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.

Biophysical dissection of high-signal phenotypes in AI-guided evolution of leucine-rich repeat binders

Junjae Park
Chungbuk National University
Saehae Choi
Osong Medical Innovation Foundation
Eun Yi Jeon
Osong Medical Innovation Foundation
Hyeonjeong Yu
Osong Medical Innovation Foundation
Cho Hyun Kim
Osong Medical Innovation Foundation
Jieun Lee
Osong Medical Innovation Foundation
Soyoung An
Osong Medical Innovation Foundation
Myungho Sohn
Osong Medical Innovation Foundation
Won-Kyu Lee
gregorio7@korea.kr

Osong Medical Innovation Foundation

Research Article

Keywords: protein language model, leucine-rich repeat, variable lymphocyte receptor, repebody, avidity, biophysical validation

Posted Date: June 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9788709/v1

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

Additional Declarations: No competing interests reported.

Abstract

Artificial-intelligence (AI)-guided protein evolution can enrich variants with stronger screening readouts, but the physical origin of those readouts may differ from the intended molecular property. This distinction is particularly important for topology-constrained repeat scaffolds, where local mutations can perturb folding cooperativity, solubility, and assembly state. Here we analyzed high-signal variants recovered during AI-guided evolution of two hen egg-white lysozyme (HEWL)-binding leucine-rich repeat (LRR) binders. In a sequence-only protein language model (PLM)-guided repebody trajectory, apparent enzyme-linked immunosorbent assay (ELISA) activity increased across rounds; the top variant C219A reached approximately 12-fold WT-normalized apparent activity but lost cooperative unfolding behavior and showed early light scattering, consistent with enrichment of assay-high but biophysically compromised variants. In a related variable lymphocyte receptor (VLR) trajectory that included structure-informed seed variants before PLM-guided optimization, A6C emerged in round 4 as a wild-type (WT)-background single mutant with approximately 10^3-fold WT-normalized apparent activity and a broadly WT-like unfolding profile. However, biolayer interferometry (BLI) under a one-to-one immobilized assay geometry showed an apparent KD of 69.61 nM for A6C compared with 31.27 nM for WT, arguing against conventional monomeric affinity maturation. Reducing/non-reducing electrophoresis, HEWL-coating-concentration-dependent ELISA, and AlphaFold3 multimer modeling supported a working model in which A6C is associated with a disulfide-linked, higher-molecular-weight assembly that enhances apparent surface binding through avidity-like effects. These results emphasize the need for orthogonal biophysical validation in topology-constrained scaffold engineering.

Introduction

Artificial-intelligence-guided protein engineering is changing how experimental protein variants are selected. Protein language models, active-learning workflows, and inverse-folding methods now allow protein engineers to prioritize variants from limited experimental datasets rather than relying solely on exhaustive library construction and screening[1–5]. This shift is particularly important for protein binders, where functional improvement often requires navigating large mutational spaces while preserving expression, folding, solubility, and binding specificity.

The central strength of AI-guided evolution is also its central vulnerability: the model optimizes the experimental label it is given. A model can efficiently enrich variants with stronger screening readouts, but a stronger readout is not necessarily equivalent to improved molecular function. This distinction is especially important for solid-phase binding assays such as ELISA. In such formats, signal can reflect monomeric binding affinity, but it can also be shaped by nonspecific adsorption, partial unfolding, aggregation-associated retention, antigen density, oligomerization, and surface-enabled avidity [6–8]. Thus, the apparent success of AI-guided evolution must be interpreted through the physical state of the selected protein, not through assay signal alone.

This problem becomes acute in topology-constrained repeat-protein scaffolds. Repeat proteins are attractive engineering platforms because their modular architectures provide regular, programmable surfaces for molecular recognition [9–11]. Designed ankyrin repeat proteins, variable lymphocyte receptors, and repebodies illustrate how repeat-based architectures can be adapted into non-immunoglobulin binding scaffolds [11–17]. However, the stability of repeat proteins is not simply the sum of independent local elements. In many repeat scaffolds, folding and stability depend on cooperative packing between neighboring repeats, terminal capping interactions, long-range coupling, and maintenance of the global solenoid architecture [18–20]. Under these conditions, a mutation that improves local surface chemistry can simultaneously disrupt foldability, unfolding cooperativity, solubility, or aggregation resistance.

Leucine-rich repeat binders provide a particularly informative setting for this problem. LRR proteins form curved solenoid architectures with concave recognition surfaces that are well suited for ligand binding [14, 15, 17, 18]. Variable lymphocyte receptors from jawless vertebrates and engineered repebodies are structurally related LRR-based scaffolds that have been developed as antibody alternatives [14–17]. Their compact single-chain formats, modularity, and programmable binding surfaces make them attractive for protein engineering. At the same time, their curved topology and inter-repeat coupling make them vulnerable to local sequence perturbations. In such scaffolds, the practical challenge is not merely to obtain a higher screening signal; it is to determine whether the selected variant remains a folded, biophysically interpretable protein.

AI-guided evolution in LRR binders therefore raises a mechanistic question: what does a high-scoring variant actually represent? One possibility is classical affinity maturation, in which mutations improve one-to-one binding between a monomeric binder and its target. A second possibility is structural failure, in which partial unfolding or aggregation produces an assay-high state that is rewarded by the screening format. A third possibility is an alternative but structurally intact solution, such as a change in oligomeric state or valency that enhances surface binding through avidity. These outcomes can appear similar in a primary ELISA screen but correspond to fundamentally different protein states.

Here we investigated the physical origins of high-signal phenotypes recovered during AI-guided evolution of two related HEWL-binding LRR scaffolds. In a sequence-only PLM-guided repebody trajectory, apparent ELISA activity increased across rounds, but orthogonal biophysical validation revealed structurally compromised variants with abnormal unfolding behavior and early light scattering. In a related VLR trajectory that included structure-informed seed variants before PLM-guided optimization, A6C emerged as a high-signal variant that retained a broadly WT-like unfolding profile. However, A6C was not explained by improved fitted one-to-one binding parameters. Instead, reducing/non-reducing electrophoresis, HEWL-coating-concentration-dependent ELISA, and multimer modeling supported a working model of disulfide-linked, avidity-like surface enhancement. We use these trajectories to illustrate two experimentally observed high-signal outcomes in related LRR scaffolds. These results show that AI-guided evolution of topology-constrained LRR binders can recover high-signal variants

through distinct physical mechanisms, and that orthogonal biophysical validation is essential for interpreting model-guided optimization outcomes.

