Programmable Icosahedral Protein Nanocages: A De Novo Design Framework for Reducing Kinetic Mis-Assembly and Enhancing Experimental Assembly Fidelity

Ashfaq Hussain *

Posted Date: 28 May 2026

doi: 10.20944/preprints202605.1980.v1

Keywords: de novo protein design; symmetric protein cages; polyhedral nanocages; icosahedral nanomaterials; self-assembling protein nanomaterials; computational protein engineering

 doi:10.20944/preprints202605.1980.v1

Article

Programmable Icosahedral Protein Nanocages: A De Novo Design Framework for Reducing Kinetic Mis-Assembly and Enhancing Experimental Assembly Fidelity

Ashfaq Hussain

Independent Researcher, Pakistan; ahussain797@gmail.com

Abstract

The de novo design of symmetric protein nanocages has emerged as a major frontier in programmable biomolecular engineering due to its potential applications in nanomedicine, vaccine development, molecular encapsulation, catalysis, and synthetic cellular systems. Despite substantial progress in computational protein design, the reliable construction of large polyhedral assemblies remains limited by kinetic mis-assembly, off-pathway oligomerization, interface instability, and low experimental success rates. This study presents a computationally integrated framework for the design of programmable icosahedral protein nanocages with improved assembly fidelity. The proposed strategy combines symmetry-guided scaffold selection, Rosetta-based symmetric docking, RFdiffusion-driven interface backbone generation, ProteinMPNN sequence optimization, and AlphaFold2-Multimer computational filtering to produce experimentally tractable cage architectures. The target system consists of a two-component assembly formed from trimeric and dimeric oligomeric building blocks organized into a 60-subunit icosahedral nanocage approximately 25 nm in diameter. The framework emphasizes reduction of non-specific intermolecular interactions through interface-specific design constraints and computational pre-screening of unstable configurations. Predicted outcomes indicate that the integration of diffusion-based backbone generation with sequence-level optimization and AF2-based structural validation substantially improves the likelihood of obtaining stable self-assembling cages while reducing experimental screening burdens. The study further evaluates how this integrated workflow compares with existing approaches in symmetric protein cage engineering. Collectively, the proposed methodology provides a scalable route toward high-fidelity programmable protein nanomaterials and contributes to the broader development of de novo self-assembling biomolecular systems.

Keywords: de novo protein design; symmetric protein cages; polyhedral nanocages; icosahedral nanomaterials; self-assembling protein nanomaterials; computational protein engineering

1. Introduction

Symmetric protein cages constitute one of the most sophisticated organizational motifs found in biological systems and represent a major focus within the broader field of de novo designed self-assembling protein nanomaterials with programmable geometry. Naturally occurring protein cages such as ferritin, encapsulins, viral capsids, lumazine synthases, and bacterial microcompartments demonstrate how biological systems exploit symmetry to achieve structural robustness, spatial organization, molecular encapsulation, and efficient biochemical processing [2,13–15]. The de novo recreation of these capabilities enables the rational construction of nanoscale architectures with programmable mechanical, chemical, and biological functions. In biological systems, symmetric cages support compartmentalization of metabolic reactions, sequestration of toxic intermediates,

 doi:10.20944/preprints202605.1980.v1

protection of nucleic acids, and long-range molecular transport, thereby illustrating how geometric order contributes directly to biological functionality [13–15].

In drug design and therapeutic delivery, polyhedral protein nanocages provide an attractive platform for programmable encapsulation and controlled release of biologically active cargo [3,4,16]. Their monodisperse geometry, tunable surface chemistry, and genetically encoded assembly properties make them particularly suitable for targeted delivery applications involving chemotherapeutics, nucleic acids, imaging agents, and immunomodulatory molecules. Icosahedral cages are especially advantageous because they maximize enclosed volume relative to surface area while maintaining high structural stability [3]. Furthermore, precisely engineered cage surfaces can display ligands, peptides, or antibody fragments in defined spatial arrangements, improving receptor targeting and multivalent binding efficiency. The ability to computationally control pore size, surface charge distribution, and internal cavity chemistry creates opportunities for highly selective therapeutic delivery systems with reduced off-target toxicity.

Within proteomics, symmetric protein cages offer powerful opportunities for molecular organization, enzyme colocalization, and synthetic compartmentalization [17]. Artificial nanocages can spatially confine catalytic pathways, thereby increasing local substrate concentrations and reducing diffusion-mediated inefficiencies. This feature is particularly important in multistep enzymatic cascades where intermediate stability limits overall pathway productivity. Engineered cages may also function as nanoscale reaction vessels capable of stabilizing transient intermediates or protecting sensitive enzymes from environmental degradation. Additionally, symmetric protein assemblies facilitate structural proteomics by serving as scaffolds for cryogenic electron microscopy (cryo-EM), improving particle visibility and symmetry averaging during high-resolution reconstruction [18]. The combination of programmable geometry and molecular specificity therefore positions protein cages as versatile tools for next-generation proteomic engineering.

In genomics and nucleic acid engineering, designed protein cages offer promising solutions for gene delivery, genome editing, and nucleic acid stabilization [19]. Protein nanocages can be engineered to selectively package DNA, RNA, CRISPR-associated ribonucleoproteins, or messenger RNA payloads while protecting these molecules from nuclease degradation during systemic transport. Unlike many viral vectors, de novo designed cages can potentially avoid preexisting immunogenicity while preserving efficient cellular uptake characteristics. Moreover, programmable assembly principles may enable sequence-specific nucleic acid recognition and compartmentalization, facilitating controlled intracellular delivery and temporally regulated gene expression systems. These capabilities connect protein cage engineering directly to emerging applications in synthetic genomics and precision medicine.

