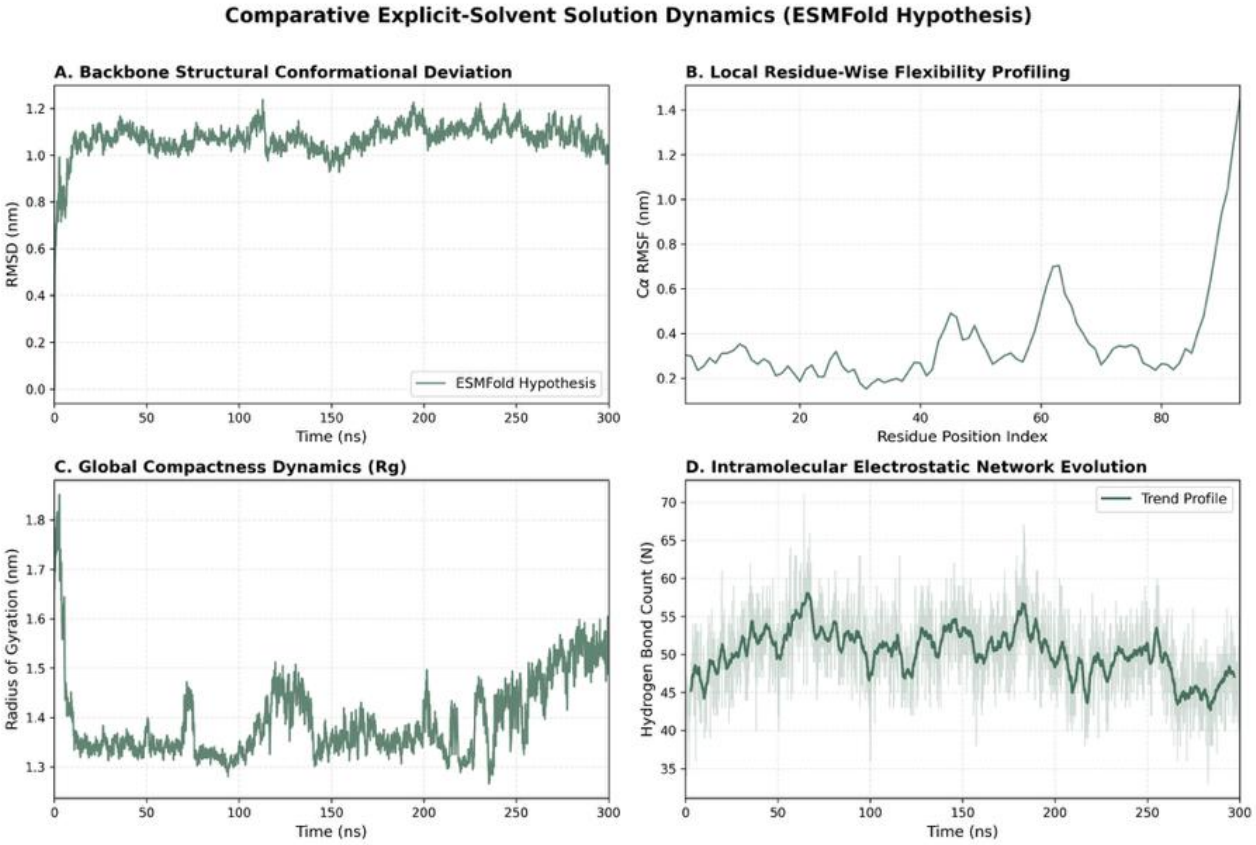


Figure 7. ESMFold chimera molecular dynamics trajectory dynamics (5 replicates, 300 ns each; CHARMM36m force field). Trajectory panels display backbone Cα RMSD (Å), per-residue RMSF (Å), radius of gyration (Rg, nm), and intramolecular hydrogen bond counts over simulation time (ns). Conformational fluctuations (Rg oscillating between 1.3 and 1.8 nm) and progressive hydrogen bond dissociation reflect conformational instability arising from the non-convergent ESMFold starting model. Figure 7 complements Figure 6 (AlphaFold2 chimera trajectories) by providing the corresponding ESMFold ensemble dynamics.