Materials and Methods

DNA synthesis, plasmid construction, and mutagenesis

All oligonucleotide primers and synthetic DNA fragments were obtained from Bionics (Seoul, Republic of Korea). Coding sequences were codon-optimized in silico for expression in Escherichia coli using the GenScript online codon-optimization tool, and optimized genes were synthesized by Bionics. Genes encoding the repebody, VLR, and derivative variants were cloned into pET-28a(+) by Gibson assembly such that the expressed proteins carried a C-terminal His6 tag and no additional N-terminal fusion sequence. PCR amplification was performed using Q5 High-Fidelity DNA Polymerase (New England Biolabs), and amplicons were purified using the QIAquick PCR Purification Kit (QIAGEN). Gibson Assembly Master Mix (New England Biolabs) was used according to the manufacturer's instructions.

Site-directed mutagenesis was performed using an overlap-based Gibson assembly strategy. Mutagenic primers were designed to generate two overlapping PCR fragments containing the desired substitution(s). The pET-28a(+) backbone was PCR-linearized, assembled with the corresponding mutant fragments, transformed into chemically competent E. coli DH5alpha cells, and sequence-verified by Sanger sequencing. Construct sequences and oligonucleotide primer sequences are provided in Online Resource 1.

Protein expression and purification

For recombinant expression, all constructs were transformed into SHuffle T7 competent Escherichia coli cells [21]. Cells were grown in LB medium supplemented with kanamycin (50 μg/mL) at 37°C until the optical density at 600 nm reached 0.6–0.9. Protein expression was induced with 1 mM isopropyl beta-D-1-thiogalactopyranoside, and cultures were incubated at 18°C for 20 h after induction in 200 mL culture volumes. Cells were harvested by centrifugation at 4,000 × g for 20 min, resuspended in phosphate-buffered saline (PBS, pH 7.4) supplemented with protease inhibitors including 1 mM PMSF, and lysed by sonication. The lysate was clarified by centrifugation at 18,000 × g for 40 min.

The clarified supernatant was applied to Ni-NTA Superflow resin (QIAGEN), washed with PBS containing 20 mM imidazole, and eluted with PBS containing 100 mM imidazole. Eluted proteins were concentrated using Amicon Ultra centrifugal filters with a 3-kDa molecular-weight cutoff (Merck). Final purification was performed by size-exclusion chromatography on a HiLoad 16/600 Superdex 75 pg column (Cytiva) equilibrated in PBS. The column was equilibrated with two column volumes of PBS, and separation was performed at a flow rate of 1.0 mL/min. Peak fractions corresponding to the principal SEC peak were pooled and used for downstream analyses. SEC was used as a preparative purification step to collect

the principal soluble protein peak and was not used to quantify the solution-state oligomeric distribution of A6C.

ELISA-based apparent binding assay

Apparent binding activity toward HEWL was measured by indirect ELISA. High-binding 96-well plates (SPL32296; SPL Life Sciences) were coated overnight at 4°C with HEWL at 1 μg/mL in PBS. Plates were washed three times with PBS containing 0.05% Tween 20 and blocked with 5% skim milk in PBS for 1 h at 37°C.

Purified WT and mutant proteins were serially diluted threefold over eight points from a starting concentration of 300 μg/mL and incubated on antigen-coated plates for 1 h at 37°C. After washing, bound proteins were detected using HRP-conjugated anti-6×His tag antibody (ab1187; Abcam) diluted 1:5,000 in blocking buffer. Signal was developed with 1-Step TMB ELISA Substrate Solution, stopped with ELISA Stop Solution, and measured at 450 nm on a VersaMax microplate reader. Dose-response curves were fitted using a four-parameter logistic model to obtain apparent EC50 values.

For the VLR ELISA panels shown in Fig. 2b, proteins were prepared from the same starting concentration. The round-4 candidate panel compared A6C with other round-4 candidates. The residue-6 comparator panel compared WT, A6C, A6N, and A6I using a 12-point 1:5 serial dilution series from 300 μg/mL to 6.14 × 10^-6 μg/mL to cover the broader A6C response range while preserving matched starting conditions.

For comparative analysis and model training, apparent activity was expressed as the ratio of WT apparent EC50 to variant apparent EC50. Larger fold-change values therefore indicate stronger apparent plate-based binding than WT, whereas values below 1 indicate weaker apparent binding. WT was measured in each experimental round and used as the within-round reference for apparent activity normalization.

Sequence-only AI-guided evolution of repebody

To evaluate sequence-only optimization in a topology-constrained LRR scaffold, the HEWL-binding repebody was subjected to PLM-guided active learning using the EVOLVEpro framework [4]. The default ESM-2 15B protein language model embedding implemented within EVOLVEpro was used for sequence representation, and variant ranking was performed using the framework's default top-layer regression model. Experimental fitness was defined from ELISA-derived apparent activity, expressed as WT-normalized fold change in apparent EC50.

Optimization was performed over five iterative rounds. We targeted 12 candidate variants per round for experimental evaluation; the exact variants and final numbers tested in each round are provided in Online Resource 2. Newly acquired measurements were appended to the cumulative dataset and used for

model retraining in the subsequent round. For orthogonal biophysical validation of the sequence-only repebody trajectory, selected ELISA-ranked variants from each round were advanced to Prometheus Panta analysis together with WT and inverse-folding-derived (IF-derived) non-trajectory control variants; representative control traces are shown in Fig. 1b, and complete traces are provided in Online Resource 5. Candidate variants were selected from EVOLVEpro-ranked nominations without imposing additional biophysical filters such as cysteine exclusion, predicted aggregation propensity, or predicted stability constraints; exact tested variants and apparent activity labels are provided in Online Resource 2.

Structure-informed initialization and VLR AI-guided evolution

To introduce a minimal structural prior for VLR evolution, we used AiCE, an inverse-folding framework that integrates structural and evolutionary constraints into mutation nomination [5]. The crystal structure of the HEWL-bound VLR complex (PDB 3G3A) was used as the source structure, and the VLR backbone chain alone was extracted as the structural input. Primary sequence nomination was performed using the default recommended structural-constraint settings (beta = 0.8, gamma = 0.5), and coupled-mutation prioritization was guided by the statistical coupling analysis implemented in AiCE.