The relevance of symmetric protein cage design to human health extends beyond delivery systems into vaccine engineering, diagnostics, biomaterials, and regenerative medicine [4,20]. Computationally designed cages have already demonstrated remarkable success as multivalent antigen presentation platforms capable of eliciting potent neutralizing immune responses against viral pathogens [4]. Ordered antigen display enhances B-cell receptor clustering and improves immunogenicity compared with soluble antigen formulations. In addition, programmable protein assemblies may function as biosensors, imaging scaffolds, or molecular carriers for tissue engineering applications. Their biodegradability, genetically encoded synthesis, and structural precision make them highly attractive alternatives to synthetic nanoparticle systems that may suffer from toxicity or poor biocompatibility.

The current state of development in symmetric protein cage engineering reflects rapid progress driven by advances in computational structural biology, machine learning, and high-resolution structural characterization techniques [1,6,9]. Early efforts focused primarily on fusion-based assembly strategies that combined naturally oligomeric domains using linker engineering and rigid-body symmetry constraints [21]. Although these approaches demonstrated proof-of-concept assembly formation, they often suffered from limited geometric flexibility and low assembly predictability. The introduction of Rosetta-based symmetric docking frameworks significantly

 doi:10.20944/preprints202605.1980.v1

improved atomic-level control over interface geometry and enabled the de novo construction of tetrahedral, octahedral, and icosahedral assemblies [1]. Subsequent developments in cryo-EM validation demonstrated that computationally designed cages could achieve near-atomic agreement with predicted models [1,6].

More recently, deep learning approaches have transformed the design landscape by introducing generative backbone modeling and neural network–guided sequence optimization [9–11]. RFdiffusion enables the generation of backbone geometries unconstrained by preexisting structural templates, thereby expanding the accessible design space for symmetric assemblies [9]. ProteinMPNN further improves sequence compatibility with designed backbones by learning sequence–structure relationships directly from large-scale protein datasets [10]. Meanwhile, AlphaFold2-Multimer provides a highly effective computational filter for evaluating foldability and interface confidence prior to experimental testing [11]. Together, these methods have substantially reduced the empirical burden associated with protein nanocage development. Nevertheless, despite these advances, experimental success rates remain limited because many computationally favorable assemblies fail to avoid kinetic traps, partial oligomerization, or non-native aggregation during real-world self-assembly conditions.

One major strategy addressing these limitations is the two-component cage design framework developed by Bale and colleagues [6]. In this approach, distinct oligomeric protein components contribute separate interface surfaces, thereby reducing unintended self-association and improving stoichiometric control over assembly formation. The method succeeds by minimizing the probability that individual components independently form incorrect oligomeric states before encountering their complementary partners. By distributing assembly interfaces across multiple components, the design effectively reduces off-pathway interactions while preserving global symmetry constraints. Furthermore, the use of independent trimeric and dimeric building blocks enables more precise control over geometric orientation and assembly kinetics. However, despite these strengths, the strategy does not completely eliminate kinetic mis-assembly because partial or malformed intermediates can still accumulate under non-ideal concentration conditions. In addition, two-component systems increase manufacturing complexity because balanced co-expression, purification, and stoichiometric mixing are required to achieve efficient assembly. The method also remains strongly dependent on accurate rigid-body interface design, which may not fully capture conformational flexibility in solution [6].

A second major research direction involves RFdiffusion-based backbone generation for interface design [9]. Unlike classical Rosetta docking methods that search within constrained rigid-body spaces, RFdiffusion employs generative diffusion modeling to create entirely new protein backbone geometries optimized for structural compatibility and interface formation. This approach succeeds by generating lower-strain, more designable interfaces that better satisfy geometric and energetic constraints simultaneously. The diffusion process allows exploration of broader conformational landscapes, enabling the discovery of interface topologies inaccessible to traditional deterministic design strategies. Experimental studies have demonstrated substantially improved success rates for designed assemblies generated using RFdiffusion compared with earlier methods [9]. However, the approach remains computationally intensive and still relies heavily on downstream filtering to identify experimentally realizable candidates. Moreover, while RFdiffusion improves interface geometry, it does not inherently solve the problem of kinetic accessibility during assembly. Generated backbones may remain vulnerable to transient aggregation states or unintended intermolecular contacts during experimental synthesis and purification.

Another influential solution is the ProteinMPNN sequence design framework combined with AlphaFold2 filtering [10]. ProteinMPNN uses graph neural networks to optimize amino acid sequences for predefined backbone structures while maintaining favorable packing, hydrogen bonding, and electrostatic compatibility. AlphaFold2-Multimer subsequently evaluates the likelihood that designed sequences will fold and assemble into the intended structures, with low RMSD and favorable predicted aligned error values serving as indicators of design quality [11]. This

 doi:10.20944/preprints202605.1980.v1

combined workflow succeeds because it dramatically reduces the number of experimentally tested failures, effectively functioning as a computational triage system that enriches for high-confidence candidates. The integration of structural prediction with sequence optimization represents a major advance in reducing empirical screening costs. Nevertheless, the approach remains fundamentally predictive rather than mechanistic. AlphaFold2 confidence metrics do not directly model assembly kinetics, intracellular expression behavior, or transient aggregation pathways. Consequently, sequences predicted to fold correctly may still fail experimentally because of solubility limitations, misfolding under cellular conditions, or non-equilibrium assembly dynamics.