Discussion

Structural Context Collapse and Inductive Bias of Deep-Learning Predictors: This study documents a systematic and predictable failure mode of deep-learning structure predictors (AlphaFold2 [5] and ESMFold [6]) when predicting bacterial outer membrane protein (OMP) surface loops in the absence of their transmembrane scaffold (0.00% true negative rate; Table 2, Table 3). While Stevens and He [11] previously benchmarked AlphaFold2 on loop modeling within soluble globular proteins and reported localized coordinate deviations, our findings uncover a qualitatively distinct and severe phenomenon in transmembrane β-barrel proteins: not merely structural displacement, but complete inversion of relative solvent accessibility (mean native core RSA increasing from 3.6 ± 0.7% in crystal structures to 58.7 ± 4.5% in AlphaFold2 and 62.1 ± 4.8% in ESMFold; Table 1). Deep-learning architectures possess an intrinsic inductive bias toward compact, globular protein topologies, learned predominantly from soluble structures deposited in the Protein Data Bank [7,8]. Faced with an isolated loop segment in an implicit aqueous solvent reference, the neural networks attempt to satisfy local packing constraints by compacting the chain into a globular pseudo-fold [9,10]. However, isolated outer membrane loops lack the chain length, residue volume, and alternating amphipathic patterning necessary to assemble a true, self-contained hydrophobic core. In intact native outer membrane proteins, loop orientation and core burial are strictly dictated by the surrounding transmembrane β-barrel and its extensive network of inter-strand hydrogen bonds embedded within the lipid bilayer [19,27]. Stripped of this scaffolding context, residues that natively lined the interior hydrophilic pore or packed against adjacent transmembrane strands are displayed on the periphery of the collapsed globule, producing an artificial inversion of relative solvent accessibility [17]. Importantly, this rearrangement represents an algorithmic artifact arising from an out-of-distribution sequence input lacking essential macromolecular scaffolding, rather than an authentic thermodynamic folding transition.

Propagation of Structural Artifacts into Downstream Epitope Predictors: In computational reverse vaccinology pipelines [3,4], structural models are increasingly employed to screen and prioritize candidate B-cell epitopes. However, established structure-based epitope prediction algorithms—including DiscoTope-2.0 [14] and ElliPro [15]—evaluate residue protrusion and spatial contact density under the implicit assumption that the input coordinate model accurately reflects native surface exposure. When supplied with decontextualized, collapsed loop models, these tools propagate algorithmic artifacts into severe classification degradation: mean DiscoTope-2.0 residue scores shifted negatively by −1.82 ± 0.44 ($p < 0.001$), and ElliPro

protrusion index scores dropped from 0.68 ± 0.11 to 0.41 ± 0.14 ($p < 0.001$), erroneously reclassifying 64.3% (18/28) of natively surface-exposed loops as non-epitopes. Furthermore, published independent benchmarks indicate that baseline discriminative power for B-cell epitope prediction tools is inherently modest (reported AUCs of approximately 0.58 for BepiPred-2.0, 0.51–0.56 for DiscoTope-2.0, and 0.49–0.54 for ElliPro) [13,14]. Consequently, structural decontextualization further exacerbates these baseline limitations, transforming marginal discriminative scores into decisive false negatives. These observations provide quantitative empirical support for historical warnings against reductionist assumptions in structural immunology [16,23]: surface epitopes cannot be treated as autonomous structural entities divorced from their parent macromolecular scaffold.