Among the highest-priority coupled mutation suggestions, Q54N/N99N and K75N/N99N were selected as seed-relevant outputs. Because residue 99 remained unchanged at the amino-acid level, these variants are referred to as Q54N and K75N. Although AiCE also ranked the Q54N/K75N pair highly, Q54N and K75N were introduced as separate single-substitution seed variants to maintain a single-substitution initialization scheme for the subsequent active-learning trajectory and to avoid introducing an additional double-mutant lineage at initialization. Q54N and K75N were introduced as structure-informed seed variants within the initial VLR screening pool rather than as independently evolved trajectories. Subsequent AI-guided evolution of the HEWL-binding VLR was carried out using EVOLVEpro over five iterative rounds. A6C was recovered in round 4 of the VLR EVOLVEpro trajectory as a single substitution relative to the VLR WT background. A6N and A6I were nominated in round 5 and were also used as residue-6 comparators to test whether the phenotype reflected a general position-6 effect or a cysteine-specific molecular consequence.

Thermal unfolding and light-scattering analysis

Thermal unfolding and aggregation-associated behavior were characterized using a Prometheus Panta instrument (NanoTemper Technologies). Purified proteins were analyzed at 1 mg/mL in 1× PBS using Prometheus Standard Capillaries. Intrinsic fluorescence and light scattering were monitored simultaneously during thermal ramping from 20 to 95°C at 1°C/min, and unfolding behavior was evaluated from the fluorescence ratio at 350/330 nm. When a clear cooperative transition was present, apparent melting temperatures were assigned from the inflection point of the unfolding trace.

For the repebody, WT, selected sequence-only top variants, and inverse-folding-derived (IF-derived) non-trajectory control variants were analyzed in parallel; representative control traces are shown in Fig. 1b.

For the VLR, WT, A6N, and A6C were analyzed under identical conditions. Light-scattering traces were used to assess the onset of aggregation-associated behavior during heating. Measurements were recorded in triplicate capillaries. Non-cooperative traces were interpreted qualitatively and were not assigned definitive melting temperatures.

Biolayer interferometry

Binding kinetics were measured by biolayer interferometry on an Octet QK384 instrument (Sartorius BioAnalytical Instruments) using Ni-NTA biosensors in HBS-EP+ buffer as the running buffer [22]. His-tagged VLR WT or A6C was loaded onto Ni-NTA biosensors at 5 µg/mL, and soluble HEWL was used as the analyte in a five-point twofold dilution series from 78.1 nM to 4.88 nM. Association and dissociation were monitored for 100 s and 60 s, respectively. Sensorgrams were fitted using a 1:1 binding model to derive apparent kon, koff, and KD values. WT and A6C were analyzed under the same loading, analyte, and fitting conditions. Because A6C is interpreted below as an assembly-prone variant under the tested conditions, fitted kinetic values for A6C are treated as apparent parameters under the tested immobilization geometry rather than definitive monomeric affinity constants. Apparent kinetic parameters and fit statistics are summarized in Online Resource 6.

AlphaFold3 multimer modeling

To assess whether A6C could adopt an arrangement compatible with interchain disulfide bond formation, structural modeling was performed using AlphaFold3 [23]. Two identical A6C chains and one HEWL molecule were submitted in multimer mode. The top-ranked model was used to inspect the geometry around residue 6 and the relative arrangement of the two VLR chains. The model was used as a geometric plausibility check for position-6 cysteine proximity rather than as proof of the solution-state assembly or the full multivalent antigen-binding geometry. HEWL was omitted from displayed structural views where appropriate for clarity. Structural figures were prepared in PyMOL version 3.1.6.1.

Reducing and non-reducing SDS-PAGE

To evaluate disulfide-linked assembly, purified proteins were analyzed by SDS-PAGE under reducing and non-reducing conditions on Any kD Mini-PROTEAN TGX Stain-Free Protein Gels (Bio-Rad). Purified WT and A6C samples were adjusted to the same nominal protein concentration, mixed with reducing or non-reducing loading buffer, and loaded in equal volumes. Reducing samples were prepared in loading buffer supplemented with beta-mercaptoethanol and heated before loading. Non-reducing samples were prepared in parallel without beta-mercaptoethanol. WT and A6C from the same purification batches were compared under the two conditions.

HEWL-coating-concentration-dependent ELISA

To assess coating-concentration-dependent binding behavior as a proxy for apparent surface antigen density, HEWL was coated onto ELISA plates starting at 50 µg/mL followed by 1:3 serial dilution. After blocking, WT VLR and A6C were applied at 100 µg/mL under otherwise identical conditions, and bound proteins were detected using the anti-His HRP-based ELISA format described above. Relative signal differences between WT and A6C across the HEWL coating-concentration series were used to assess avidity-like behavior as a function of coating concentration. Measurements were performed in triplicate wells.

Replication and data analysis

Unless otherwise stated, ELISA-based measurements were performed in triplicate wells and thermal unfolding measurements were recorded in triplicate capillaries using the indicated purified protein preparations. ELISA data were fitted using a four-parameter logistic model, and summary values are reported as mean +/- standard deviation where applicable. Apparent activity values used for AI-guided optimization were calculated as WT-normalized fold change in apparent EC50. Because apparent activity was normalized to the WT measured within each experimental round, fold changes are best interpreted as within-round comparisons and as qualitative round-to-round trajectories.

Results

Sequence-only AI-guided evolution recovers assay-high but biophysically compromised repebody variants

To examine how a sequence-only AI-guided evolution workflow behaves in a topology-constrained LRR scaffold, we first evolved a HEWL-binding repebody using a PLM-guided active-learning framework. Across five rounds of optimization, we targeted 12 variants per round for expression and ELISA screening; the exact variants and final numbers tested in each round are provided in Online Resource 2. Apparent binding activity, expressed as WT-normalized fold change in apparent EC50, progressively increased over successive rounds, giving the initial impression of assay-level improvement (Fig. 1a). The top-performing variant, C219A, reached an approximately 12-fold increase in apparent activity relative to WT.

Biophysical characterization did not support a simple interpretation as improved binder performance. WT repebody displayed a well-defined cooperative unfolding transition in Prometheus Panta measurements, with an apparent melting temperature near 69°C and low baseline light scattering. In contrast, selected top variants recovered from the sequence-only trajectory failed to preserve this behavior. C219A, in particular, did not exhibit a clear cooperative unfolding transition, and its light-scattering signal rose sharply from an anomalously low temperature of approximately 43°C (Fig. 1b; Table 1). This profile is consistent with pronounced structural destabilization and entry into an aggregation-prone state before canonical thermal unfolding.