Negative design of off-target interfaces constitutes another important strategy for improving assembly specificity [1]. In this framework, protein surfaces not intended for productive assembly are deliberately engineered to disfavor nonspecific interactions through the incorporation of charged or polar residues. The method succeeds by reducing hydrophobic surface patches that might otherwise promote aggregation or kinetic trapping. By explicitly destabilizing alternative low-energy interaction modes, negative design enhances the energetic funnel directing proteins toward the intended assembly state. This principle represents a critical conceptual advance because it recognizes that successful self-assembly depends not only on stabilizing desired interactions but also on destabilizing undesired ones. However, negative design introduces its own limitations. Excessive surface polarity may compromise overall protein stability or interfere with productive interface formation. Additionally, predicting all possible off-pathway interactions remains computationally difficult because the number of alternative assembly configurations grows exponentially with increasing system complexity. As a result, some kinetically favorable mis-assembly pathways may remain undetected during computational design.

Computational screening through molecular dynamics simulations has also been widely employed to evaluate interface stability and assembly robustness prior to experimental testing [12]. Molecular dynamics approaches succeed by providing atomistic insight into interface flexibility, solvent exposure, hydrogen bonding persistence, and thermodynamic stability over time. Simulations can identify weak interfaces prone to dissociation or conformational distortion under physiological conditions, thereby eliminating unstable candidates before synthesis. Moreover, molecular dynamics analyses provide mechanistic information regarding cage breathing motions, interface fluctuations, and strain accumulation that static structural predictions cannot capture. However, molecular dynamics simulations remain computationally expensive, particularly for megadalton-scale symmetric assemblies. Sampling limitations also constrain the ability to fully model long-timescale assembly pathways or rare kinetic events. Consequently, although molecular dynamics screening improves thermodynamic confidence, it does not completely solve the challenge of predicting experimentally successful assembly behavior in crowded biological environments.

The present study focuses on addressing the central unresolved challenge in the de novo design of symmetric protein cages: the simultaneous minimization of kinetic mis-assembly, off-pathway oligomerization, and low experimental success rates. The proposed solution integrates symmetry-aware scaffold selection, Rosetta-guided symmetric docking, RFdiffusion-based interface generation, ProteinMPNN sequence optimization, AlphaFold2-Multimer structural filtering, and experimentally compatible expression constraints into a unified design framework. Unlike prior approaches that optimize isolated aspects of assembly behavior independently, the present framework attempts to integrate geometric designability, energetic specificity, foldability prediction, and kinetic accessibility into a single end-to-end workflow. The overarching objective is to transform abstract symmetry targets into experimentally realizable protein sequences encoding robust self-assembly instructions for highly stable icosahedral nanocages with programmable geometry and minimized off-pathway interactions.

2. Methodology and Predicted Results

The methodology developed in this study is designed to address one of the principal unresolved barriers in de novo symmetric protein cage engineering: the inability to reliably convert

 doi:10.20944/preprints202605.1980.v1

computationally designed assemblies into experimentally validated nanocages with high fidelity and low rates of kinetic failure. The proposed framework integrates symmetry analysis, rigid-body docking, generative backbone modeling, neural-network sequence optimization, structural confidence filtering, and experimentally compatible expression validation into a unified multistage pipeline. The workflow is specifically optimized for the computational generation of a two-component icosahedral protein nanocage consisting of sixty total subunits and exhibiting an approximate external diameter of 25 nm.

The starting design conditions are intentionally constrained to maximize experimental realism and computational tractability. The target geometry is an icosahedral nanocage possessing global I symmetry and assembled from trimeric and dimeric oligomeric building blocks. Component A serves as a C3-symmetric trimeric scaffold, whereas Component B serves as a C2-symmetric dimeric scaffold. The use of a two-component system is intended to reduce spontaneous single-component aggregation and improve stoichiometric control during self-assembly [6]. Expression conditions are assumed to involve cytoplasmic production in *Escherichia coli* without disulfide-bond stabilization, thereby ensuring compatibility with high-throughput recombinant expression workflows. Purification is assumed to occur through His-tag affinity chromatography under native aqueous conditions. Candidate scaffold proteins are selected from the Protein Data Bank (PDB) based on their thermostability, solubility, helical bundle architecture, and experimentally validated oligomerization behavior [22]. The assembly environment is restricted to phosphate-buffered saline at physiological ionic strength and neutral pH in order to model biologically relevant aqueous assembly conditions. Computational filtering assumptions further specify that AlphaFold2-Multimer predictions must exhibit per-residue pLDDT values greater than 85 and interface predicted aligned error values below 10 Å to qualify as high-confidence assemblies [11].

Step 1: Definition of Target Symmetry and Identification of Scaffold Proteins

The first stage of the proposed workflow transforms the abstract geometric objective of constructing an icosahedral nanocage into a set of explicit symmetry constraints and experimentally tractable scaffold candidates. This step is necessary because successful self-assembly requires precise alignment between local oligomeric symmetry axes and the global symmetry architecture of the target polyhedron [1,6]. Icosahedral symmetry contains rotational axes corresponding to twofold, threefold, and fivefold operations, and therefore requires oligomeric building blocks capable of geometrically compatible spatial arrangement. The selected design strategy employs trimeric and dimeric protein scaffolds because C3 and C2 symmetry axes naturally map onto icosahedral rotational symmetry elements while minimizing interface redundancy.

Candidate scaffold proteins are identified from the RCSB Protein Data Bank using filters emphasizing small size, high solubility, thermal stability, and predominantly α-helical secondary structure. Helical bundle proteins are prioritized because their relatively rigid backbones and modular geometry simplify interface engineering and reduce conformational entropy penalties during assembly [1]. Experimentally characterized trimeric and dimeric coiled-coil proteins are specifically targeted because they provide preorganized oligomerization states that reduce the complexity of de novo symmetry generation. Structural symmetry analysis is subsequently performed using symmetry-space orientation tools to identify scaffold geometries compatible with icosahedral lattice placement [23].