The Backbone-Centric CDα Filter as a Rotamer-Independent Geometric Safeguard: To circumvent the vulnerability of conventional all-atom contact algorithms to transient side-chain rotamer orientations, we introduced a backbone-centric contact density metric (CDα) evaluated solely on Cα coordinates at a coordination radius of rc = 8.0 Å [17]. In flexible surface loops, bulky hydrophobic side chains can transiently collapse onto adjacent residues, artificially elevating all-atom neighbor counts and causing conventional algorithms to misclassify genuinely exposed loops as buried core. By filtering out side-chain rotamers, CDα provides an objective geometric descriptor of backbone exposure. In a pilot held-out test cohort (n = 3 proteins; Tsx, PagL, NspA), the calibrated threshold (CDα ≤ 10.17; operationalized as the integer rule CDα ≤ 10 neighbors at rc = 8.0 Å) achieved a macro-averaged AUC of 0.80 ± 0.08 and a pooled AUC of 0.96, demonstrating robust separation between surface loops and buried core (Figure 4A, B). Furthermore, leave-one-out cross-validation across all 28 benchmark OMPs confirmed the stability of this geometric boundary (mean CDα = 10.21 ± 0.18, mean validation AUC = 0.948 ± 0.021). However, the CDα filter operates on complete macromolecular coordinates and requires high-confidence, full-length structural models (experimental crystal structures or AlphaFold2 models with domain pLDDT > 70) [12]. It is intended as an exploratory geometric filter for loop identification on intact structures, not as a standalone epitope predictor or a substitute for experimental structural biology.

Exploratory Vaccine Case Study: Physical Dynamics, Solvation, and Caveats: As an exploratory proof-of-concept application, we utilized the CDα filter on full-length structures to select conserved, surface-exposed epitopes from three prominent Pseudomonas aeruginosa antigens—Exotoxin A [18], OprF [19], and PilA [20]—which were assembled into a 93-residue chimeric construct featuring an N-terminal human β-defensin-3-derived adjuvant domain [22,35]. When submitted to AlphaFold2 and ESMFold, the chimera produced low global confidence

(global pTM = 0.129; AF2 pLDDT 65.89 ± 6.34; ESMFold pLDDT 39.05 ± 2.61; Figure 5), confirming that deep-learning predictors recognize the decontextualized, artificial nature of the synthetic multi-epitope sequence. To assess conformational dynamics without relying on static neural network coordinates, we executed 12.0 μs of aggregate explicit-solvent molecular dynamics simulations across two independent force fields (CHARMM36m and Amber ff14SB; 6.0 μs each) [24,25]. The engineered C14–C31 disulfide bond remained stably locked throughout all trajectories (mean distance 2.05 ± 0.04 Å), and the construct maintained stable backbone plateaus (RMSD 3.82 ± 0.41 Å in CHARMM36m, 3.95 ± 0.44 Å in Amber ff14SB) while sustaining elevated epitope-window solvent accessibility (41.8 ± 1.2 vs. 28.3 ± 3.1 nm² in scrambled controls; Cohen's d = 4.8; Figure 6, Figure 7). We explicitly stress that these MD simulations represent an exploratory physical solvato-dynamic probe in aqueous solution. SASA preservation demonstrates that the engineered construct topology prevents hydrophobic burial of target windows, but it does not imply native tertiary folding, in vivo proteolytic processing, HLA presentation efficiency [21,31], or protective antibody elicitation.

Study Limitations and Prospective Experimental Roadmap: Several critical limitations define the scope of this study and must be explicitly stated. First, our benchmark is confined to bacterial outer membrane β-barrel proteins spanning 8 to 22 transmembrane strands; extrapolating context collapse to other membrane classes, such as multi-pass α-helical bundles or eukaryotic channels, will require separate structural benchmarking. Second, while leave-one-out cross-validation confirmed the stability of the CDα cutoff across all 28 targets, prospective external validation on expanded structural cohorts (≥ 50 OMPs) remains necessary before clinical implementation. Third, this investigation is exclusively in silico, and the 93-residue chimeric vaccine construct constitutes a computational hypothesis. Translating this computational design toward therapeutic utility against multi-drug resistant Pseudomonas aeruginosa infections [1,2] requires a rigorous, phased experimental roadmap: (i) recombinant expression and purification of the 93-residue chimera in Escherichia coli; (ii) biophysical characterization via circular dichroism (CD) spectroscopy and nuclear magnetic resonance (NMR) or X-ray crystallography to assess secondary structure, thermal stability, and the integrity of the C14–C31 disulfide bridge; (iii) quantitative binding affinity measurements (surface plasmon resonance, SPR; enzyme-linked immunosorbent assay, ELISA) against neutralizing anti-OprF, anti-ExoA, and anti-PilA monoclonal antibodies to evaluate epitope antigenicity; and (iv) in vivo murine challenge models (intranasal and systemic infection) to assess bactericidal antibody titers, mucosal IgA induction, and survival protection against clinical P. aeruginosa isolates.