Table 1
Representative high-signal states and biophysical interpretation

Scaffold	Variant	Context	ELISA	Orthogonal validation	Interpretation
Repebody	WT	Reference	1.0	Tm ~ 69°C; low scattering	Folded reference
Repebody	C219A	Sequence-only PLM	~ 12-fold	No clear transition; scattering onset ~ 43°C	Assay-high; compromised / aggregation-prone
VLR	WT	Reference	1.0	WT-like thermal profile; BLI apparent KD = 31.27 nM; no higher-MW band	Folded reference
VLR	A6C	Round 4; WT-background single mutant	~ 10^3-fold	WT-like thermal profile; BLI apparent KD = 69.61 nM; higher-MW non-reducing band	Disulfide-linked higher-MW species; avidity-like surface binding

Values are representative summary values used to interpret the physical state of selected variants; BLI values are apparent, geometry-dependent parameters under the tested immobilization conditions.

The key observation is that the strongest apparent screening phenotypes in this trajectory were associated with poor biophysical behavior. The model therefore did not fail to optimize the supplied experimental label; rather, it optimized a label that was only partially aligned with the intended molecular property. Once structurally compromised or aggregation-prone states can generate high plate-based signal, active learning can be drawn toward assay-high but biophysically compromised variants. This is the operational failure mode captured by the repebody trajectory.

Because the ELISA readout measures retained His-tagged material on an antigen-coated solid phase after washing rather than solution affinity alone, structurally compromised repebody variants could increase apparent signal through several nonexclusive mechanisms, including exposure of sticky surface patches, enhanced retention on the coated plate, aggregation-associated local enrichment, or altered wash resistance. The Prometheus data do not prove aggregation during the ELISA incubation itself, but they identify a compromised, aggregation-prone state compatible with such plate-retention mechanisms.

Structure-informed seed variants define the initial VLR search pool

The repebody trajectory indicated that high apparent activity can be decoupled from protein quality in an LRR scaffold. We therefore examined a related HEWL-binding VLR trajectory in which the initial search

pool included a small number of structure-informed seed variants. Using AiCE on the VLR backbone derived from PDB 3G3A, Q54N/N99N and K75N/N99N were identified among the highest-priority coupled mutation suggestions in the highest-confidence output class. Because residue 99 remained unchanged at the amino-acid level, these outputs are referred to as Q54N and K75N. These variants were introduced into the initial VLR screening pool before subsequent PLM-guided active learning.

This structural input was used only at initialization; subsequent variant nomination, experimental screening, and round-to-round model updating were performed by PLM-guided active learning using ELISA-derived apparent activity as the experimental label.

A VLR trajectory identifies A6C, a cysteine-dependent high-signal variant with WT-like cooperative thermal behavior

Across the VLR rounds, the screen recovered several higher-signal variants, with A6C emerging as the dominant outlier in round 4. A6C was recovered in the VLR EVOLVEpro trajectory as a single substitution relative to the VLR WT background and showed an approximately three-orders-of-magnitude increase in WT-normalized apparent activity (Fig. 2a,b; Table 1). In the round-4 dose-response comparison, A6C was highlighted as the high-signal outlier relative to other tested round-4 candidates (Fig. 2b, left). To assess whether this phenotype reflected cysteine-specific chemistry or general mutability at position 6, we examined additional residue-6 variants. A6N and A6I were nominated in round 5 of the AI-guided trajectory and also served as residue-6 comparators in a direct comparison with WT and A6C (Fig. 2b, right). Neither variant reproduced the A6C high-signal phenotype, supporting the interpretation that the effect was specific to cysteine substitution rather than to residue-6 substitution alone.

The central question was whether A6C represented another structurally compromised state analogous to the repebody top variants. To test this, we compared the thermal behavior of WT, A6C, and A6N using Prometheus Panta. WT and A6C showed broadly similar cooperative thermal transitions and did not exhibit the low-temperature collapse-associated scattering pattern observed in the sequence-only repebody trajectory (Fig. 2c). A6N is shown as an additional residue-6 comparator and likewise did not reproduce the repebody-like collapse-prone profile. These data indicate that the high-signal A6C phenotype was not accompanied by overt low-temperature aggregation-associated scattering detectable by this assay.

A6C high apparent activity is not explained by fitted one-to-one binding parameters

To determine whether the A6C phenotype reflected conventional monomeric affinity maturation, we compared VLR WT and A6C by BLI under the same Ni-NTA immobilization geometry. VLR WT showed an apparent KD of 31.27 nM, with an apparent kon of 1.303 x 10^6 M^-1 s^-1 and koff of 4.076 x 10^-2 s^-1. In contrast, A6C showed an apparent KD of 69.61 nM, with an apparent kon of 8.735 x 10^5 M^-1 s^-1 and koff of 6.080 x 10^-2 s^-1 (Fig. 2d; Online Resource 6). Thus, the large increase in apparent ELISA signal was not accompanied by a corresponding improvement in fitted one-to-one binding parameters.

Because later analyses indicated that A6C is associated with a disulfide-linked higher-molecular-weight species, these BLI fits are interpreted as apparent parameters under the tested immobilization geometry rather than definitive monomeric affinity constants. The BLI experiment was used to test whether A6C showed a commensurate improvement under a fitted one-to-one geometry, not to define a complete binding model for the assembly-prone species. Nevertheless, the BLI results argue against interpreting A6C as a classical affinity-matured monomer and motivated further analysis of the mechanism underlying its plate-based signal.

A6C forms a disulfide-linked higher-molecular-weight species with avidity-like surface binding

The mismatch between the ELISA and BLI results suggested that A6C increased plate-based signal through a mechanism other than improved monomeric affinity. AlphaFold3 multimer modeling provided a geometric plausibility check for whether the position-6 cysteines could approach an interchain disulfide-compatible arrangement (Fig. 3a). Non-reducing SDS-PAGE showed that A6C displayed a prominent higher-molecular-weight species, whereas reducing conditions restored the major band near the monomeric position (Fig. 3b). This reducing-sensitive shift supports a disulfide-linked assembly associated with the cysteine introduced at position 6.

We next asked whether this altered assembly state could account for the binding phenotype. In HEWL-coating-concentration-dependent ELISA, A6C displayed the strongest advantage over WT at high HEWL coating concentration, whereas the difference diminished as coating concentration decreased (Fig. 3c). This coating-concentration dependence is more consistent with an avidity-like surface-binding phenotype than with a simple increase in monomeric affinity alone [7, 8]. Together, these results support a working model in which A6C is associated with a disulfide-linked, higher-molecular-weight assembly that enhances apparent surface binding through avidity-like effects.

Discussion

This study shows that high screening signals recovered during AI-guided evolution of topology-constrained protein binders can arise from distinct physical mechanisms. In the LRR systems examined here, stronger ELISA signals did not map onto a single molecular interpretation. In one trajectory, high apparent activity was associated with structural compromise and early aggregation behavior. In another, high apparent activity was associated with a broadly WT-like thermal profile and an assembly-dependent, avidity-like surface phenotype. The practical implication is direct: in structurally constrained scaffolds, the output of AI-guided evolution cannot be interpreted from the primary screening signal alone.