This stage accomplishes several critical objectives simultaneously. First, it constrains the design search space to experimentally realistic protein architectures. Second, it establishes the geometric compatibility necessary for subsequent docking operations. Third, it reduces the probability of catastrophic assembly incompatibilities arising during later stages of interface optimization. The transformation achieved in this step converts an abstract symmetry target into a physically realizable collection of oligomeric scaffold candidates with known structural behavior and experimentally validated folding properties.

Step 2: Symmetric Docking to Generate Cage Configurations

 doi:10.20944/preprints202605.1980.v1

Following scaffold selection, the next stage involves rigid-body symmetric docking using the Rosetta macromolecular modeling suite [24]. The objective of this step is to generate candidate cage configurations in which trimeric and dimeric building blocks occupy geometrically compatible orientations consistent with global icosahedral symmetry. This stage is essential because even highly stable oligomeric scaffolds cannot form productive assemblies unless their relative rotational and translational relationships satisfy strict geometric constraints.

Rosetta symmetric docking protocols systematically explore rotational and translational degrees of freedom while enforcing predefined symmetry operations corresponding to the target icosahedral architecture [24]. During this process, the C3 trimeric building blocks are positioned along threefold symmetry axes, whereas the C2 dimeric components are aligned along twofold axes. Interface regions are evaluated according to shape complementarity, steric compatibility, hydrogen bonding potential, buried surface area, and overall energetic favorability. Candidate docked configurations exhibiting steric clashes or insufficient interface contact areas are eliminated.

This stage accomplishes the transformation of isolated oligomeric scaffolds into preliminary cage-like assemblies possessing coherent global symmetry. Importantly, the docking process also establishes the initial interface geometries that later serve as substrates for generative backbone optimization. Without this step, interface generation would occur in geometrically unconstrained conformational space, substantially increasing the likelihood of nonphysical assembly solutions. The use of explicit symmetry constraints further reduces computational complexity by limiting the search space to arrangements consistent with icosahedral rotational operations.

Step 3: De Novo Interface Design Using RFdiffusion

The third stage employs RFdiffusion to generate de novo interface backbone structures optimized for symmetric assembly formation [9]. This stage addresses a major limitation of classical rigid-body docking approaches, namely that docked interfaces often exhibit geometric strain, insufficient packing quality, or unrealistic backbone conformations. RFdiffusion overcomes these limitations by using diffusion-based generative modeling to create entirely new backbone segments capable of mediating stable intercomponent interactions.

The process begins with the docked cage configurations generated in Step 2. Interface regions between trimeric and dimeric components are designated as generative design targets. RFdiffusion iteratively denoises randomly initialized backbone conformations while conditioning the generation process on the spatial geometry of the target assembly [9]. Through this process, the model constructs novel interface helices, loops, and connecting motifs that maximize structural compatibility while minimizing strain energy.

This stage is necessary because successful self-assembly requires interfaces capable of simultaneously satisfying geometric specificity, energetic stability, and kinetic accessibility. Traditional rigid-body optimization frequently fails because naturally occurring scaffolds are not evolutionarily optimized for artificial polyhedral assembly. RFdiffusion effectively expands the accessible design landscape by generating backbone architectures specifically tailored for the intended cage geometry. Furthermore, because the generative process explores broad conformational distributions rather than local deterministic minima, the resulting interfaces are predicted to exhibit improved designability and reduced frustration.

The transformation achieved during this stage converts geometrically plausible but energetically incomplete docked assemblies into structurally coherent nanocage architectures possessing de novo engineered interfaces specifically optimized for symmetric assembly formation.

Step 4: Sequence Design Using ProteinMPNN

Once backbone geometries have been generated, amino acid sequences compatible with the designed structures must be identified. This objective is accomplished using ProteinMPNN, a graph neural network–based sequence design framework trained on large-scale protein structural datasets [10]. ProteinMPNN evaluates local and global geometric relationships within the designed backbone and predicts amino acid identities capable of stabilizing the structure while preserving assembly specificity.

 doi:10.20944/preprints202605.1980.v1

The generated backbone structures from RFdiffusion are provided as inputs to ProteinMPNN. Residues located at designed interfaces receive special attention because these positions determine assembly specificity, oligomerization kinetics, and interface stability. Hydrophobic packing interactions are optimized within buried regions, whereas solvent-exposed regions are enriched in polar or charged residues to reduce nonspecific aggregation. Sequence generation is performed iteratively, producing multiple candidate sequence sets for each backbone architecture.

This step is necessary because backbone geometry alone does not guarantee physical realizability. The designed sequences must support correct folding, maintain oligomeric stability, and preserve assembly specificity under experimental conditions. ProteinMPNN succeeds in this regard because it implicitly learns evolutionary constraints linking sequence patterns to structural stability. Additionally, the neural-network framework allows efficient exploration of sequence space without exhaustive combinatorial enumeration.

The transformation achieved in this stage converts abstract backbone structures into fully specified protein sequences encoding the structural and assembly information necessary for nanocage formation. The resulting sequences represent experimentally synthesizable genetic blueprints for self-assembling protein architectures.

Step 5: AF2-Multimer Computational Filtering

Following sequence generation, the proposed workflow employs AlphaFold2-Multimer computational filtering to evaluate foldability, interface confidence, and assembly accuracy prior to experimental synthesis [11]. This stage functions as a computational quality-control checkpoint designed to eliminate unstable or geometrically inconsistent designs before costly laboratory validation.