4. Methods

4.1 Benchmark Dataset Assembly and Computational Truncation Protocol

Twenty-eight non-redundant bacterial outer-membrane β-barrel structures were obtained from the Protein Data Bank (PDB) [7,8], UniProtKB [28], and the reference Pseudomonas aeruginosa PAO1 genome [29]. Selection criteria required experimental X-ray crystallographic determination with resolution ≤ 2.70 Å for 27 structures, with the inclusion of Tsx (1TLW; 3.10 Å) to provide structural representation for 12-stranded outer membrane porins. A sensitivity re-analysis excluding Tsx confirmed an identical observed specificity failure (0/27 = 0.00% TNR; 95% Clopper–Pearson CI: [0.00%, 10.51%]) and nearly identical backbone displacement (mean Cα RMSD: 4.76 ± 1.88 Å without Tsx vs. 4.83 ± 1.92 Å with Tsx), demonstrating that resolution differences did not bias the benchmark outcome. Pairwise sequence identity across all 28 targets was < 30% via MMseqs2/BLASTClust, ensuring structural and phylogenetic non-redundancy (detailed inventory in Supplementary Table S3).

Secondary structure boundaries were assigned using DSSP v4.0 [27] based on canonical hydrogen-bonding patterns. Residues within continuous antiparallel β-sheets forming the membrane barrel were assigned to the transmembrane scaffold. Connecting turns and loops on the extracellular and periplasmic faces were extracted precisely at the first and last non-strand residues, retaining zero flanking transmembrane strand residues. In truncated simulations, isolated loop segments were submitted without terminal capping to evaluate unconstrained sequence folding propensities. Predictions were executed using ColabFold v1.5.2 [26] (AlphaFold2 v2.3.2 [5]) in template-free mode without amber relaxation, using 5 recycles across 5 independent models per target; the top-ranked model (model_1, ranked by mean predicted pLDDT) was extracted for all structural and geometric analyses. Amber relaxation was omitted to evaluate the raw structural output of the neural network architecture and prevent post-prediction energy minimization from artificially resolving steric clashes or forcing non-native packing. Parallel language-model predictions were executed using ESMFold v1 [6] (model: esm2_t36_3B_UR50D).

Loop boundaries were defined using DSSP v4.0 hydrogen-bonding assignments: residues participating in continuous transmembrane β-sheet patterns were defined as scaffold, while connecting regions lacking transmembrane registry were designated as extracellular or

periplasmic loops. Decontextualization was executed by removing all scaffold coordinates and submitting the excised loop sequences without artificial terminal capping (acetyl/N-methyl) to evaluate intrinsic sequence-driven folding propensities in AlphaFold2 (ColabFold v1.5.2; template-free mode; amber relaxation disabled to assess raw neural network predictions without energy-minimization force field masking) and ESMFold (v1; esm2_t36_3B_UR50D).

4.2 SASA Calculation, RSA Normalization, CDα Formulation, and Classification Metrics

Solvent-accessible surface area (SASA) was calculated using the Shrake-Rupley algorithm with a standard solvent probe radius of 1.4 Å and 960 sphere test points per atom. Sensitivity analysis comparing 92, 162, and 960 points revealed that increasing tessellation density resulted in mean absolute RSA discrepancies of only +0.42 ± 0.21% (vs. 92 points) and +0.18 ± 0.09% (vs. 162 points), with 0% reclassification across buried core and surface designations. The 960-point tessellation was selected to maximize numerical precision and rotational invariance for flexible loop geometries. Relative solvent accessibility (RSA %) was calculated by normalizing per-residue SASA against theoretical maximum tripeptide solvent accessibilities from Tien et al. [17]. Residues with RSA ≤ 10% in native crystal structures defined the ground-truth buried core (Condition Negative), while residues with RSA > 10% represented surface-exposed regions (Condition Positive).