The repebody trajectory illustrates a failure mode that is likely to be broadly relevant to model-guided protein engineering. Sequence-only PLM-guided evolution increased apparent ELISA activity across rounds, giving the appearance of assay-level improvement. However, the top-scoring variants lost cooperative unfolding behavior and showed early light scattering. These data indicate that the model enriched variants that performed well under the experimental label but poorly as folded proteins. The

problem was not that the model failed to optimize. Rather, the model optimized a screening signal that was partially uncoupled from the intended molecular property. When structural damage can produce or enhance plate-based signal, active learning can be pulled toward assay-high but biophysically compromised variants. Related directed-evolution studies have emphasized that aggregation, poor stability, and developability liabilities often need to be actively selected against rather than assumed to be absent [24, 25].

This distinction is important because protein language models are often discussed as if sequence plausibility, functional score, and molecular quality are naturally aligned. Recent Protein Journal work applying ESM-2 embeddings to protein property classification illustrates the expanding use of PLM-derived representations for protein-level prediction tasks [27]. In many systems, sequence-derived signals may be sufficiently aligned with useful engineering objectives. In topology-constrained repeat scaffolds, however, this alignment can break down. LRR binders are modular but not structurally independent at each repeat; comparative analyses of LRR architectures further emphasize that repeat class, variable-segment geometry, and inter-repeat organization shape the final solenoid [28]. Local substitutions can influence neighboring repeat packing, terminal capping, curvature, and global solenoid stability. A variant can therefore appear favorable in sequence space and in an assay readout while leaving the structurally admissible region required for a usable binder. For such scaffolds, biophysical validation is not a secondary characterization step; it is part of defining whether evolution has succeeded.

The VLR A6C trajectory provides the complementary lesson. A6C showed a large increase in apparent ELISA activity while retaining a broadly WT-like unfolding profile. This indicates that an assay-high phenotype can occur without the collapse-associated behavior observed in the sequence-only repebody trajectory. However, A6C was not a straightforward affinity-matured monomer. BLI under the tested Ni-NTA immobilization geometry did not show a commensurate improvement in fitted one-to-one binding parameters, and the combined biochemical and structural evidence supported a disulfide-linked higher-molecular-weight assembly compatible with avidity-like surface binding. Thus, A6C represents a noncanonical, assembly-dependent high-signal solution class whose apparent signal is driven by valency and assay geometry rather than by improved monomeric affinity.

This finding should not be dismissed as merely an artifact. In protein engineering, assembly state and valency can influence effective binding, surface retention, and target engagement [8, 26]. The important point is not that avidity-like solutions are intrinsically good or bad, but that they are mechanistically different from monomeric affinity maturation. For diagnostic capture, immobilized-target detection, or applications involving high target surface density, an assembly-dependent avidity-like binder may be useful. For applications requiring defined monomeric affinity, solution homogeneity, or predictable pharmacokinetics, the same phenotype would require further engineering and characterization. AI-guided evolution can therefore uncover useful solution classes, but those classes must be identified experimentally.

Several limitations define the scope of our conclusions. First, the repebody and VLR trajectories were performed in related but non-identical LRR scaffolds and should not be interpreted as a controlled comparison between sequence-only and structure-informed evolution. Second, A6C assembly is inferred from non-reducing electrophoresis, HEWL-coating-concentration-dependent ELISA, and structural modeling; these data support an avidity-like working model but do not quantitatively define the solution-state oligomeric distribution. This cautious interpretation is consistent with recent discussions of AlphaFold as a hypothesis-generating complement to, rather than a replacement for, experimental structural and biophysical validation [29]. Third, because ELISA detection used an anti-His antibody, the A6C signal may include contributions from both increased effective valency and increased detectable epitope number. Finally, BLI parameters were obtained under an immobilized 1:1 fitting geometry and are therefore treated as apparent, geometry-dependent parameters.

Overall, these results argue for a protein-level interpretation of AI-guided evolution. The model proposes variants and the assay assigns scores, but the protein determines what those scores mean. In LRR binders, high apparent activity can reflect structurally compromised states or assembly-dependent avidity-like surface binding rather than monomeric affinity maturation. The correct interpretation emerges only when the screening signal is connected to folding behavior, aggregation propensity, binding kinetics, and assembly state. For topology-constrained scaffolds, orthogonal biophysical validation is therefore not optional; it is required to define the molecular outcome of AI-guided protein evolution.

Conclusion

AI-guided evolution can rapidly enrich high-signal variants, but in topology-constrained LRR binders those signals can originate from different physical states. In a sequence-only repebody trajectory, increased apparent ELISA activity was associated with loss of cooperative unfolding and early light scattering, consistent with enrichment of assay-high but biophysically compromised variants. In a VLR trajectory that included structure-informed seed variants, A6C retained a broadly WT-like unfolding profile but gained apparent surface signal through a disulfide-linked higher-molecular-weight assembly and avidity-like surface effects rather than improved fitted one-to-one affinity. These findings show that high-throughput screening output alone is insufficient to define engineering success in structurally constrained scaffolds. AI-guided evolution of repeat-protein binders should be paired with orthogonal biophysical validation to determine whether recovered variants improve the intended molecular property, damage the scaffold, or acquire an alternative assay-defined mechanism.

Declarations

Acknowledgements

The authors thank Jeoungbin Kim, Yerin Chi, YooKun Won, and Sung Ik Cho at KAIST, and Soojae Lee at Chungbuk National University, for helpful discussions and technical assistance.

Statements and Declarations

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (grant no. RS-2023-00235511).

Competing interests

The authors declare no competing interests.

Author Contributions

Junjae Park: Conceptualization; Methodology; Investigation; Formal analysis; Visualization; Writing – original draft; Writing – review & editing.

Saehae Choi: Investigation; Writing – review & editing.

Eun Yi Jeon: Investigation; Writing – review & editing.
Hyeonjeong Yu: Investigation; Writing – review & editing.

Cho Hyun Kim: Investigation; Writing – review & editing.

Jieun Lee: Investigation; Writing – review & editing.

Soyoung An: Investigation; Writing – review & editing.

Myungho Sohn: Investigation; Writing – review & editing.

Won-Kyu Lee: Conceptualization; Supervision; Project administration; Funding acquisition; Writing – review & editing.