Each candidate sequence pair is evaluated using AlphaFold2-Multimer or ColabFold implementations configured for multimeric assembly prediction. Predicted structures are compared against the intended design models using root-mean-square deviation metrics, interface predicted aligned error values, and per-residue pLDDT (predicted Local Distance Difference Test) confidence scores. Only assemblies exhibiting RMSD values below 1.5 Å, pLDDT values above 85, and interface PAE (Predicted Aligned Error) values below 10 Å are advanced for experimental consideration.

This filtering stage is necessary because even highly optimized sequence designs may fail to adopt the intended conformations under realistic folding conditions. AlphaFold2-Multimer predictions provide an additional layer of structural validation by evaluating whether the designed sequences encode sufficiently strong energetic information to reproduce the intended assembly geometry. Importantly, this stage also helps identify interfaces vulnerable to conformational ambiguity or competing assembly states.

The transformation accomplished during this step narrows a large candidate pool into a high-confidence subset of experimentally tractable nanocage designs. By computationally eliminating likely failures, the workflow substantially reduces the empirical screening burden associated with protein nanomaterial development.

Step 6: Gene Synthesis, *E. coli* Expression, Purification, and Assembly Validation

The final stage of the workflow transitions from computational prediction to experimental realization. Synthetic genes encoding the optimized Component A and Component B sequences are codon optimized for *E. coli* expression and inserted into plasmid vectors containing inducible promoters and C-terminal His-tags [25]. Expression is assumed to occur within the cytoplasm of *E. coli* strains optimized for recombinant protein production.

Following expression, proteins are purified using nickel-affinity chromatography under native conditions. Purified components are subsequently mixed in stoichiometric ratios within phosphate-buffered saline containing 150 mM NaCl at pH 7.4. Assembly formation is monitored using size-exclusion chromatography, dynamic light scattering, native gel electrophoresis, and cryogenic electron microscopy. Successful cage formation is predicted to produce monodisperse particles approximately 25 nm in diameter consistent with the designed icosahedral architecture.

 doi:10.20944/preprints202605.1980.v1

This stage is necessary because computational predictions alone cannot fully capture intracellular folding behavior, translational kinetics, or solution-phase assembly dynamics. Experimental validation therefore serves as the ultimate determinant of design success. The use of *E. coli* expression systems additionally ensures compatibility with scalable recombinant production workflows.

The transformation achieved in this stage converts digitally encoded structural information into physically realized nanoscale protein assemblies capable of experimental characterization and downstream functional application.

Final Predicted Outcome: Transformation of Starting Materials

The final predicted outcome of the proposed workflow is the successful transformation of generic oligomeric scaffold proteins and an abstract icosahedral symmetry target into experimentally realizable protein sequences encoding a stable 60-subunit self-assembling nanocage. Through sequential symmetry analysis, rigid-body docking, RFdiffusion backbone generation, ProteinMPNN sequence optimization, and AlphaFold2-Multimer computational filtering, the workflow is predicted to generate highly specific interfaces capable of directing efficient assembly while minimizing off-pathway oligomerization and kinetic trapping.

The resulting nanocage is predicted to exhibit several key characteristics. First, the assembly should maintain geometric fidelity to the intended icosahedral architecture with an external diameter near 25 nm. Second, the two-component design is predicted to substantially reduce nonspecific self-association relative to single-component systems. Third, RFdiffusion-generated interfaces are expected to exhibit lower geometric strain and improved energetic complementarity compared with interfaces generated solely through rigid-body docking. Fourth, ProteinMPNN optimization combined with AlphaFold2-Multimer filtering is predicted to increase the proportion of experimentally realizable sequences by eliminating candidates with poor foldability or ambiguous interface geometry. Finally, the integrated workflow as a whole is expected to significantly improve experimental success rates relative to conventional protein cage design methodologies.

Collectively, the proposed framework represents a comprehensive transformation pipeline in which generic structural scaffolds and abstract symmetry specifications are progressively converted into programmable biomolecular assemblies encoding robust self-organization behavior at the nanoscale.

3. Discussion

The predicted outcome generated by the integrated workflow proposed in this study differs fundamentally from earlier symmetric protein cage design strategies because it combines geometric programming, generative backbone optimization, neural-network sequence engineering, structural confidence filtering, and experimentally compatible assembly constraints into a single continuous design architecture. Whereas previous approaches generally optimized isolated aspects of assembly formation independently, the present framework attempts to address the full sequence of events leading from abstract symmetry definition to experimentally realizable self-assembly. The resulting 60-subunit icosahedral nanocage is predicted to exhibit reduced kinetic frustration, minimized off-pathway oligomerization, enhanced foldability confidence, and improved assembly specificity relative to existing methodologies.

The first major comparison concerns the proposed framework versus the two-component cage design strategy developed by Bale and colleagues [6]. The earlier two-component method represented a major conceptual advance because it reduced spontaneous single-component aggregation through stoichiometric interface partitioning. By separating assembly interfaces across distinct oligomeric species, the approach improved control over self-assembly kinetics and reduced some forms of off-pathway association. However, the method remained strongly dependent on classical rigid-body interface engineering, which constrained interface geometry to conformations accessible through deterministic docking procedures. In contrast, the present workflow retains the advantages of two-component assembly while incorporating RFdiffusion-generated interfaces

 doi:10.20944/preprints202605.1980.v1

capable of exploring substantially broader conformational landscapes [9]. This addition is predicted to reduce geometric strain and improve interface complementarity beyond what rigid-body optimization alone can achieve. Furthermore, the integration of AlphaFold2-Multimer filtering introduces a predictive foldability validation layer absent from the original two-component strategy. Consequently, the proposed framework is predicted not only to preserve stoichiometric assembly control but also to improve structural realizability and reduce experimentally observed assembly failure rates.