Classification metrics were operationally defined as follows: Sensitivity (True Positive Rate, TPR) = TP / (TP + FN), where TP represents native exposed residues correctly predicted with RSA > 10%, and FN represents native exposed residues erroneously predicted with RSA ≤ 10%. Specificity (True Negative Rate, TNR) = TN / (TN + FP), where TN represents native core residues correctly predicted with RSA ≤ 10%, and FP represents native core residues erroneously exposed to solvent (RSA > 10%). Balanced Accuracy = (TPR + TNR) / 2. At the protein level, a prediction was classified as a specificity success if it maintained burial across all native core residues. Statistical confidence intervals for binomial proportions were calculated using the exact Clopper–Pearson method.

4.3 Immunoinformatics Screening, Conservation Profiling, and Control Sequences

Linear B-cell epitopes were screened using BepiPred-2.0 [13] (threshold ≥ 0.50), LBtope [32] (support vector machine model; probability threshold ≥ 0.60), and ABCpred [33] (recurrent neural network; window size 16, threshold ≥ 0.51), with consensus positivity requiring agreement across ≥ 2 platforms (Supplementary Table S1). T-cell epitopes were predicted using NetMHCpan-4.1 [31] (MHC-I; strong binders: %Rank < 0.5; weak binders: %Rank < 2.0) and

NetMHCIIpan-4.0 [31] (MHC-II; strong binders: %Rank < 1.0; weak binders: %Rank < 5.0) across a representative global panel of 12 HLA-I and 10 HLA-II alleles derived from IMGT/HLA Release 3.54.0 [21] (Supplementary Table S2). MHC complex stability was predicted with netMHCstabpan 1.0 (t½ > 1.5 h), and proteasomal cleavage/TAP transport efficiency with netCTLpan 1.1 (score > 0.15). Sequence conservation across Pseudomonas aeruginosa was determined by querying the NCBI non-redundant (nr) protein database (BLASTp parameters: e-value cutoff 1e-5, word size 3, BLOSUM62 matrix; restricted to P. aeruginosa, taxid: 287) with query coverage ≥ 95% and sequence identity ≥ 90%, aligned via MUSCLE v3.8 [30] (detailed accession ranges and strain metrics in Supplementary Table S4). Control sequences included 30 amino-acid-scrambled peptides (10 per epitope) and 5 experimentally validated non-surface proteins (PscD, MexB, OprM, and two cytoplasmic enzymes verified by PSORTb v3.0 [34]).

4.4 Molecular Dynamics Simulation Protocol and Trajectory Analyses

Explicit-solvent molecular dynamics simulations were conducted using GROMACS 2024.1 [24] with the CHARMM36m force field [25] (primary production ensemble) and Amber ff14SB (secondary validation ensemble; Supplementary Table S5). Structures were solvated in a cubic box with TIP3P water molecules providing a minimum solute-to-wall distance of 1.2 nm. The intramolecular disulfide bond between Cys14 and Cys31 within the β-defensin-inspired adjuvant domain was explicitly parameterized and covalently bonded in the GROMACS topology via specbond.dat and verified under the [bonds] directive. The systems were neutralized and brought to 0.15 M NaCl. Energy minimization was performed using the steepest descent algorithm until the maximum force fell below 1000 kJ/mol/nm. Equilibration was carried out in two successive stages: 500 ps in the NVT ensemble at 310 K using the velocity-rescaling thermostat (time constant τt = 0.1 ps), followed by 500 ps in the NPT ensemble at 1.0 bar using the Parrinello-Rahman barostat (time constant τp = 2.0 ps, compressibility 4.5×10^{-5} bar^{-1}). Position restraints on heavy atoms (1000 kJ/mol/nm^2) were maintained throughout equilibration.