Data availability

Processed data supporting the main conclusions of this study are included in the manuscript and Supporting Information. Construct sequences, oligonucleotide primer sequences, ELISA-derived apparent activity values used for AI-guided optimization, raw Prometheus traces, filtered AiCE output tables, and BLI fitted kinetic parameters are provided as Online Resource files. Representative BLI sensorgrams and global fits are shown in the main figures. Additional raw data retained by the authors can be provided by the corresponding author upon reasonable request.

Use of artificial intelligence tools

A generative artificial-intelligence tool was used for language editing, drafting assistance, and organization of text. The tool was not used to generate experimental data, analyze raw data, create

primary scientific conclusions, or produce final figures. All scientific content, experimental design, data analysis, interpretation, references, and final manuscript text were reviewed and verified by the authors.

References

1. 1. Yang KK, Wu Z, Arnold FH (2019) Machine-learning-guided directed evolution for protein engineering. Nat Methods 16:687–694. https://doi.org/10.1038/s41592-019-0496-6
2. 2. Biswas S, Khimulya G, Alley EC, et al (2021) Low-N protein engineering with data-efficient deep learning. Nat Methods 18:389–396. https://doi.org/10.1038/s41592-021-01100-y
3. 3. Rives A, Meier J, Sercu T, et al (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci 118:e2016239118. https://doi.org/10.1073/pnas.2016239118
4. 4. Jiang K, Yan Z, Di Bernardo M, et al (2024) Rapid in silico directed evolution by a protein language model with EVOLVEpro. Science 387:eadr6006. https://doi.org/10.1126/science.adr6006
5. 5. Fei H, Li Y, Liu Y, et al (2025) Advancing protein evolution with inverse folding models integrating structural and evolutionary constraints. Cell 188:4674–4692.e19. https://doi.org/10.1016/j.cell.2025.06.014
6. 6. Güven E, Duus K, Lydolph MC, et al (2014) Non-specific binding in solid phase immunoassays for autoantibodies correlates with inflammation markers. J Immunol Methods 403:26–36. https://doi.org/10.1016/j.jim.2013.11.014
7. 7. Lehtonen O-P, Viljanen MK (1980) Antigen density in ELISA; Effect on avidity dependency. J Immunol Methods 36:63–70. https://doi.org/10.1016/0022-1759(80)90094-0
8. 8. Erlendsson S, Teilum K (2021) Binding Revisited-Avidity in Cellular Function and Signaling. Front Mol Biosci 7:615565. https://doi.org/10.3389/fmolb.2020.615565
9. 9. Andrade MA, Perez-Iratxeta C, Ponting CP (2001) Protein Repeats: Structures, Functions, and Evolution. J Struct Biol 134:117–131. https://doi.org/10.1006/jsbi.2001.4392
10. 10. Main ER, Jackson SE, Regan L (2003) The folding and design of repeat proteins: reaching a consensus. Curr Opin Struct Biol 13:482–489. https://doi.org/10.1016/S0959-440X(03)00105-2
11. 11. Boersma YL, Plückthun A (2011) DARPins and other repeat protein scaffolds: advances in engineering and applications. Curr Opin Biotechnol 22:849–857. https://doi.org/10.1016/j.copbio.2011.06.004
12. 12. Binz HK, Stumpp MT, Forrer P, et al (2003) Designing Repeat Proteins: Well-expressed, Soluble and Stable Proteins from Combinatorial Libraries of Consensus Ankyrin Repeat Proteins. J Mol Biol 332:489–503. https://doi.org/10.1016/S0022-2836(03)00896-9
13. 13. Binz HK, Amstutz P, Plückthun A (2005) Engineering novel binding proteins from nonimmunoglobulin domains. Nat Biotechnol 23:1257–1268. https://doi.org/10.1038/nbt1127
14. 14. Pancer Z, Cooper MD (2006) The evolution of adaptive immunity. Annu Rev Immunol 24:497–518. https://doi.org/10.1146/annurev.immunol.24.021605.090542

15. 15. Han BW, Herrin BR, Cooper MD, Wilson IA (2008) Antigen Recognition by Variable Lymphocyte Receptors. Science 321:1834–1837. https://doi.org/10.1126/science.1162484
16. 16. Lee S-C, Park K, Han J, et al (2012) Design of a binding scaffold based on variable lymphocyte receptors of jawless vertebrates by module engineering. Proc Natl Acad Sci 109:3299–3304. https://doi.org/10.1073/pnas.1113193109
17. 17. Waters EA, Shusta EV (2018) The variable lymphocyte receptor as an antibody alternative. Curr Opin Biotechnol 52:74–79. https://doi.org/10.1016/j.copbio.2018.02.016
18. 18. Bella J, Hindle KL, McEwan PA, Lovell SC (2008) The leucine-rich repeat structure. Cell Mol Life Sci 65:2307–2333. https://doi.org/10.1007/s00018-008-8019-0
19. 19. Kloss E, Courtemanche N, Barrick D (2008) Repeat-protein folding: New insights into origins of cooperativity, stability, and topology. Arch Biochem Biophys 469:83–99. https://doi.org/10.1016/j.abb.2007.08.034
20. 20. Tokuriki N, Tawfik DS (2009) Protein Dynamism and Evolvability. Science 324:203–207. https://doi.org/10.1126/science.1169375
21. 21. Lobstein J, Emrich CA, Jeans C, et al (2012) SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm. Microb Cell Factories 11:753. https://doi.org/10.1186/1475-2859-11-56
22. 22. Kamat V, Rafique A (2017) Designing binding kinetic assay on the bio-layer interferometry (BLI) biosensor to characterize antibody-antigen interactions. Anal Biochem 536:16–31. https://doi.org/10.1016/j.ab.2017.08.002
23. 23. Abramson J, Adler J, Dunger J, et al (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493–500. https://doi.org/10.1038/s41586-024-07487-w
24. 24. Famm K, Winter G (2006) Engineering aggregation-resistant proteins by directed evolution. Protein Eng Des Sel 19:479–481. https://doi.org/10.1093/protein/gzl032
25. 25. Buchanan A, Ferraro F, Rust S, et al (2012) Improved drug-like properties of therapeutic proteins by directed evolution. Protein Eng Des Sel 25:631–638. https://doi.org/10.1093/protein/gzs054
26. 26. Yip SH-C, Foo J-L, Schenk G, et al (2011) Directed evolution combined with rational design increases activity of GpdQ toward a non-physiological substrate and alters the oligomeric structure of the enzyme. Protein Eng Des Sel 24:861–872. https://doi.org/10.1093/protein/gzr048
27. 27. Narang K, Nath A, Hemstrom W, et al (2024) HaloClass: Salt-Tolerant Protein Classification with Protein Language Models. Protein J 43:1035–1044. https://doi.org/10.1007/s10930-024-10236-7
28. 28. Batkhishig D, Bilguun K, Enkhbayar P, et al (2018) Super Secondary Structure Consisting of a Polyproline II Helix and a beta-Turn in Leucine Rich Repeats in Bacterial Type III Secretion System Effectors. Protein J 37:223–236. https://doi.org/10.1007/s10930-018-9767-9
29. 29. Li MQC, Wang S, Lin SR, et al (2026) Advantages and Limitations of AlphaFold in Structural Biology: Insights from Recent Studies. Protein J 45:22–38. https://doi.org/10.1007/s10930-025-10310-8