The second comparison involves RFdiffusion-based backbone generation itself [9]. RFdiffusion represented a transformative development because it enabled the generative creation of novel protein interfaces unconstrained by existing structural templates. Experimental evidence demonstrated that diffusion-generated interfaces often exhibit superior designability and reduced conformational strain relative to earlier Rosetta-derived assemblies. Nevertheless, RFdiffusion alone does not constitute a complete assembly solution because successful backbone generation does not guarantee compatible sequence encoding, favorable expression behavior, or experimentally stable self-assembly. The present study extends the utility of RFdiffusion by embedding it within a broader multistage design pipeline that includes scaffold preselection, symmetry-aware docking, ProteinMPNN sequence optimization, and AlphaFold2-Multimer structural filtering. This integration is significant because it links backbone generation directly to downstream experimental realizability constraints. As a result, the proposed workflow is predicted to achieve greater assembly reliability than RFdiffusion operating in isolation. In particular, the sequential filtering process reduces the likelihood that geometrically attractive but experimentally unstable interfaces advance to synthesis stages.

The third comparison concerns the ProteinMPNN and AlphaFold2 filtering paradigm introduced by Dauparas and colleagues [10]. The ProteinMPNN framework substantially improved sequence optimization by learning structure–sequence compatibility relationships from large-scale protein datasets. Combined with AlphaFold2-based structural validation, the method dramatically reduced the number of low-confidence experimental candidates requiring laboratory testing [10,11]. However, the approach remains primarily sequence-centric and predictive rather than mechanistically assembly-oriented. ProteinMPNN optimizes amino acid identities for a given backbone, whereas AlphaFold2 evaluates structural plausibility, but neither method explicitly models the kinetic pathways through which self-assembly occurs. The present framework addresses this limitation by integrating sequence optimization into a larger assembly-aware architecture beginning with symmetry-compatible scaffold selection and continuing through diffusion-based interface generation. Because interface topology and assembly geometry are co-optimized before sequence design begins, the resulting sequences are predicted to encode more kinetically accessible assembly pathways. Additionally, the proposed workflow explicitly incorporates negative surface design principles and aqueous assembly constraints, thereby extending beyond purely structural prediction toward experimentally informed assembly engineering.

A fourth comparison may be drawn with the negative design framework developed in early computational nanomaterial engineering studies [1]. Negative design represented an essential conceptual advancement because it recognized that successful self-assembly requires suppression of alternative low-energy interaction pathways in addition to stabilization of intended interfaces. By introducing charged or polar residues on nonproductive surfaces, previous studies successfully reduced aggregation and nonspecific oligomerization [1]. Nevertheless, negative design alone cannot fully eliminate kinetic traps because many unintended interactions arise from transient conformational states or partially assembled intermediates that are difficult to predict computationally. The present framework incorporates negative design implicitly during both ProteinMPNN optimization and AlphaFold2 filtering while also addressing upstream interface geometry quality through RFdiffusion. This distinction is important because poorly optimized backbone geometries may create frustrated interaction landscapes that cannot be completely corrected through surface polarity engineering alone. Consequently, the integrated workflow is

 doi:10.20944/preprints202605.1980.v1

predicted to achieve more comprehensive suppression of off-pathway assembly by combining interface geometry optimization with sequence-level negative design principles.

The fifth comparison involves molecular dynamics (MD)–based computational screening approaches [12,26]. Molecular dynamics simulations provide valuable insight into thermodynamic stability, interface flexibility, solvent interactions, and conformational fluctuations. Earlier studies demonstrated that MD simulations can identify weak interfaces and unstable assemblies prior to experimental testing, thereby improving overall design reliability [12,26]. However, MD-based screening remains computationally expensive, particularly for megadalton-scale polyhedral assemblies containing dozens of subunits. Moreover, finite simulation timescales limit the ability to fully sample rare kinetic events or long-timescale assembly transitions. The present workflow differs by emphasizing predictive structural confidence and generative interface quality rather than exhaustive atomistic trajectory analysis. While molecular dynamics simulations remain useful for downstream refinement, the proposed framework seeks to reduce dependence on costly MD sampling by integrating machine-learning–based structural confidence metrics earlier in the design pipeline. This strategy is predicted to increase throughput while preserving high assembly fidelity. Additionally, because AlphaFold2-Multimer predictions provide ensemble-informed structural confidence estimates, the workflow may capture many instability signatures without requiring prohibitively expensive long-timescale simulations.

Collectively, these comparisons demonstrate that the present framework does not simply replace earlier methodologies but rather synthesizes their strongest elements into a unified design strategy specifically optimized for minimizing kinetic mis-assembly and improving experimental realization rates. The proposed solution integrates stoichiometric control from two-component systems, geometric flexibility from RFdiffusion, sequence optimization from ProteinMPNN, structural validation from AlphaFold2-Multimer, and aggregation suppression from negative design principles. The resulting workflow therefore represents a systems-level approach to programmable symmetric nanocage engineering rather than a collection of isolated optimization procedures. Most importantly, the framework explicitly attempts to bridge the persistent gap between computationally favorable structures and experimentally successful self-assembly, which remains one of the central unresolved challenges in de novo protein nanomaterial design.

4. Future Directions

Future research in symmetric protein cage engineering will likely focus on improving the dynamic adaptability, environmental responsiveness, and functional complexity of de novo designed nanocages. One important direction involves the development of stimulus-responsive assemblies capable of undergoing reversible structural transitions in response to pH, ionic strength, temperature, redox state, ligand binding, or light exposure. Such systems could enable programmable cargo release, conditional enzymatic activation, or environment-specific therapeutic delivery. Integrating conformational switching mechanisms into highly symmetric cage architectures remains difficult because local structural perturbations must propagate coherently across large multimeric assemblies without destabilizing the global architecture. Advances in diffusion-based generative modeling and allosteric network design may provide new solutions to this challenge.

