PROTEAN: BEYOND DE NOVO GENERATION — CAN LLM AGENTS PERFORM TEXT-GUIDED PROTEIN STRUCTURE EDITING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative AI has achieved remarkable success in *de novo* protein design, yet protein engineering typically begins with existing, characterized proteins. The primary demand is targeted local editing, e.g., engineering a specific region to optimize affinity, alter immunogenicity, or improve stability, while strictly preserving core catalytic motifs and overall scaffold integrity. However, no standardized framework currently exists for carrying out and evaluating such edits. In current generative-design workflows, researchers often need to manually specify complex low-level parameters and rely on repetitive, trial-and-error sampling and evaluation across generative backends to search for viable conformations. While large language models (LLMs) offer an intuitive interface to express these goals using natural language, prevailing structure-generation models cannot directly interpret high-level human instructions. To bridge this gap, we introduce PROTEAN, a standardized benchmark for text-guided protein structural editing. PROTEAN comprises 380 sub-tasks spanning six intervention families and evaluates a two-stage pipeline in which an LLM translates natural-language goals into backend-specific geometric constraints and a generative model remodels the protein backbone. A two-tier evaluation separates geometric compliance, including target attainment and off-target scope fidelity, from biophysical realizability assessed by inverse folding and refolding consistency. Across five frontier LLMs paired with three backends, strict end-to-end biophysical success starts at just 7.0%. Ablation studies show that LLM constraint compilation is already highly accurate, whereas the main bottlenecks are backend-specific scope leakage and refolding failure. PROTEAN provides an open testbed for developing autonomous, multi-backend macromolecular design agents.

1 INTRODUCTION

Artificial intelligence has fundamentally transformed structural biology. Deep learning systems such as AlphaFold2 and AlphaFold3 have turned proteome-wide structure prediction into standard practice (Jumper et al., 2021; Abramson et al., 2024), while 3D generative architectures like RFdiffusion (Watson et al., 2023a), Chroma (Ingraham et al., 2023), and ESM3 (Hayes et al., 2025) have unlocked programmable *de novo* backbone generation from scratch. These generative frameworks, however, operate primarily in an unconstrained regime—synthesizing full proteins from stochastic noise or unconstrained boundary priors.

In real-world protein engineering, however, projects rarely start from scratch. Practitioners almost invariably build upon existing, evolutionarily vetted proteins to inherit their natural stability, solubility, and proven expression profiles—qualities that *de novo* designs frequently lack. Across both therapeutic development and industrial biocatalysis, the primary demand is *targeted structural editing*: remodeling a specific local region to optimize affinity, alter immunogenicity, or improve stability, while strictly preserving core catalytic motifs and overall scaffold integrity. Yet, executing targeted edits computationally remains an empirical ordeal. Because existing protein models provide limited localized control, often permitting unintended structural drift outside the target site, practitioners must manually configure low-level parameters and rely on repetitive, trial-and-error sampling across fragmented backends.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Large language models (LLMs) provide an intuitive interface to streamline this process, allowing engineers to express nuanced structural objectives in natural language. Yet, prevailing **3D generative engines cannot directly parse high-level human instructions**; they require specialized, low-level geometric constraints and heterogeneous parameter configurations. Bridging this gap generally demands an two-stage LLM-plus-engine collaborative paradigm: an upstream reasoning LLM translates abstract natural-language instructions into backend-specific geometric constraints, and a downstream generative backend executes the physical 3D backbone remodeling. While prior biology-oriented language models have addressed sequence-level editing or tool invocation for *de novo* design (Liu et al., 2023; Weng et al., 2025; Ghafarollahi & Buehler, 2024), no standardized benchmark currently exists to formalize, execute, and evaluate text-guided 3D backbone editing under strict conservation contracts across diverse generative backends.

To establish a principled foundation for this emerging problem, we introduce PROTEAN, the first standardized benchmark for text-guided protein structural editing. PROTEANestablishes 380 tasks spanning eight structural intervention families: secondary-structure shifting (helix and sheet enrichment), inter-domain distance modulation (approach and separation), reference-fragment mimicry, global shape control, backbone cyclization, and motif scaffolding. These tasks collectively evaluate an agent's ability to drive purposeful deformation while preserving non-target regions under a rigid geometric contract (C_α-RMSD ≤ 0.5 Å).

Crucially, PROTEANinstitutes a rigorous two-tier evaluation protocol that decouples geometric compliance from biophysical realizability: **Tier 1: Geometric Compliance** verifies macroscopic task execution directly on the generated backbone, measuring both target attainment (whether the instructed structural transformation succeeded) and off-target scope fidelity (whether unedited scaffolds strictly satisfied the conservation threshold); **Tier 2: Biophysical Realizability** subjects geometrically compliant backends to sequence redesign via ProteinMPNN (Dauparas et al., 2022b) and orthogonal refolding via ESMFold (Lin et al., 2023b), examining whether the computationally edited backbone can stably self-assemble in physical solution.

PROTEAN fundamentally redefines computational protein editing along two core dimensions. First, it transforms targeted structural editing from an *ad hoc*, trial-and-error craft into a formalized task governed by rigid geometric conservation contracts. Second, it establishes natural language as an expressive, high-level control interface, thereby democratizing macromolecular design by empowering researchers to direct complex, multi-backend 3D remodeling through scientific intent rather than fragile parameter hacking. (Fig. 1) Beyond its utility as an execution framework, PROTEAN provides the foundational diagnostic testbed required to rigorously interrogate and catalyze next-generation, physically realizable autonomous design agents.

Through comprehensive benchmarking across five frontier LLMs paired with three heterogeneous generative backends (BoltzGen (Stark et al., 2025), ODesign (Zhang et al., 2025), and RFdiffusion3 (Butcher et al., 2025)), our experiments uncover that autonomous structural editing is viable yet severely bottlenecked. Specifically, even the top-performing agent (Claude-Opus-5) attains only 46.0% joint geometric compliance ($G_1 \wedge G_3$, Table 1) with competitor baselines falling below 28%, while achieving end-to-end biophysical refolding self-consistency poses an even steeper sampling barrier (Table 2).

Systematic bottleneck analyses reveal that this performance collapse stems not from semantic misunderstanding, but from generator-specific failures, including pervasive non-local contact strain, severe scope leakage (collateral drift in designated invariant scaffolds), and persistent Pareto deadlocks during complex remodeling. This work delivers three primary contributions:

1. We formalize text-guided protein structural editing under hard conservation boundaries, establishing a two-stage collaborative paradigm that connects high-level natural language to low-level 3D generative backends across eight structural intervention families.
2. We introduce PROTEAN, comprising 380 verified editing sub-tasks across six intervention families and eight concrete tasks. It is paired with an evaluation harness that decouples macroscopic geometric compliance from microscopic biophysical realizability.
3. We provide extensive empirical evaluations of a set of paired agent-backend toolchains, isolating backend coordinate leakage and refolding collapse as the primary bottlenecks, thereby offering clear directions for developing autonomous macromolecular design agents.

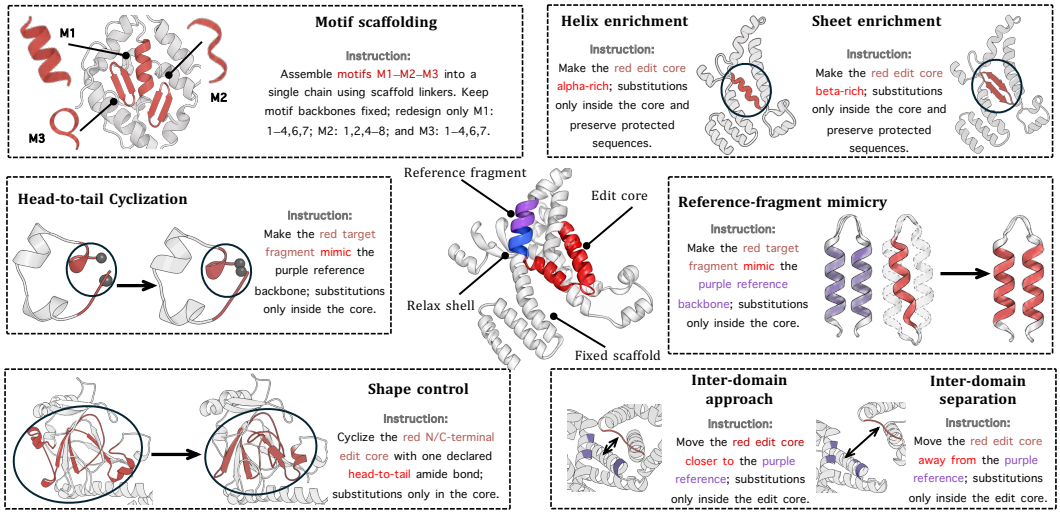


Figure 1: **Overview of the eight protein structure editing tasks in PROTEAN.** The benchmark covers motif scaffolding, helix enrichment, sheet enrichment, head-to-tail cyclization, reference-fragment mimicry, shape control, inter-domain approach, and inter-domain separation. Across the tasks, red denotes the editable core, blue denotes the surrounding relaxation shell, gray denotes the fixed scaffold, and purple denotes the reference fragment or region. Representative structures and natural-language instructions illustrate the editing objectives and the associated sequence and geometric constraints.