Figures

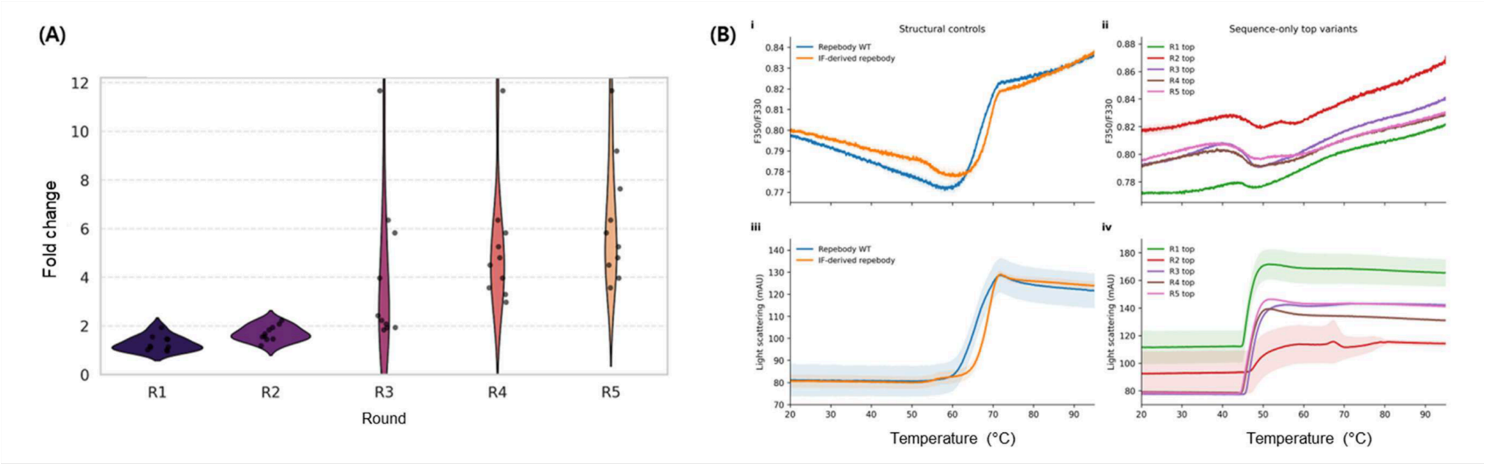


Figure 1

Sequence-only AI-guided evolution recovers assay-high but biophysically compromised repebody variants. (a) Apparent ELISA activity trajectories across five rounds of sequence-only PLM-guided evolution of a HEWL-binding repebody. (b) Prometheus Panta analysis of WT, selected top variants, and a representative inverse-folding-derived (IF-derived) non-trajectory control. Representative traces are shown; complete Prometheus traces are provided in Online Resource 5. Top sequence-only variants showed abnormal unfolding profiles and early light scattering, consistent with structural compromise rather than straightforward improvement in molecular binding quality.