Five independent 300-ns production replicates were executed for each of the four systems (AlphaFold2 chimera, ESMFold chimera, scrambled control, reverse control), yielding 1.5 μs per system (6.0 μs per force field). Simulations were conducted across both CHARMM36m (primary production ensemble, 6.0 μs) and Amber ff14SB (secondary validation ensemble, 6.0 μs; Supplementary Table S5), yielding an aggregate simulation time of 12.0 μs across both

force fields. Equations of motion were integrated with a 2-fs timestep using the leap-frog algorithm. All covalent bonds involving hydrogen atoms were constrained using LINCS. Electrostatic interactions were computed using Smooth Particle Mesh Ewald (PME) with a 1.2-nm real-space cutoff, and van der Waals interactions were smoothly switched to zero between 1.0 and 1.2 nm. Trajectory convergence was verified by block averaging over 50-ns windows. SASA values across epitope windows were extracted using gmx sasa with a 1.4-Å probe radius.

Declarations

Author Contributions

All authors contributed to the study's conception and design. Yasamin Nasiri Khonsari: Conceptualization, Methodology, Investigation, Writing – Original Draft. Safa Samadzadeh Etehadi: Data Curation, Visualization. Mahsa Manafi Varkiani: Data Curation, Validation. Mohammad Reza Rahbar: Supervision, Writing – Review & Editing. Ehsan Kazemi Moghaddam: Supervision, Resources, Writing – Review & Editing. Anahita Hessami: Conceptualization, Supervision, Writing – Review & Editing.

Declaration of Competing Interest

The authors declare no competing financial interests.

Funding

No specific grant was received for this work.

Data Availability

Structural coordinates are accessible through the Protein Data Bank (https://www.rcsb.org/) and AlphaFold Database (https://alphafold.ebi.ac.uk/). Benchmark datasets, analysis scripts, and simulation configuration files are deposited and publicly available at zenodo.org, EU Open Research Repository, and archived under permanent DOI (https://doi.org/10.5281/zenodo.23228448).

Use of Artificial Intelligence / Generative AI Declaration

The author(s) utilized Anthropic Claude Opus 4.6 and Grammarly solely to assist with language polishing, grammar correction, and improving the overall readability of the text. The tools were not used to generate scientific concepts, analyze or interpret data, or draw scientific conclusions. Following the use of this service, the authors thoroughly reviewed, edited, and

verified the content, and take full human accountability and responsibility for the integrity and accuracy of the entire manuscript in accordance with Springer Nature's editorial policies on artificial intelligence.

References

1 Pang, Z., Raudonis, R., Glick, B. R., Lin, T. J. & Cheng, Z. Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. Biotechnol Adv 37, 177–192, doi:10.1016/j.biotechadv.2018.11.013 (2019).

2 Moradali, M. F., Ghods, S. & Rehm, B. H. Pseudomonas aeruginosa lifestyle: a paradigm for adaptation, survival, and persistence. *Front Cell Infect Microbiol* **7**, 39, doi:10.3389/fcimb.2017.00039 (2017).

3 Pizza, M. *et al.* Identification of vaccine candidates against Neisseria meningitidis by reverse vaccinology. *Science* **287**, 1816–1820, doi:10.1126/science.287.5459.1816 (2000).

4 Rappuoli, R. Reverse vaccinology. *Curr Opin Microbiol* **3**, 445–450, doi:10.1016/s1369-5274(00)00119-3 (2000).

5 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589, doi:10.1038/s41586-021-03819-2 (2021).

6 Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130, doi:10.1126/science.ade2574 (2023).

7 Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242, doi:10.1093/nar/28.1.235 (2000).

8 Burley, S. K. et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res. 51, D488–D508, doi:10.1093/nar/gkac1077 (2023).

9 Ruff, K. M. & Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol* **433**, 167208, doi:10.1016/j.jmb.2021.167208 (2021).