2 RELATED WORK

2.1 STRUCTURE GENERATION AND HETEROGENEOUS DESIGN BACKENDS

Recent structural generative models (Watson et al., 2023a; Ingraham et al., 2023; Hayes et al., 2025) have demonstrated remarkable capabilities in *de novo* backbone synthesis and motif scaffolding. However, these systems rely heavily on human experts manually constructing complex, backend-specific inputs. Moreover, current state-of-the-art generators exhibit fragmented, specialized operational domains rather than interchangeable capabilities: BoltzGen (Stark et al., 2025) natively handles topological head-to-tail cyclization through explicit covalent bond constraints; ODesign (Zhang et al., 2025) excels at partial diffusion and hotspot-conditioned inter-domain distance modulation; and RFdiffusion3 provides robust generalized backbone generation under spatial coordinate bounds. PROTEAN treats these heterogeneous engines as frozen tools, evaluating how effectively LLMs can autonomously select and compile their incompatible native specifications.

2.2 LANGUAGE-DRIVEN PROTEIN ENGINEERING

A growing body of work leverages natural language to steer macromolecular generation. Although existing language-driven methods (Liu et al., 2023; Weng et al., 2025; Wang et al., 2024; Lv et al., 2024) edit 1D protein sequences, text-guided 3D structure editing remains the core imperative for functional design. Operating solely on sequences neglects explicit geometric constraints, offers no spatial preservation for unedited regions, and fails to yield native 3D representations. On the other hand, scientific agent frameworks such as ScienceAgentBench (Chen et al., 2025), ChemCrow (Bran et al., 2024), and BioDesignBench (Kim & Romero, 2026) evaluate tool-use workflows, but predominantly target general coding or *de novo* binder generation pipelines.

2.3 PROTEIN DESIGN BENCHMARKS & DECOUPLED EVALUATION

Existing protein design benchmarks focus on isolated stages, such as representation learning (e.g., TAPE, PEER) (Rao et al., 2019b; Xu et al., 2022), fitness prediction (ProteinGym) (Notin et al., 2023a), or inverse folding (ProteinInvBench) (Gao et al., 2023). Consequently, they fail to diagnose the multi-tier failure modes of text-guided editing, often conflating geometric non-compliance with biophysical instability under a monolithic score. PROTEAN bridges this gap by introducing a decou-

pled evaluation protocol that rigorously disentangles 3D structural geometric fidelity from downstream biophysical plausibility, establishing a fine-grained diagnostic benchmark for autonomous protein engineering.

3 THE PROTEAN BENCHMARK

PROTEAN evaluates a model's ability to perform targeted structural interventions on an existing protein while strictly preserving its unedited scaffold, with natural language. The benchmark comprises eight task categories (380 sub-tasks): seven length-preserving structure-editing tasks (350 sub-tasks) and one motif-scaffolding task (30 sub-tasks) governed by a variable-length output contract.

3.1 TASK FORMULATION

A PROTEAN task is formally defined as a tuple:

$$\tau = \left(\underbrace{(s, X)}_{\text{source protein}} , \underbrace{\mathcal{M}}_{\text{edit scope}} , \underbrace{u}_{\text{instruction}} \right), \qquad s \in \mathcal{A}^L, \quad X \in \mathbb{R}^{L\times 4\times 3},$$

where s denotes the amino-acid sequence of length L over the canonical alphabet $\mathcal{A}$, X contains the backbone heavy-atom coordinates $(\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{O})$, u is a natural-language editing instruction, and $\mathcal{M} = (\mathcal{M}_{\mathrm{edit}}, \mathcal{M}_{\mathrm{relax}}, \mathcal{M}_{\mathrm{prot}})$ partitions the residues into editable, flexible (relaxation-allowed), and strictly protected regions.

For the seven length-preserving tasks, valid solutions must retain the source sequence length, chain topology, and canonical residue ordering (i.e., insertions, deletions, and non-canonical amino acids are strictly disallowed). Sequence substitutions are restricted exclusively to $\mathcal{M}_{\mathrm{edit}}$, whereas residues in $\mathcal{M}_{\mathrm{prot}}$ must preserve their original identities. Spatially, after anchored superposition onto the scaffold, all backbone atoms in $\mathcal{M}_{\mathrm{prot}}$ must exhibit an RMSD within 0.1 Å relative to the source.

By enforcing an explicit edit scope, sequence constraints, and deterministic success criteria, PROTEAN strictly demarcates targeted *structural editing* from unconstrained regeneration: achieving a design objective by reconstructing unperturbed regions violates the preservation contract. PROTEAN thus benchmarks whether an agent or model can reliably ground language-guided structural interventions into executable design pipelines under stringent structural preservation constraints. This complements existing benchmarks in structure prediction, inverse folding, and fitness prediction (Rao et al., 2019a; Notin et al., 2023b), text-guided sequence design (Liu et al., 2025), and constrained generative models such as RFdiffusion, BoltzGen, and ODesign.

3.2 TASK TAXONOMY AND CRITERIA

To systematically probe localized structural control, we formulate six intervention families across eight concrete tasks that span orthogonal biophysical dimensions, from local secondary structure and core packing to inter-domain kinematics, chain topology, and functional motif scaffolding. Evaluation is grounded in reproducible structural contracts: task success verifies that a candidate strictly satisfies the prescribed geometric constraints.

Secondary-structure enrichment (2 tasks). These tasks require remodeling the editable region to become either α-helix-rich (**50 sub-tasks**) or β-strand-rich (**50 sub-tasks**). Secondary structures are assigned via DSSP (Kabsch & Sander, 1983; Touw et al., 2015), and the objective is quantified by the propensity contrast $\Delta f = f_{\mathrm{target}} - f_{\mathrm{opposing}}$. Passing requires $\Delta f \geq 0.35$ for helix enrichment and $\Delta f \geq 0.10$ for sheet enrichment. These absolute thresholds evaluate whether a model can actively reshape local hydrogen-bonding networks and steric rigidity, rather than merely generating sequences statistically biased toward a particular fold.

Inter-domain distance modulation (2 tasks). These tasks evaluate spatial repositioning by moving an editable structural element either toward (*approach*, **50 sub-tasks**) or away from (*separation*, **50 sub-tasks**) a protected reference patch. In a shared global frame aligned to the source scaffold,

success requires a median distance change of at least 1.0 Å (decreasing for approach, increasing for separation; see Table 3). This abstracts conformational switching—such as allosteric domain closure or relief of autoinhibition (Changeux & Edelstein, 2005)—into a localized hinge- or loop-remodeling problem under strict contextual preservation.

Reference-fragment mimicry (1 task). Given a target backbone fragment and explicit residue mappings, the model must remodel the editable region to reproduce the prescribed geometry (**50 sub-tasks**). A design passes if the mapped backbone RMSD is ≤ 2.0 Å, covering $\geq 80\%$ of the designated alignment with at least three aligned residues. This task benchmarks structural transplantation, formalizing practical workflows such as CDR loop grafting (Jones et al., 1986), catalytic-site transfer, and epitope presentation (Ofek et al., 2010) within a fixed scaffold.

Core burial (1 task). Starting from an exposed 12-residue window, the model must pack the segment into a more buried and compact microenvironment (**50 sub-tasks**). Passing criteria require the median relative solvent accessibility (RSA, normalized via SASA (Tien et al., 2013)) to decrease by at least 0.10 to an absolute $\mathrm{RSA} \leq 0.40$, alongside bounded variations in the local radius of gyration. This reflects hydrophobic core repacking—the primary thermodynamic driver of protein folding and stability (Chothia, 1984; Rose et al., 1985).

Head-to-tail cyclization (1 task). Cyclization tasks define the protein termini as editable and require linking the N- and C-termini through a peptide bond (**50 sub-tasks**). A valid edit must form exactly one new covalent bond with a terminal backbone C–N distance between 1.20 and 1.50 Å. Unlike purely geometric alterations, cyclization modifies chain topology; it evaluates whether a model can perform topological closure to enhance proteolytic resistance or preorganize binding conformations (Rettie et al., 2025a;b) while maintaining scaffold integrity.

Motif scaffolding (1 task). Operating under a variable-length contract, motif scaffolding constructs a new single-chain scaffold around one or more disconnected functional segments (e.g., catalytic motifs, metal-binding sites), this task has 30 sub-tasks. Following MotifBench (Zheng et al., 2025b), success requires motif backbone preservation ($\mathrm{RMSD} \leq 1.0$ Å) and computational self-consistency: at least one inverse-folded sequence must refold into the designed structure. This serves as the structural dual of targeted editing: holding the functional geometry invariant while rebuilding the entire structural carrier (Watson et al., 2023b).

3.3 Benchmark Curation Pipeline

As illustrated in Figure 2, PROTEAN instances are constructed via a three-stage curation funnel:

Data curation. Starting from public protein structures PDB (Berman et al., 2000), we build a high-quality, non-redundant source pool. Sequence deduplication and family-based clustering prevent homologous structures from entering the benchmark as spurious independent instances (detailed filtering thresholds are provided in Appendix B.1.1 and Appendix B.1.3). The motif-scaffolding task is adopted directly from MotifBench (Zheng et al., 2025b).