Another major direction concerns the incorporation of functional internal environments within designed nanocages. Most current cage systems primarily emphasize structural assembly rather than active biochemical functionality. Future designs may incorporate catalytic centers, electron-transfer pathways, substrate channels, or molecular sorting mechanisms within cage interiors. Achieving this objective will require simultaneous optimization of global symmetry, local catalytic geometry, and molecular transport properties. The integration of enzyme engineering with programmable nanocage assembly could enable synthetic organelle systems capable of supporting highly organized metabolic pathways or artificial biosynthetic compartments.

The application of large-scale generative artificial intelligence models represents another promising research frontier. Current methods such as RFdiffusion and ProteinMPNN already

 doi:10.20944/preprints202605.1980.v1

demonstrate the power of machine learning in protein engineering, but future systems may integrate sequence generation, backbone generation, thermodynamic prediction, kinetic modeling, and evolutionary optimization within unified multimodal architectures. Such models could potentially simulate entire assembly trajectories rather than predicting only final structures. The ability to computationally model assembly kinetics directly would represent a major advance in overcoming the persistent problem of kinetic mis-assembly.

Further improvements in experimental validation technologies are also expected to accelerate the field. High-throughput cryogenic electron microscopy, single-particle analysis, native mass spectrometry, and microfluidic assembly screening platforms may substantially increase the speed at which candidate assemblies can be experimentally characterized. Coupling these experimental approaches with active-learning optimization algorithms could enable iterative closed-loop protein nanomaterial engineering workflows in which computational predictions are continuously refined using experimental feedback.

An additional future direction involves the design of asymmetric or quasi-symmetric hybrid assemblies capable of combining multiple functional modules within a single nanostructure. While highly symmetric cages provide substantial advantages in stability and manufacturability, biological systems frequently exploit controlled asymmetry to achieve directional transport, regulatory gating, or spatial compartmentalization. Future programmable nanocages may therefore integrate symmetric structural cores with asymmetric functional domains to achieve more sophisticated behaviors.

Finally, future research will likely emphasize biomedical translation and manufacturability. Many computationally designed assemblies remain difficult to produce at industrial scale because of solubility limitations, expression bottlenecks, or purification inefficiencies. Incorporating manufacturability constraints directly into early computational design stages may therefore become increasingly important. Likewise, improved understanding of immunogenicity, biodistribution, pharmacokinetics, and long-term biocompatibility will be essential for translating de novo protein nanocages into clinical applications.

5. Conclusions

The de novo design of symmetric protein cages with programmable geometry represents one of the most ambitious and rapidly advancing areas within computational protein engineering. Although major progress has been achieved through developments in symmetric docking, generative backbone modeling, neural-network sequence optimization, and machine-learning–based structural prediction, the reliable experimental realization of complex polyhedral nanocages remains constrained by kinetic mis-assembly, off-pathway oligomerization, and low assembly success rates.

This study presented a comprehensive technical analysis of symmetric protein cage design and critically evaluated five influential solution strategies currently used to address assembly fidelity challenges: two-component cage design, RFdiffusion-generated interfaces, ProteinMPNN sequence optimization with AlphaFold2 filtering, negative design of off-target interactions, and molecular dynamics–based computational screening. Each approach contributes important advances toward improving assembly precision, yet each also exhibits limitations when applied independently.

To address these limitations, the present study proposed an integrated multistage design framework combining symmetry-aware scaffold selection, Rosetta symmetric docking, RFdiffusion interface generation, ProteinMPNN sequence optimization, AlphaFold2-Multimer computational filtering, and experimentally compatible Escherichia coli expression constraints. The predicted outcome of this workflow is the generation of highly stable and experimentally realizable 60-subunit icosahedral protein nanocages exhibiting reduced kinetic frustration, minimized nonspecific oligomerization, and enhanced assembly specificity.

The proposed framework differs from earlier methodologies because it attempts to integrate geometric compatibility, energetic optimization, foldability prediction, and assembly accessibility into a unified systems-level design strategy. Rather than optimizing isolated assembly parameters

 doi:10.20944/preprints202605.1980.v1

independently, the workflow progressively transforms abstract symmetry objectives into experimentally tractable protein sequences encoding robust self-assembly behavior.

As computational structural biology, generative artificial intelligence, and experimental validation technologies continue to advance, the ability to engineer programmable nanoscale protein architectures with atomic precision is expected to expand substantially. The development of reliable symmetric protein cages may ultimately enable transformative applications in targeted therapeutics, vaccine engineering, synthetic biology, catalysis, diagnostics, molecular transport, and artificial cellular organization. The continued integration of machine learning, structural biophysics, and experimental protein engineering will therefore remain central to the future evolution of de novo self-assembling protein nanomaterials with programmable geometry.