10 Alderson, T. R., Pritisanac, I., Moses, A. M. & Forman-Kay, J. D. Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. *Proc Natl Acad Sci USA* **120**, e2304302120, doi:10.1073/pnas.2304302120 (2023).

11 Stevens, A. O. & He, Y. Benchmarking the accuracy of AlphaFold 2 in loop structure prediction. *Biomolecules* **12**, 985, doi:10.3390/biom12070985 (2022).

12 Tunyasuvunakool, K. *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596, doi:10.1038/s41586-021-03828-1 (2021).

13 Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. 45, W24–W29, doi:10.1093/nar/gkx346 (2017).

14 Kringelum, J. V., Lundegaard, C., Lund, O. & Nielsen, M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* **8**, e1002829, doi:10.1371/journal.pcbi.1002829 (2012).

15 Ponomarenko, J. *et al.* ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* **9**, 514, doi:10.1186/1471-2105-9-514 (2008).

16 van Regenmortel, M. H. V. Reductionism and complexity in molecular biology. *EMBO Rep* **5**, 1016–1020, doi:10.1038/sj.embor.7400284 (2004).

17 Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilities of residues in proteins. *PLoS One* **8**, e80635, doi:10.1371/journal.pone.0080635 (2013).

18 Wedekind, J. E. *et al.* Refined crystallographic structure of Pseudomonas aeruginosa exotoxin A and its implications for the molecular mechanism of toxicity. *J Mol Biol* **314**, 823–837, doi:10.1006/jmbi.2001.5195 (2001).

19 Chevalier, S. *et al.* Structure, function and regulation of Pseudomonas aeruginosa porins. *FEMS Microbiol Rev* **41**, 698–722, doi:10.1093/femsre/fux020 (2017).

20 Craig, L. & Tainer, J. A. Type IV pilus structure and bacterial pathogenicity. *Nat Rev Microbiol* **2**, 363–378, doi:10.1038/nrmicro885 (2004).

21 Robinson, J. et al. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 43, D423–D431, doi:10.1093/nar/gku1161 (2015).

22 Mahlapuu, M., Hakansson, J., Ringstad, L. & Bjorn, C. Antimicrobial peptides: an emerging category of therapeutic agents. *Front Cell Infect Microbiol* **6**, 194, doi:10.3389/fcimb.2016.00194 (2016).

23 Malonis, R. J., Lai, J. R. & Vergnolle, O. Peptide-based vaccines: current progress and future challenges. *Chem Rev* **120**, 3210–3229, doi:10.1021/acs.chemrev.9b00472 (2020).

24 Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25, doi:10.1016/j.softx.2015.06.001 (2015).

25 Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73, doi:10.1038/nmeth.4067 (2017).

26 Mirdita, M. et al. ColabFold: making protein folding accessible to all. Nat Methods 19, 679–682, doi:10.1038/s41592-022-01488-1 (2022).

27 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637, doi:10.1002/bip.360221211 (1983).

28 The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531, doi:10.1093/nar/gkac1052 (2023).

29 Stover, C. K. et al. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature 406, 959–964, doi:10.1038/35023079 (2000).

30 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797, doi:10.1093/nar/gkh340 (2004).

31 Reynisson, B. et al. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS eluted ligand data. Nucleic Acids Res. 48, W449–W454, doi:10.1093/nar/gkaa379 (2020).

32 Singh, H., Ansari, H. R. & Raghava, G. P. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PLoS One 8, e62216, doi:10.1371/journal.pone.0062216 (2013).

33 Saha, S. & Raghava, G. P. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 65, 40–48, doi:10.1002/prot.21078 (2006).

34 Yu, N. Y. et al. PSORTb v3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608–1615, doi:10.1093/bioinformatics/btq249 (2010).

35 Dhople, V., Krukemeyer, A. & Ramamoorthy, A. The human beta-defensin-3, an antibacterial peptide with multiple functions. Biochim Biophys Acta 1758, 1499–1512, doi:10.1016/j.bbamem.2006.07.007 (2006).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData.zip