Task pre-classification. We then evaluate each curated structure's suitability across edit families. By coupling LLM reasoning with structural analysis tools, this stage assigns candidate edit types and identifies the requisite structural context (e.g., editable vs. protected regions, reference inputs). This tailors general proteins into well-defined, type-specific editing problems rather than treating them interchangeably across tasks.

Feasibility check. Finally, candidate problems are filtered against the predefined constraints of their assigned edit type. Only validated source–task pairs are retained. The resulting PROTEAN benchmark comprises 380 equal-length editing tasks (50 for each of the seven edit types and 30 motif-scaffolding cases). To minimize potential protein-data leakage—particularly the possibility that an LLM may have encountered related protein information during pretraining—we applied a protein-blinding procedure (see Appendix B.1.4).

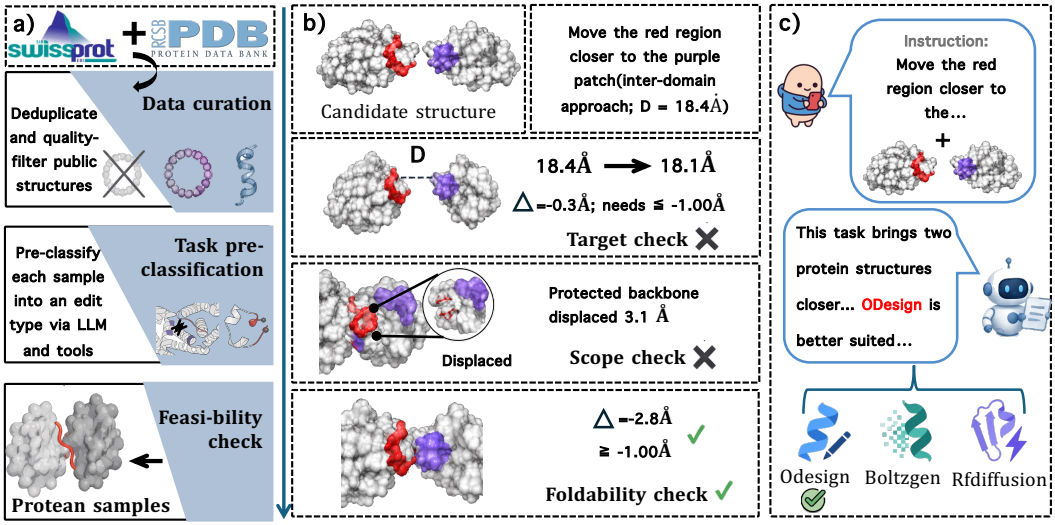


Figure 2: **Overview of the PROTEAN benchmark construction and model-routing pipeline. (a)** Protein structures are collected from Swiss-Prot and the PDB, deduplicated, quality-filtered, pre-classified by edit type, and screened for feasibility. **(b)** Candidate edits are validated through target-achievement, scope compliance, and refolding self-consistency checks to retain feasible samples. **(c)** Given a protein structure and a natural-language instruction, the model interprets the editing task and selects an appropriate backend, such as ODesign, BoltzGen, or RFdiffusion.

3.4 MULTI-GATED EVALUATION PROTOCOL

We evaluate protein editing across three complementary dimensions, formalized as the corresponding validation gates:

Target attainment (G_1). Assesses whether the intended structural modification is achieved according to type-specific criteria (Table 3). All evaluation thresholds are frozen prior to model inference. For example, an inter-domain approach task is claimed to pass the gate only when its source-referenced median distance decreases by at least 1.0 Å

Refolding self-consistency (G_2). Verifies the biophysical plausibility of the edited backbone via a fixed redesign-and-refolding loop. For length-preserving tasks, the editable region is redesigned with ProteinMPNN (Dauparas et al., 2022a) while preserving the protected context. Four candidate sequences are folded using ESMFold (Lin et al., 2023a); passing requires at least one sequence to meet frozen whole-chain and protected-context RMSD thresholds. This serves as a finite-budget proxy, meaning failure denotes a sampling miss rather than definitive biological incompatibility. (Cyclization is evaluated via a separate protocol.)

Scope compliance (G_3). Ensures the output is structurally sound and strictly constrained within the declared edit boundaries. Protected sequence identities, residue mapping, and backbone geometry must adhere to benchmark specifications. Candidates that fulfill target criteria by altering off-target regions are deemed invalid.

Overall success requires passing all applicable gates. Furthermore, execution outcomes (e.g., malformed outputs, backend failures) are reported separately to avoid conflating runtime anomalies with algorithmic editing failures (see Appendix B.8 for full scoring details).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate five frontier large language models: Claude-Opus-5 (Anthropic, 2025), GPT-5.6-Sol (OpenAI, 2025), DeepSeek-V4-Pro (Liu et al., 2024), Qwen3.8Max (Yang et al., 2024), and

Table 1: Per-task gate-wise pass rates across 7 in-scaffold editing tasks (pass@16, best backend). G_1: target attainment, G_3: scope compliance, $\wedge$: joint $G_1 \wedge G_3$. Thresholds: Approach/Separation ($\Delta\text{dist} \geq 1.0$ Å), Mimicry ($\text{RMSD} \leq 2.0$ Å, $\text{cov} \geq 80\%$), Core burial ($\Delta\text{RSA} \geq 0.10, \text{RSA} \leq 0.40$), Cyclization (C–N $\in [1.2, 1.5]$ Å). Red and blue denote highest and second-highest scores per task.

	DeepSeek-V4-Pro			Qwen3.8Max			Claude-Opus-5			GPT-5.6-Sol			Kimi-K3		
Task / Family	G_1	G_3	$\wedge$	G_1	G_3	$\wedge$	G_1	G_3	$\wedge$	G_1	G_3	$\wedge$	G_1	G_3	$\wedge$
Secondary-structure enrichment															
α-Helix	86	100	66.0	88	100	58.0	98	100	90.0	96	100	84.0	88	100	76.0
β-Sheet	62	94	44.0	60	94	36.0	84	94	60.0	82	94	58.0	60	94	40.0
Inter-domain distance modulation															
Approach	36	100	18.0	38	100	12.0	56	100	36.0	46	100	26.0	48	100	14.0
Separation	46	100	20.0	44	100	12.0	48	100	34.0	52	100	28.0	50	100	18.0
Localized structural transplantation & repacking															
Fragment mimicry	18	78	16.0	18	78	22.0	76	78	56.0	58	78	38.0	18	78	14.0
Core burial	34	96	30.0	36	96	28.0	66	96	46.0	50	96	34.0	34	96	20.0
Chain topological alteration															
Cyclization	100	100	0.0	100	100	0.0	100	100	0.0	100	100	0.0	100	100	0.0
Overall	54.6	95.4	27.7	54.9	95.4	24.0	75.4	95.4	46.0	69.1	95.4	38.3	56.9	95.4	26.0

Kimi-K3 (Moonshot AI, 2025), they act as autonomous constraint compilers across three structurally heterogeneous generation engines: BoltzGen, ODesign, and RFdiffusion3. Evaluations are conducted on the standardized PROTEAN suite curated in Section 3, comprising 380 tasks across the two complementary paradigms: 350 length-preserving editing tasks (50 instances across each of the seven structural intervention types) and 30 de novo motif-scaffolding cases, evaluated strictly under the formulation and criteria established in Sections 3.1 and 3.2.

Evaluation Protocol. We use pass@K as the evaluation metric. Furthermore, we tailor evaluation criteria to the distinct physical nature of each task:

- **Structural Editing ($\mathcal{G}_{\text{edit}} = G_1 \wedge G_3$):** Because equal-length editing operates on native protein contexts, the central challenge is executing local deformations without corrupting global structural integrity. We thus assess task success via *Joint Geometric Compliance*, requiring designs to concurrently fulfill target modifications (*Target Attainment*, G_1) and respect off-target boundary constraints (*Scope Compliance*, G_3). Refolding self-consistency (G_2) is decoupled and analyzed as a downstream sampling bottleneck in Appendix B.10.
- **Motif Scaffolding ($\mathcal{G}_{\text{motif}} = G_1 \wedge G_2 \wedge G_3$):** In contrast, motif scaffolding generates backbones *de novo*. Lacking native scaffolds to ensure refolding self-consistency, purely geometric criteria risk admitting physically unrealizable artifacts. Adhering to community benchmarks (Zheng et al., 2025a), we mandate orthogonal biophysical verification for candidate validity: backbones must preserve motif geometry ($G_1 \wedge G_3$) and produce at least one sequence via ProteinMPNN that refolds via ESMFold with whole-chain C_α-RMSD < 2.0 Å and motif C_α-RMSD < 1.0 Å (G_2).