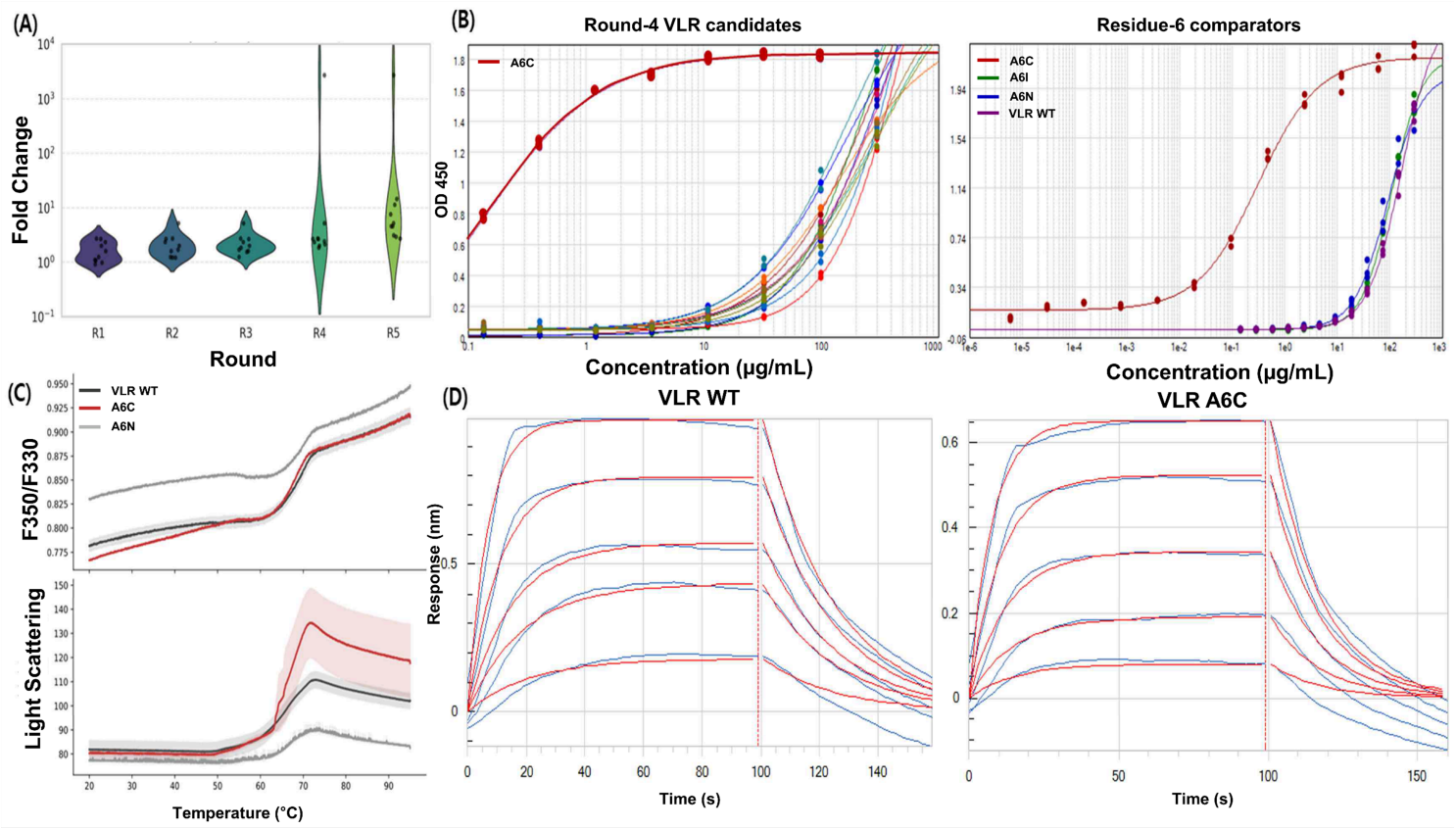


Figure 2

A VLR trajectory identifies A6C, a cysteine-dependent high-signal variant with WT-like cooperative thermal behavior but no fitted one-to-one affinity gain. (a) Apparent activity trajectories during VLR evolution. (b) Left, ELISA dose-response analysis of representative round-4 VLR candidates, with A6C highlighted in red and other round-4 candidates shown as thin unlabeled curves. Right, direct comparison of WT, A6C, A6N, and A6I using a 12-point 1:5 serial dilution series from the same starting concentration. A6N and A6I were nominated in round 5 and used as residue-6 comparators. Exact tested variant identities are provided in Online Resource 2. (c) Prometheus Panta analysis of WT, A6C, and A6N. (d) BLI sensorgrams and global 1:1 fits for VLR WT and A6C; blue traces indicate observed sensorgrams and red traces indicate fitted 1:1 curves. Fitted values are interpreted as apparent parameters under the tested Ni-NTA immobilization geometry.

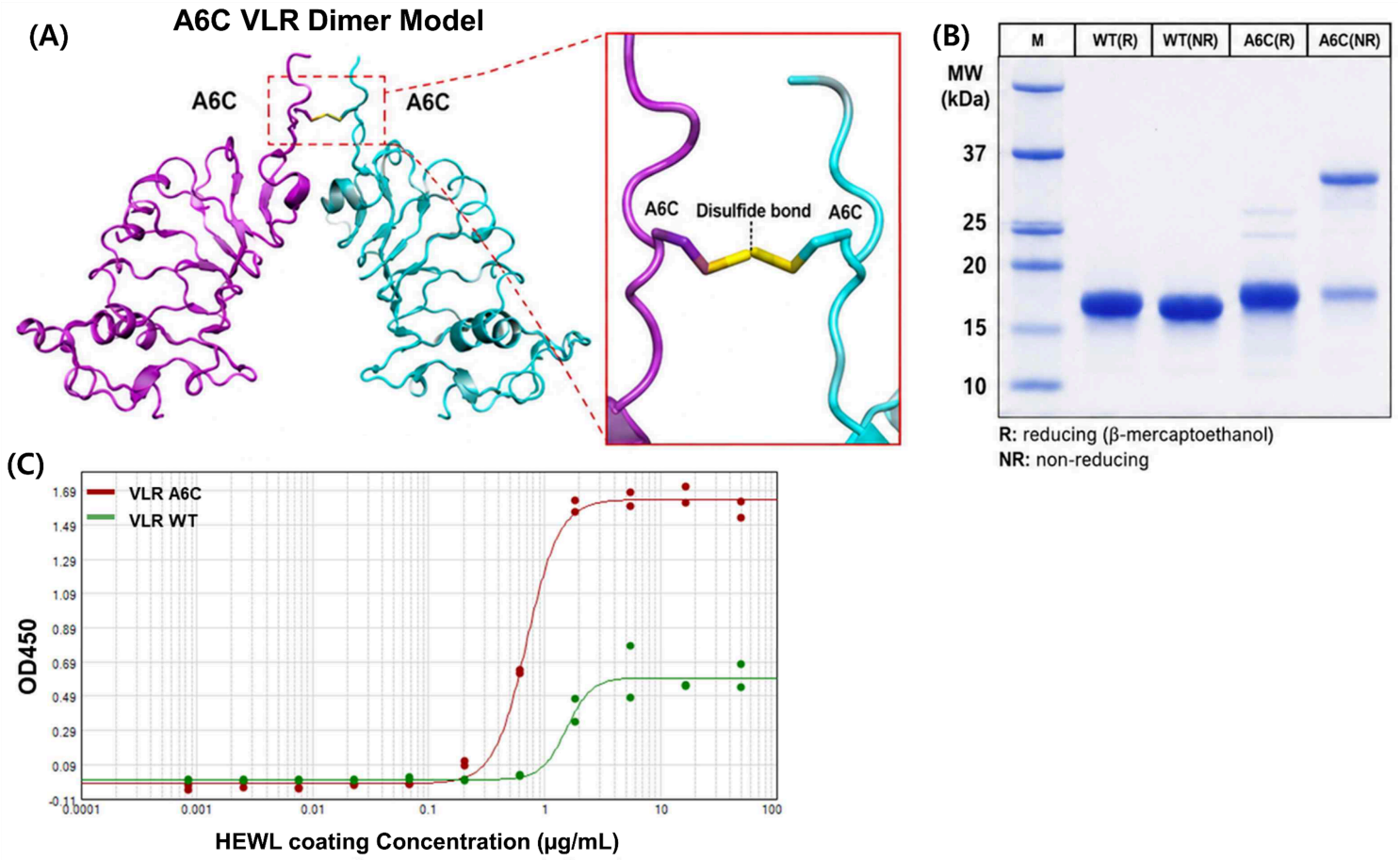


Figure 3

A6C forms a disulfide-linked higher-molecular-weight species with coating-concentration-dependent, avidity-like surface binding. (a) AlphaFold3 multimer model of an A6C disulfide-linked dimeric arrangement, used as a geometric plausibility check for interchain disulfide formation between position-6 cysteines. (b) Reducing and non-reducing SDS-PAGE analysis of VLR WT and A6C. A6C showed a prominent higher-molecular-weight species under non-reducing conditions and returned near the monomeric position under reducing conditions. (c) HEWL-coating-concentration-dependent ELISA of WT

and A6C. A6C showed the strongest signal advantage at high HEWL coating concentration, and this advantage decreased as coating concentration was reduced, supporting a coating-concentration-dependent, avidity-like surface-binding interpretation.

Supplementary Files

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

- Table1.xlsx
- OnlineResource1.xlsx
- OnlineResource3.xlsx
- OnlineResource2.xlsx
- OnlineResource6.xlsx
- OnlineResource4.xlsx
- OnlineResource5.xlsx