References

1. King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., André, I., Gonen, T., Yeates, T. O., and Baker, D. "Computational design of self-assembling protein nanomaterials with atomic level accuracy," *Science*, vol. 336, no. 6085, pp. 1171–1174, 2012.
2. Yeates, T. O., Liu, Y., and Laniado, S. L. "The design of symmetric protein nanomaterials comes of age in theory and practice," *Current Opinion in Structural Biology*, vol. 39, pp. 134–143, 2016.
3. Uchida, M., McCoy, M. V., Fukuto, H., Yang, L., Yoshimura, P. M., Miettinen, T., LaFrance, B., Patterson, D. E., Schwarz, B., and Douglas, T. "Modular self-assembly of protein cage lattices for multistep catalysis," *ACS Nano*, vol. 12, no. 9, pp. 942–953, 2018.
4. Walls, A. C., Fiala, M., Schäfer, M., Wrenn, M., Pham, D., Murphy, A., Tse, M., Shehata, K., O'Connor, M., Chen, C., et al. "Elicitation of potent neutralizing antibody responses by designed protein nanoparticle vaccines," *Cell*, vol. 183, no. 5, pp. 1367–1382, 2020.
5. Baker, D. "What has de novo protein design taught us about protein folding and biophysics?" *Protein Science*, vol. 28, no. 4, pp. 678–683, 2019.
6. Bale, J. B., Gonen, S., Liu, Y., Sheffler, W., Ellis, D., Thomas, C., Cascio, D., Yeates, T. O., King, N. P., and Baker, D. "Accurate design of megadalton-scale two-component icosahedral protein complexes," *Science*, vol. 353, no. 6297, pp. 389–394, 2016.
7. Hsia, Y., Bale, J. B., Gonen, S., Shi, D., Sheffler, W., Fong, K. K., Nattermann, U., Xu, C., Huang, P. S., Ravichandran, R., et al. "Design of a hyperstable 60-subunit protein icosahedron," *Nature*, vol. 535, pp. 136–139, 2016.
8. Meador, R., et al. "A suite of designed protein cages using machine learning and protein fragment-based protocols," *Structure*, vol. 32, pp. 751–765, 2024.
9. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Pulavarti, S. A., Bortoli, L. B., Correia, A. P. L., Ovchinnikov, A. M., et al. "De novo design of protein structure and function with RFdiffusion," *Nature*, vol. 620, pp. 1089–1100, 2023.
10. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B., Courbet, A., de Haas, R., Bethel, N., et al. "Robust deep learning–based protein sequence design using ProteinMPNN," *Science*, vol. 378, no. 6615, pp. 49–56, 2022.
11. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. "Protein complex prediction with AlphaFold-Multimer," *bioRxiv*, 2021.
12. Bale, J. B., Park, R. U., Liu, Y., Gonen, S., Gonen, T., Cascio, D., King, N. P., Yeates, T. O., and Baker, D. "Structure of a designed tetrahedral protein assembly variant engineered to have improved soluble expression," *Protein Science*, vol. 24, no. 10, pp. 1695–1701, 2015.
13. Sutter, D., Boehringer, M., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M., and Stetter, M. "Structural basis of enzyme encapsulation into a bacterial nanocompartment," *Nature Structural & Molecular Biology*, vol. 15, pp. 939–947, 2008.
14. Yeates, T. O., Crowley, C. S., and Tanaka, S. "Bacterial microcompartment organelles: protein shell structure and evolution," *Annual Review of Biophysics*, vol. 39, pp. 185–205, 2010.
15. Kerfeld, D., and Melnicki, M. R. "Assembly, function and evolution of cyanobacterial carboxysomes," *Current Opinion in Plant Biology*, vol. 31, pp. 66–75, 2016.

 doi:10.20944/preprints202605.1980.v1

16. Uchida, M., Klem, D., Allen, S., Suci, M., Flenniken, T., Gillitzer, M., Varpness, Z., Douglas, T., and Young, M. "Targeting of cancer cells with ferrimagnetic ferritin cage nanoparticles," *Journal of the American Chemical Society*, vol. 129, no. 29, pp. 9102–9108, 2007.
17. Glasgow, J. A., and Tullman-Ercek, D. "Production and applications of engineered viral capsids," *Applied Microbiology and Biotechnology*, vol. 98, pp. 5847–5858, 2014.
18. Liu, Y., Huynh, D., and Yeates, T. O. "A 3.8 Å resolution cryoEM structure of a designed protein cage assembled by way of de novo–designed coiled coils," *Protein Science*, vol. 28, no. 10, pp. 1860–1867, 2019.
19. Guo, W., et al. "Modeling Viral Capsid Assembly: A Review of Computational Strategies and Applications," *Cells*, vol. 13, 2088, 2024.
20. Yang, W., Wang, S., Lee, G. R., et al. "The past, present and future of de novo protein design," *Nature*, vol. 652, pp. 1139–1152, 2026.
21. Padilla, J. E., Colovos, C., and Yeates, T. O. "Nanohedra: using symmetry to design self assembling protein cages, layers, crystals, and filaments," *Proceedings of the National Academy of Sciences USA*, vol. 98, no. 5, pp. 2217–2221, 2001.
22. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., and Bourne, P. "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.
23. Laniado, J., and Yeates, T. O. "A complete rule set for designing symmetry combination materials from protein molecules," *Proceedings of the National Academy of Sciences USA*, vol. 117, pp. 31817–31823, 2020.
24. Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P., Smith, C., Sheffler, W., et al. "ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules," *Methods in Enzymology*, vol. 487, pp. 545–574, 2011.
25. Gao, R., Zhang, X.-E., and Li, F. "Generation and characterization of self-assembled protein nanocages based on β-carboxysomes in Escherichia coli," *Acta Biochimica et Biophysica Sinica*, vol. 53, no. 7, pp. 943–949, 2021.
26. Shoemark, D. K., Ibarra, A. A., Ross, J. F., Beesley, J. L., Bray, H. E. V., Mosayebi, M., and Sessions, R. B. "The dynamical interplay between a megadalton peptide nanocage and solutes probed by microsecond atomistic MD; implications for design," *Physical Chemistry Chemical Physics*, vol. 21, no. 1, pp. 137–147, 2019.