4.2 Structural Geometric Compliance Across Task Families

Macro Stratification and the Joint-Gate Imperative. As detailed in Table 1, evaluating structural editing across frontier reasoning agents reveals distinct capability tiers under the joint compliance criterion ($G_1 \wedge G_3$ at pass@16). Claude-Opus-5 defines the benchmark frontier (46.0% overall), followed by GPT-5.6-Sol (38.3%), while remaining models cluster substantially lower (24.0–27.7%). Crucially, isolated target attainment (G_1) or boundary compliance (G_3) paints a deceptively optimistic picture: marginal G_3 rates saturate across all architectures due to pass@16 sampling diversity. The actual spatial reasoning bottleneck appears exclusively under the joint gate ($G_1 \wedge G_3$), which penalizes agents that satisfy design goals at the expense of unedited scaffold integrity within the *identical* candidate (Appendix B.9).

Table 2: Motif scaffolding performance comparing frontier model ceilings against competitor baselines across pipelines (ODesign, Boltzgen, RFdiffusion3). Values denote success rates (%) under strict biophysical refolding self-consistency ($G_1 \wedge G_2 \wedge G_3$). Red and blue highlights indicate the highest and second-highest scores per column, respectively.

Model	s@1	s@2	s@4	s@8	s@16
Claude-Opus-5	13.0	20.0	23.0	40.0	47.0
GPT-5.6-Sol	17.0	23.0	27.0	40.0	43.0
DeepSeek-V4-Pro	13.0	17.0	20.0	27.0	33.0
Qwen3.8Max	10.0	20.0	20.0	23.0	23.0
Kimi-K3	7.0	13.0	20.0	23.0	27.0

Contact Order Governs Secondary-Structure Remodeling. Performance on secondary-structure enrichment is fundamentally bounded by physical contact order. Helix design consistently outperforms sheet design across all models by 20–30 percentage points. Because α-helices depend on strictly local, contiguous hydrogen bonds ($i \rightarrow i + 4$), generative backends nucleate them readily without disrupting tertiary topology. Conversely, β-sheet enrichment requires coordinating non-local strand pairing and registry with the preserved framework. Generative backends repeatedly deform flanking conserved residues to satisfy inter-strand geometries, triggering pervasive scope violations.

Reference Mimicry as the Benchmark Discriminator. Fragment mimicry exhibits the steepest performance dispersion, spanning a four-fold divergence between Claude-Opus-5 (56.0%) and Kimi-K3 (14.0%). Mimicry represents the benchmark's most demanding test of spatial reasoning, requiring models to simultaneously solve discrete residue-index mapping and continuous rigid-body superposition ($\mathrm{RMSD} \leq 2.0$ Å over $\geq 80\%$ coverage) under invariant boundary bounds. Sub-frontier models predominantly fail at instruction translation, either emitting misaligned residue mappings or proposing structurally incompatible grafts whose insertion strain shatters contextual tolerances.

Mechanical Strain in Relational and Compaction Edits. Inter-domain modulation (Approach/Separation) and Shape control (Core Burial) both expose the mechanical tension between metric optimization and coordinate invariance. In relational modulation, baseline models frequently command rigid-body translations without designing compensating flexibility into transition hinges, transmitting excessive torque into the protected core. Similarly, driving core burial via hydrophobic compaction ($\Delta\mathrm{RSA} \geq 0.10$ to ≤ 0.40) inherently exerts compressive forces on the backbone, creating structural friction that severely caps joint compliance across all architectures.

The Pareto Deadlock of Topological Cyclization. Head-to-tail cyclization uncovers an absolute Pareto incompatibility: while all models reach 100% pass rates on marginal G_1 (forming the peptide bond) and G_3 (boundary preservation), joint compliance collapses completely to **0.0%** ($G_1 \wedge G_3$). Under rigid Euclidean bounds, these objectives are mutually exclusive: closing the separated termini (1.20–1.50 Å) inevitably causes global backbone relaxation that breaches G_3, whereas strictly pinning scaffold coordinates immobilizes the termini (G_1 failure). This universal 0.0% impasse demonstrates that cyclic topological edits demand concerted multi-scale deformation planning rather than isolated localized re-sampling (Appendix B.6).

4.3 Functional Motif Scaffolding and Biophysical Realizability

The Refolding Imperative in De Novo Carrier Design. While fixed-scaffold editing leverages evolutionary backbones that anchor global stability, motif scaffolding requires generating entirely new single-chain carriers around discontinuous functional segments. In this variable-length regime, generative engines easily hallucinate backbones that satisfy static spatial alignments with the motif but represent thermodynamically unviable conformations. Following structural biology standards (Zheng et al., 2025b), Table 2 mandates orthogonal computational refolding verification (G_2, via sequence redesign with ProteinMPNN and forward folding with ESMFold). A design is considered successful only under joint validity ($G_1 \wedge G_2 \wedge G_3$), filtering out non-physical hallucinations.

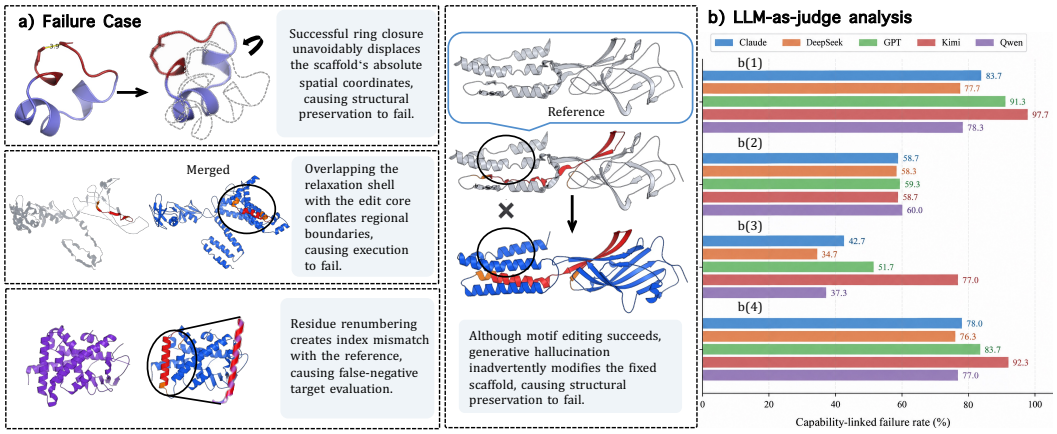


Figure 3: **Failure mode analysis and LLM-as-judge capability evaluation in ODesign. (a) Representative failure cases in protein editing.** Detailed mechanistic breakdowns of structural and operational failures, including absolute coordinate displacement during ring closure, scaffold modifications induced by generative hallucination, regional boundary conflation from overlapping relaxation shells, and evaluation mismatches caused by residue renumbering. **(b) Capability-linked failure profiling via LLM-as-judge.** Quantitative evaluation across leading LLMs categorized into four distinct failure modes: *(1) Edit-intent translation failure*, *(2) Spatial-relation planning failure*, *(3) Structural-boundary preservation failure*, and *(4) Multi-constraint coordination failure*. Unsupported tasks and out-of-memory (OOM) instances are excluded; percentages indicate capability-linked failure rates and may overlap across categories.

Frontier Ceilings and Linker Arithmetic. Under the full candidate budget (s@16), Claude-Opus-5 and GPT-5.6-Sol establish the generative frontier (43.0–47.0% self-consistent success), substantially outpacing competitor baselines (23.0–33.0%). In motif scaffolding, the reasoning agent must deduce the length partitioning of inter-motif linkers while satisfying global sequence constraints. Frontier models excel at inferring linker lengths that naturally accommodate canonical secondary-structure elements, thereby relieving torsional strain. In contrast, baseline models frequently prescribe over-compacted or excessively extended linkers that induce severe conformational tension, ultimately precipitating refolding collapse.

Further implementation details in motif scaffolding Because of the different nature of this task, we apply some distinct way to implement. We shall show the implementation details and further discussion in Appendix B.11

4.4 Mechanistic Failure Mode Profiling via LLM-as-Judge

To diagnose the bottlenecks behind failed designs, we dissect candidate breakdowns through mechanistic case analyses and LLM-as-judge capability profiling across four distinct failure modes (Fig. 3). Qualitative inspection confirms that errors stem from physical incompatibilities, such as ring-closure coordinate displacement, regional boundary conflation, and generative hallucinations mutating conserved scaffolds (Fig. 3a). Systematic evaluation reveals that while spatial planning fails uniformly across models ($\approx 60\%$), multi-constraint coordination failure (76.3%–92.3%) and intent translation failure remain the primary discriminators separating frontier models from baselines (Fig. 3b).

5 Conclusion

PROTEANestablishes a rigorous foundation for measuring autonomous macromolecular editing under strict physical constraints. By formalizing explicit boundary preservation contracts, source-isolated task splits, and multi-candidate evaluation, our work provides an operational benchmark to guide the development of next-generation biomolecular AI agents.

Reproducibility Statement

All task definitions, canonical source coordinates, and frozen evaluation thresholds are standardized within the benchmark framework. Candidate alignments follow deterministic Kabsch superposition, sequence redesign utilizes fixed seeds via ProteinMPNN, and structural refolding is verified using ESMFold. Standard pass@k metrics across candidate budgets ($k \in \{1, 2, 4, 8, 16\}$) evaluate empirical rollout success rates under uniform conditions. Detailed backend adapter schemas, compilation prompts, and evaluation procedures are documented in Appendix B.1–B.8.

Ethics Statement

This work uses publicly available datasets, benchmark materials, or interaction records in accordance with their respective licenses and terms of use. We do not intentionally collect personally identifiable information or access private intermediate system states beyond what is necessary for evaluation. Any released data, logs, or trace files should be limited to the information required for reproducibility and analysis, and residual identifiers should be removed where applicable.

AI Use Statement

The language models evaluated in this study were accessed via standard API interfaces under fixed system prompts and deterministic decoding settings. Generative AI tools were also used for textual refinement and formatting assistance during manuscript preparation. All experimental protocols, structural scoring pipelines, and analytical evaluations were implemented and verified by the authors.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Jason Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024.

Anthropic. The Claude 3.5 and beyond model family. `https://www.anthropic.com/claude`, 2025.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. ChemCrow.

Nathan Butcher, Joseph L. Watson, Nathaniel R. Bennett, Jueun Lee, Brian Sumida, Frank Yang, Brian Coventry, and David Baker. De novo design of all-atom biomolecular interactions with RFdiffusion3. *bioRxiv preprint*, 2025. doi: 10.1101/2025.09.18.676967.

Jean-Pierre Changeux and Stuart J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308(5727):1424–1428, 2005. doi: 10.1126/science.1108595.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Ziao Lu, et al. ScienceAgentBench: Toward rigorous assessment of language agents

for data-driven scientific discovery. In *International Conference on Learning Representations*, 2025.

Cyrus Chothia. Principles that determine the structure of proteins. *Annual Review of Biochemistry*, 53:537–572, 1984. doi: 10.1146/annurev.bi.53.070184.002541.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Nathaniel Bethel, Po-Ssu Jason Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Huey Nguyen, Alex Kang, Banumathi Sankaran, David Baker, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022a. doi: 10.1126/science.add2187.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Robert J. de Haas, Neville Bethel, Paul J. Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Fátima Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022b.

Zhangyang Gao, Cheng Tan, and Stan Z. Li. ProteinInvBench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. In *Advances in Neural Information Processing Systems*, 2023.

Alireza Ghafarollahi and Markus J. Buehler. ProtAgents: Protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery*, 2024.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Rafael S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Marinka Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Emilien R. Van Vlack, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Andrew R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Peter T. Jones, Paul H. Dear, Jefferson Foote, Michael S. Neuberger, and Greg Winter. Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature*, 321(6069):522–525, 1986. doi: 10.1038/321522a0.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. doi: 10.1002/bip.360221211.

S. Kim and P. A. Romero. BioDesignBench: Evaluating language agents on expert-curated biomolecular design workflows. *arXiv preprint*, 2026. Bibliographic details to be finalized.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023a. doi: 10.1126/science.ade2574.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023b.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Chen, Chengqi Deng, Chong Chen, Chunjing Zhao, Wenfeng Liang, et al. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Shengchao Liu, Yuxuan Li, Zhuoxinran Li, Anthony Gitter, and Jian Tang. A text-guided protein design framework. In *Advances in Neural Information Processing Systems*, 2023. ProteinDT.

Shengchao Liu, Yanjing Li, Zhuoxinran Li, Anthony Gitter, Yutao Zhu, Jiarui Lu, Zhao Xu, Weili Nie, Arvind Ramanathan, Chaowei Xiao, Jian Tang, Hongyu Guo, and Anima Anandkumar. A text-guided protein design framework. *Nature Machine Intelligence*, 7:580–591, 2025. doi: 10.1038/s42256-025-01011-z.

Liuzhenghao Lv et al. ProLLaMA: A protein large language model for multi-task protein language processing. *arXiv preprint*, 2024.

Moonshot AI. Kimi and Moonshot architecture series. `https://www.moonshot.ai`, 2025.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Wektoro, Jonathan Frazer, Yarin Gal, and Debora Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, 2023a.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379, 2023b. doi: 10.52202/075280-2810. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/cac723e5ff29f65e3fcbb0739ae91bee-Abstract-Datasets_and_Benchmarks.html`.

Gilad Ofek, F. Javier Guenaga, William R. Schief, Jeff Skinner, David Baker, Richard Wyatt, and Peter D. Kwong. Elicitation of structure-specific antibodies by epitope scaffolds. *Proceedings of the National Academy of Sciences*, 107(42):17880–17887, 2010. doi: 10.1073/pnas.1004728107.

OpenAI. GPT-4 and frontier generative architecture series. `https://openai.com/research`, 2025.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9689–9701, 2019a. URL `https://proceedings.neurips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html`.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 2019b.

Stephen A. Rettie, Katelyn V. Campbell, Asim K. Bera, Alex Kang, Simon Kozlov, Yensi Flores Bueso, Joshmyn De La Cruz, Maggie Ahlrichs, Suna Cheng, Stacey R. Gerben, Mila Lamb, Analisa Murray, Victor Adebomi, Guangfeng Zhou, Frank DiMaio, Sergey Ovchinnikov, and Gaurav Bhardwaj. Cyclic peptide structure prediction and design using AlphaFold2. *Nature Communications*, 16:4730, 2025a. doi: 10.1038/s41467-025-59940-7.

Stephen A. Rettie, David Juergens, Victor Adebomi, Yensi Flores Bueso, Qinqin Zhao, Alexandria N. Leveille, Andi Liu, Asim K. Bera, Joana A. Wilms, Alina Üffing, Alex Kang, Evans Brackenbrough, Mila Lamb, Stacey R. Gerben, Analisa Murray, Paul M. Levine, Maika Schneider, Vibha Vasireddy, Sergey Ovchinnikov, Oliver H. Weiergräber, Dieter Willbold, Joshua A. Kritzer, Joseph D. Mougous, David Baker, Frank DiMaio, and Gaurav Bhardwaj. Accurate de novo design of high-affinity protein-binding macrocycles using deep learning. *Nature Chemical Biology*, 21(12):1948–1956, 2025b. doi: 10.1038/s41589-025-01929-w.

George D. Rose, Allan R. Geselowitz, George J. Lesser, Richard H. Lee, and Mark H. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985. doi: 10.1126/science.4023714.

Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward universal binder design. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689494. Preprint.

Matthew Z. Tien, Austin G. Meyer, Dariya K. Sydykova, Stephanie J. Spielman, and Claus O. Wilke. Maximum allowed solvent accessibilites of residues in proteins. *PLOS ONE*, 8(11):e80635, 2013. doi: 10.1371/journal.pone.0080635.

Wouter G. Touw, Coos Baakman, Jon Black, Tim A. H. te Beek, Elmar Krieger, Robbie P. Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368, 2015. doi: 10.1093/nar/gku1028.

Zeyuan Wang et al. InstructProtein: Aligning human and protein language via knowledge instruction. *arXiv preprint*, 2024.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Eric Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023a.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023b. doi: 10.1038/s41586-023-06415-8.

Yunjia Weng et al. DisProtEdit: Improving diffusion-based protein generation via disentangled textual guidance. *arXiv preprint*, 2025. Bibliographic details to be finalized.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding. In *Advances in Neural Information Processing Systems*, 2022.

An Yang, Baosong Yang, Binyuan Zhang, Bofei Hui, Bo Zheng, Bowen Yu, Dayi Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Odin Zhang, Xujun Zhang, Haitao Lin, Cheng Tan, Qinghan Wang, Yuanle Mo, Qiantai Feng, Gang Du, Yuntao Yu, Zichang Jin, Ziyi You, Peicong Lin, Yijie Zhang, Yuyang Tao, Shicheng Chen, Jack Xiaoyu Chen, Chenqing Hua, Weibo Zhao, Runze Ma, Yunpeng Xia, Kejun Ying, Jun Li, Yundian Zeng, Lijun Lang, Peichen Pan, Hanqun Cao, Zihao Song, Bo Qiang, Jiaqi Wang,

Pengfei Ji, Lei Bai, Jian Zhang, Chang-Yu Hsieh, Pheng Ann Heng, Siqi Sun, Tingjun Hou, and Shuangjia Zheng. ODesign: A world model for biomolecular interaction design. *arXiv preprint arXiv:2510.22304*, 2025.

Zaixiang Zheng et al. MotifBench: A comprehensive benchmark for functional motif scaffolding. *bioRxiv preprint*, 2025a.

Zhuoqi Zheng, Bo Zhang, Kieran Didi, Kevin K. Yang, Jason Yim, Joseph L. Watson, Hai-Feng Chen, and Brian L. Trippe. MotifBench: A standardized protein design benchmark for motif-scaffolding problems, 2025b.

Table 3: The eight PROTEAN edit types. Thresholds are frozen before model evaluation. The final column gives the biological motivation for each structural intervention rather than an experimentally validated outcome.

Edit type	Sub-tasks	Additional input	Primary criterion	Biological motivation
Helix enrichment	50	–	secondary structure propensity contrast ≥ 0.35	local rigidification and presentation of interaction surfaces
Sheet enrichment	50	–	secondary structure propensity contrast ≥ 0.10	extended recognition geometries and strand-rich structural states
Inter-domain approach	50	reference patch	source-referenced median distance decreases by at least 1.0 Å	domain closure, interface formation, and allosteric rearrangement
Inter-domain separation	50	reference patch	source-referenced median distance increases by at least 1.0 Å	domain opening, relief of autoinhibition, and changes in multivalent geometry
Reference-fragment mimicry	50	reference fragment and residue map	backbone RMSD ≤ 2.0 Å, at least 3 mapped residues, coverage ≥ 0.80	transfer of binding, catalytic, or epitope geometries
Core burial	50	–	median RSA decreases by at least 0.10, final RSA ≤ 0.40, with constrained local compaction	packing, solvent shielding, and stabilization-oriented redesign
Head-to-tail cyclization	50	terminal editable region	one new head-to-tail amide bond with N–C distance in $[1.20, 1.50]$ Å	conformational constraint and proteolytic stabilization of peptides
Motif scaffolding	30	conserved motif segments	motif preservation plus the MotifBench refolding criterion	transfer of a functional geometry onto a new structural carrier

A EIGHT PROTEAN EDIT TYPES.

See Table 3. We choose some hyperparameters such as 0.35, 0.10 because of the corresponding statistical quantile.

B WORKED EXAMPLES AND TASK CONSTRUCTION

This appendix grounds the benchmark definition in concrete artifacts: how the tasks were constructed (Section B.1), what a task card contains for each of the eight edit types (Section B.2), the exact prompt a model receives and the specification it returned (Sections B.3 and B.4), rendered source/candidate structures for two successful edits and one cyclization rollout that closes the ring yet fails the scope tolerance (Section B.5), pseudocode for the three gates (Section B.8), and detailed bottleneck ablations tracking structural leakage and biophysical refolding attrition (Section B.10).

B.1 TASK CONSTRUCTION PIPELINE

Figure 4 summarizes how the evaluation set was produced. Parameters quoted on the edges are the frozen construction values; all are recorded in the construction scripts and build manifests.

B.1.1 SOURCES.

The experimental leg collects cross-referenced PDB entries for reviewed UniProt accessions (at most three entries per accession, X-ray and electron-microscopy preferred) and inherits the quality filters of the interface-loop harvest that seeded the editable-region geometry: X-ray resolution ≤ 3.0 Å, $R_{\text{free}} \leq 0.30$ and $R_{\text{free}} - R_{\text{work}} \leq 0.08$, editable chain 40–300 residues and fixed chains 80–700 residues, total modelled length ≤ 900 residues, buried surface area 500–2,500 Å^2, and at least eight interface residues on each side. The predicted leg uses reviewed human Swiss-Prot entries of length 100–800 whose AlphaFold Database model has sequence coverage ≥ 0.90 and pLDDT ≥ 70 both

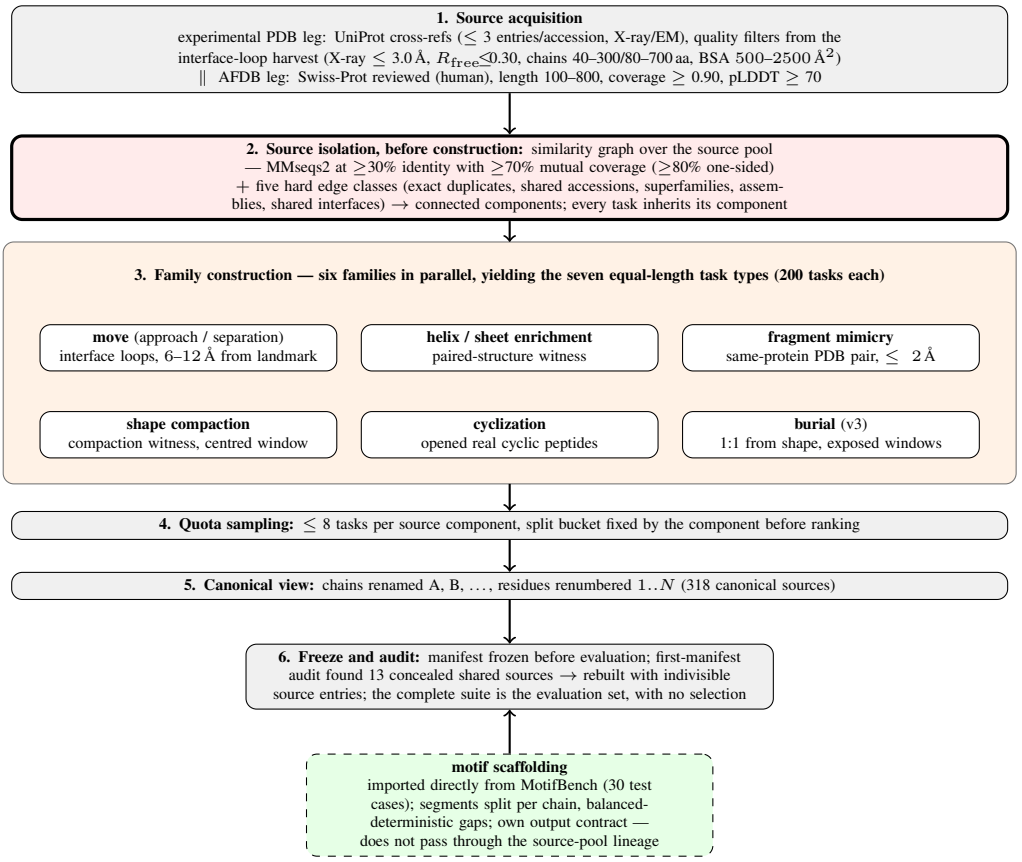


Figure 4: Task-construction pipeline. Sources come from two filtered legs, and *each source is grouped at intake*. The load-bearing step is the source-isolation graph, computed on the source pool *before* any task is constructed: near-duplicates and related structures are merged into connected components, and every task inherits its component. The seven equal-length task types are then built in parallel from the grouped pool—six construction families, with the move and secondary-structure families each yielding two task types, and burial derived one-to-one from the shape family. Motif scaffolding is *imported directly from MotifBench* under its own output contract; it does not pass through the source-acquisition or isolation lineage and joins the suite only at the frozen evaluation set. Quota sampling caps tasks per component ($\leq$ 8, split bucket fixed by the component), the canonical view renumbers structures for model input, and the manifest is frozen after audit—the complete suite (350 + 30 motif) forms the evaluation set.

globally and over the edit region. The release mixes 1,142 experimental and 258 predicted tasks across the seven equal-length families.

B.1.2 Per-family construction.

Editable windows are selected under family-specific feasibility conditions: distance tasks require a source-to-patch median distance in 6–12 Å (centred at 9 Å by the selection score); secondary-structure tasks pair the source with another experimental structure of the same protein whose helix (or strand) count exceeds the source's by at least three over the window; mimicry tasks align a same-accession PDB pair globally and take 24–48-residue blocks whose backbone Kabsch RMSD is at most 2 Å (the residue map is the aligned block); shape tasks use a paired witness with a 5% smaller radius of gyration and rebuild a centred window of width $\min(80, \max(40, n/3))$; cyclization tasks are *opened* real cyclic peptides—nine witness chains of 13–40 residues whose termini are rotated in 5° steps until the N–C distance lands in 2.0–4.0 Å, with the first and last four residues forming the editable region.

B.1.3 MOTIF IMPORT.

The 30 motif-scaffolding cases are MotifBench test cases: the multi-segment motif is split into per-segment structures, and scaffold gaps distribute the remaining length deterministically over segment slots (case 1: $52 + 21 + 52 = 125$ residues). These canonical gaps are construction bookkeeping only: the model-facing instruction leaves the scaffold lengths open and requires only arithmetic self-consistency with the public total length (Table 4).

B.1.4 CANONICAL VIEW, QUOTAS, ISOLATION, FREEZE.

For model input and scoring, every source is canonicalized (chains renamed A, B, ..., residues renumbered $1..N$; 318 canonical sources). Quota sampling yields 200 tasks per family with at most eight tasks per source group. The source-isolation graph (MMseqs2 at $\geq$30% identity with $\geq$70% mutual coverage, $\geq$80% one-sided for short sequences, plus the five hard edge classes of Appendix A) groups related sources; the manifest was frozen before evaluation after an audit found thirteen task groupings that concealed a shared source. The final evaluation set is the complete suite with no held-out selection.

B.2 TASK CARDS, ONE PER EDIT TYPE

Each task ships as a card carrying the instruction (verbatim, English and Chinese), the role partition as per-chain residue ranges, the sequence policy, and the public target criteria. Table 4 gives one real card per edit type (compact excerpt; long index arrays elided). The card received by a model is the same object minus all routing metadata.

The full card of the inter-domain-approach example used throughout this appendix:


```
{"task_id": "ptask-1c11ae980dba27dd3a12", "action": "move_closer",
 "instruction": {"en": "Move the red edit core closer to the purple
   reference; substitutions only inside the edit core.", "zh": "..."},
 "roles": {
   "edit_core":        {"B": {"label_seq_ranges": "99-110"}},
   "relax_shell":      {"B": {"label_seq_ranges": "91-98,111-118"}},
   "hard_protected":   {"B": {"label_seq_ranges": "1-90,119-128"}},
   "fixed_context":    {"A": {"label_seq_ranges": "1-202"}},
   "reference_patch":  {"A": {"label_seq_ranges": "7-12,193,196-197,201"}}},
 "sequence_policy": {"profile": "core_redesign_v1", "mode": "core_redesign",
   "editable_regions": ["edit_core"],
   "protected_regions": ["relax_shell", "reference_patch",
                         "fixed_context", "hard_protected"],
   "max_substitution_fraction": null, "preserve_length": true,
   "preserve_chain_ids": true, "preserve_residue_order": true,
   "preserve_atom_mapping": true, "forbid_noncanonical": true},
 "target_criteria": {"primary_metric_id": "median_min_pseudo_cb_distance_delta_v2",
   "target": {"metric_id": "median_min_pseudo_cb_distance_delta_v2",
     "move_away":  {"delta_threshold_angstrom": 0.5},
     "move_closer": {"delta_threshold_angstrom": -0.5},
     "source_distance_angstrom": 8.954}},
 "source": {...}, "split": "test", "benchmark_split": "test",
 "source_content_group_id": "isolation-0daaf9b03d0388fabc69"}
```


The card carries the thresholds of the profile that constructed it; under the frozen evaluation profile the move gate is ± 1.0 Å and the card-level value is ignored (the scorer flags such rows).

B.3 WHAT THE MODEL SEES

Each attempt is a single API call with two messages. The system prompt is fixed per backend:

```
You are a protein design engineer.  Translate the benchmark
task into a {backend} input specification.  Respond with a
single JSON object and nothing else.
```

The user message concatenates four segments (12,462 characters for the worked example):

```
# Backend Capability Card
<markdown card: input schema, contig/design syntax, hotspot and bond
 channels, expressivity boundary, legality rules>
```

Table 4: One real task card per edit type (excerpt). Ranges are `chain:label-seq`; the editable region is listed under "edit scope" with the reference landmark where applicable.

Type / task id	Edit scope (chain:residues)	Instruction (English, verbatim)
Helix enrichment `ptask-0227c7e0...`	`D:313-360`	"Make the red edit core alpha-rich; substitutions only inside the core and preserve protected sequences."
Sheet enrichment `ptask-00201762...`	`A:97-144`	"Make the red edit core beta-rich; substitutions only inside the core and preserve protected sequences."
Inter-domain approach `ptask-1c11ae98...`	`B:99-110`; ref. `A:7-12,193,196-197,201`	"Move the red edit core closer to the purple reference; substitutions only inside the edit core."
Inter-domain separation `ptask-00a70700...`	`A:113-118`; ref. `B:29,111,114-116`	"Move the red edit core away from the purple reference; substitutions only inside the edit core."
Reference-fragment mimicry `ptask-01425f04...`	`A:49-96`; ref. `2w8b_A` + residue map	"Make the red target fragment mimic the purple reference backbone; substitutions only inside the core."
Core burial `burial_psource-3c1d...`	`A` (12-res. window)	"Bury the red edit core: reduce its solvent accessibility (median RSA drop) while compacting it; substitutions only inside the edit core."
Head-to-tail cyclization `ptask-030bf23a...`	`A:1-4,13-16`	"Cyclize the red N/C-terminal edit core with one declared head-to-tail amide bond; substitutions only in the core."
Motif scaffolding `motif_case_12`	motifs `M0` (7 res.) + `M1` (17 res.); redesign `M0`:1,7, `M1`:1,3,8,16,17	"Scaffold the fixed motifs M0 (7 residues) and M1 (17 residues) into a single chain of 100 residues, choosing the scaffold segment lengths yourself; the motif backbone geometry must remain fixed and residues 1, 7 of M0 and residues 1, 3, 8, 16, 17 of M1 may undergo sequence redesign."

```
# Scoring Criteria
<markdown card rendered from the frozen scorer constants>

# Task Card
<the task-card JSON above, with all routing metadata removed:
 routing, backend_hint, required_primitives, backend_support,
 allowed_backends, preferred_backend, support_reason>

# Feedback on your previous attempt        (round >= 2, on failure only)
<public gate outcomes and failure classes>

Output the JSON input specification now.
```

Sampling is deterministic given the seed (temperature 0.2); at most one structured retry is issued for malformed output. Because the protocol forbids prose, the model's response *is* the final backend specification—there is no separate reasoning trace to display. Every request and response is logged verbatim per task and round, and the examples below quote those logs.

B.4 ROLLOUTS: PROMPT, SPECIFICATION, AND OUTCOME

Success (inter-domain approach, DeepSeek × ODesign). For the worked card above, the model returned the ODesign specification

```
{"chains": [{"chain_type": "proteinChain", "sequence": "A/1-202", "if_cyc": false},
            {"chain_type": "proteinChain", "sequence": "B/1-98,12-12,B/111-128",
             "if_cyc": false}],
 "hotspot": "A/7,A/8,A/9,A/10,A/11,A/12,A/193,A/196,A/197,A/201",
 "partial_diff": "B/99-110"}
```

The specification is worth reading closely. Chain A (the fixed context) is kept in full; on chain B the model kept residues 1–98 and 111–128 and asked for an anonymous twelve-residue redesign segment (the `12-12` contig, matching the editable region's length) whose position makes it cover exactly `B/99-110`; the diffusion region `partial_diff` names the same span; and the ten hotspots are precisely the reference-patch residues listed on the card. In other words, the model derived the keep/redesign split, the diffusion window, and the landmark conditioning from the task card—the compilation skill the benchmark measures.

Success (helix enrichment, DeepSeek × ODesign). For the card whose edit core is `A:97-144` (48 residues), the model returned the ODesign specification

```
{"chains": [{"chain_type": "proteinChain",
             "sequence": "A/1-96,48-48,A/145-282", "if_cyc": false}],
 "hotspot": "", "partial_diff": "A/97-144"}
```

The anonymous 48-residue redesign contig (`48-48`) is positioned so that it covers exactly `A/97-144`, and the diffusion region names the same span—the same keep/redesign compilation skill as the approach example, at six times the region size. The candidate's editable region, coil throughout in the source, is remodelled into helices covering 56% of the window (Figure 6).

Cyclization: target pass, scope fail (DeepSeek × BoltzGen). For a 22-residue opened cyclic peptide (edit core `A:1-4,19-22`), the model returned the BoltzGen specification

```
{"entities": [{"file": {"include": [{"chain": {"id": "A"}}],
                        "design": [{"chain": {"id": "A", "res_index": "1..4"}},
                                   {"chain": {"id": "A", "res_index": "19..22"}}]}}],
 "constraints": [{"bond": {"atom1": ["A", 22, "C"],
                           "atom2": ["A", 1, "N"]}}]}
```

The two design intervals cover the terminal edit core exactly, and the bond constraint declares the head-to-tail amide bond between `A22:C` and `A1:N`—the bond channel this task type exists to test. The returned candidate closes the ring: exactly one new bond is formed and $d_{\text{N-C}} = 1.32$ Å, inside the $[1.20, 1.50]$ Å band, so the target gate passes (Figure 7), and both substitutions stay inside the core. The scope gate nevertheless fails the candidate: ring closure relaxes the protected backbone by up to 0.21 Å against the 0.10 Å tolerance. This failure mode is mechanical rather than a compilation error—indeed, no candidate from any backend passes all three gates on this task type in the frozen run, because a genuinely closed ring cannot leave fourteen protected residues rigid.

B.5 GENERATED STRUCTURES

Figures 5–7 render the source and candidate structures of the three rollouts above (PyMOL cartoons; editable region in firebrick red, reference patch in purple, remaining protected context in slate; the cyclization example additionally draws the two terminal residues as sticks and marks the N$\cdots$C distance). Candidates are aligned to the source by a rigid fit over protected Cα anchors (post-alignment anchor RMSD 0.000 Å for the two redesign examples; 0.075 Å for the cyclization example, whose ring closure necessarily relaxes the protected backbone).

B.6 CYCLIZATION DECOMPOSITION: SHARED SCOPE CANDIDATES AND CLOSURE

The head-to-tail family contains 50 tasks and is expressible only on a backend with an explicit covalent-bond primitive. We therefore report the BoltzGen lane separately and mark ODesign and RFd3 as "–" rather than treating an unavailable input language as a model failure. The audit below

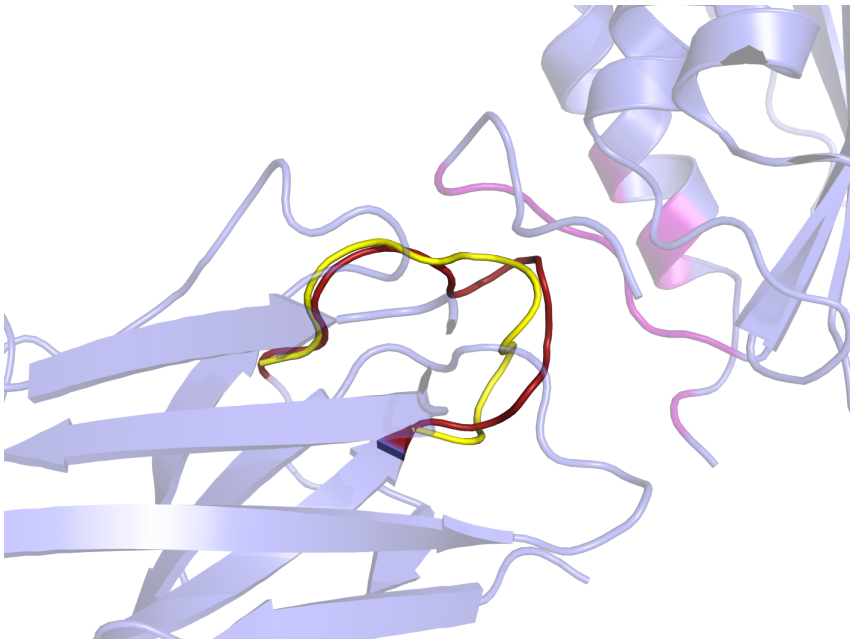


Figure 5: Inter-domain approach, successful. The source and candidate structures are overlaid in a single view. The editable region (red, `B:99–110`) moves toward the reference patch (purple, on chain A). The dashed line marks the closest patch contact of edit-core residue `B:101` (to `A:201`), which shortens from 10.7 Å to 6.0 Å, and the median source-referenced distance drops from 8.95 Å to 7.48 Å ($\Delta = -1.43$ Å, gate requires ≤ -1.0 Å). All three gates pass.

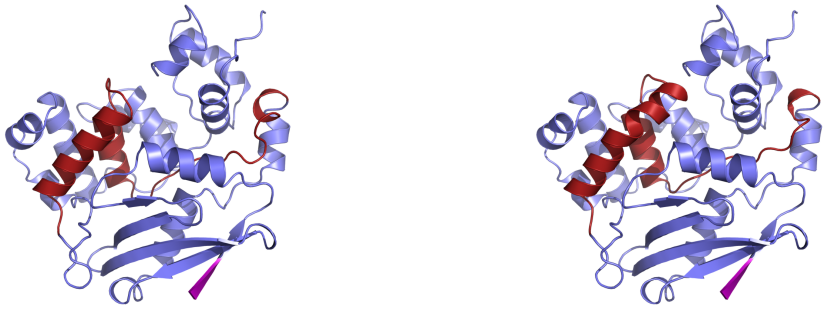


Figure 6: Helix enrichment, successful. The 48-residue editable region (red, `A:97–144`) is assigned no helix in the source (DSSP helix fraction 0.00 over the window); in the candidate it is remodelled into helices covering 56% of the window (helix fraction 0.56, balance $0.54 \geq 0.35$), with 42 of 48 residues substituted inside the core and the protected context emitted unchanged (maximum protected-backbone displacement 3×10^{-5} Å). All three gates pass.

uses the five available model–BoltzGen score files; the Gemini score file was not present at the time of this appendix freeze. Each model contributes 850 auditable candidate rows (17 candidates per task). "Topology" requires exactly one new native A:22:C–A:1:N head-to-tail bond; "geometry" additionally requires a 1.20–1.50 Å terminal distance; "scope" is the protected context gate; and "joint" is the same-candidate conjunction.

The 50 scope-pass rows are effectively shared across models. They are all rollout 1; after canonicalizing atom coordinates and topology, all 50 task-matched candidates are identical across the five models. Forty-nine of 50 CIF files are byte-identical as well. The only byte-level difference is



Figure 7: Head-to-tail cyclization. The 22-residue peptide's edit core (red) spans the first and last four residues; the terminal residues are drawn as sticks and the dashed line marks the N$\cdots$C distance. Left: opened source (3.9 Å); right: candidate, in which the declared head-to-tail amide bond closes the ring ($d_{\text{N-C}} = 1.32$ Å, inside the $[1.20, 1.50]$ Å target band; exactly one new bond). The target gate passes; the scope gate does not: closing the ring relaxes the protected backbone by up to 0.21 Å (tolerance 0.10 Å), while both substitutions stay inside the core and none of the four structural invariants changes.

Model	**Topology** 850/850	**Geometry** /850	**Scope** /850	**Joint** tasks / 50	**Any geom.** tasks / 50	**Any scope** tasks / 50
Claude Opus 5	100.0%	774 (91.1%)	50 (5.9%)	0	50	50
DeepSeek V4 Pro	100.0%	773 (90.9%)	50 (5.9%)	0	50	50
GPT-5.6-Sol	100.0%	778 (91.5%)	50 (5.9%)	0	50	50
Kimi K3	100.0%	766 (90.1%)	50 (5.9%)	0	50	50
Qwen3.8Max	100.0%	779 (91.6%)	50 (5.9%)	0	50	50

Unsupported backends: ODesign and RFd3 are "–" (no expressible cyclization specification).

Table 5: Cyclization gate decomposition for the five available BoltzGen score files. Topology, geometry, and scope are candidate-level counts; "any" columns are task-level best-of-17 diagnostics. ODesign and RFd3 do not expose an explicit covalent-bond primitive and are therefore shown as unsupported ("–").

a single GPT file whose occupancy metadata differs; its atom coordinates, residue identities, and bond record are unchanged. These rows therefore represent the common source/adapter-preserving candidate rather than five independent model-specific solutions. They should not be used to claim that one language model preserves the protected region better than another.

The model-dependent signal appears in the remaining candidates. Qwen3.8Max has the highest geometry-pass rate ($779/850 = 91.6\%$) and the smallest non-closing tail; GPT-5.6-Sol follows at $778/850 = 91.5\%$. Claude Opus 5 and DeepSeek V4 Pro are close behind, whereas Kimi K3 has the lowest geometry-pass rate ($766/850 = 90.1\%$) and the largest non-closing tail. These differences are small (1.5 percentage points) and arise only after the explicit bond constraint is compiled successfully; they are not differences in protected scope preservation.

For the closed candidates, the aligned protected-backbone displacement has a median of approximately 0.43–0.45 Å, while the frozen scope tolerance is 0.10 Å. These candidates preserve length, residue order, backbone mapping, and edit all substitutions inside the declared core, but they trigger `protected_coordinate_changed`. The unchanged scope-pass candidate leaves the termini approximately 3.0–4.0 Å apart and consequently misses the cyclization distance. The strict $0/50$ result is therefore the empty intersection of two candidate sets: one set preserves the context and one set closes the bond.

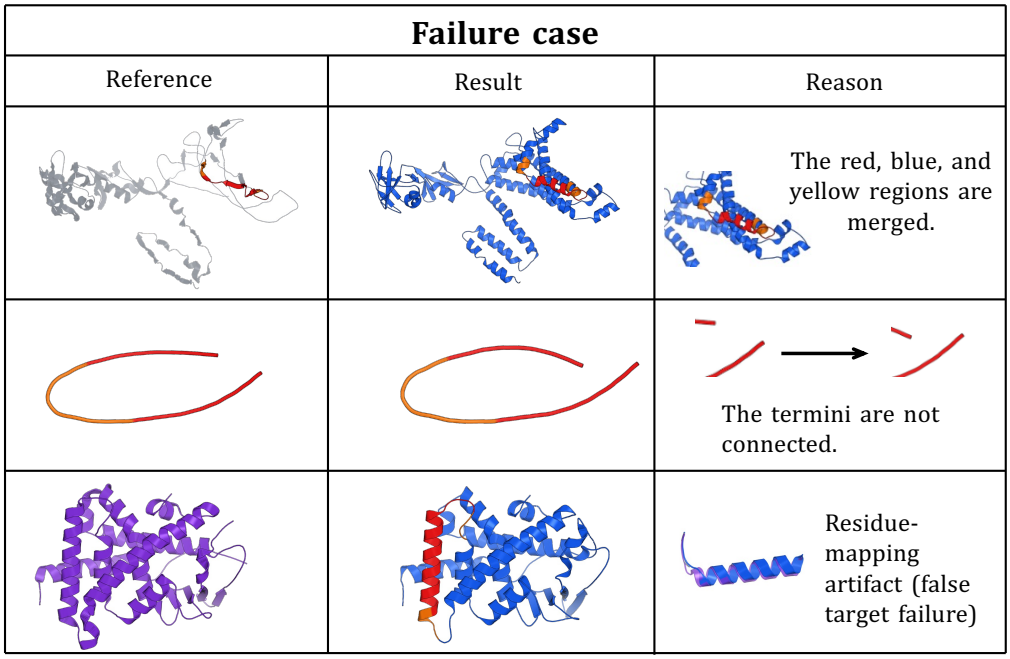


Figure 8: Representative cyclization failure cases. Top: the editable and protected color regions are merged, so the region partition required by the contract is not preserved. Middle: the generated chain remains open and the termini are not connected, which is a genuine topology/target failure once the structure and residue mapping are valid. Bottom: a residue-mapping artifact creates an apparent target failure even though the visual structure is not the source of the discrepancy; such rows must be remapped and rescored before they are attributed to a model.

This case is a coupled-contract diagnostic. The models can emit the required topology and usually attain the local bond geometry, while the frozen contract does not permit the local relaxation needed to close the termini. A cyclization-specific report should therefore list topology, geometry, scope, and strict joint gates separately, and should protect only genuinely fixed context while allowing a calibrated terminal/relax-shell motion. The existing CIFs can be re-scored under that relaxed cyclization contract without repeating the GPU generation; the strict frozen-context score can remain as a secondary stress test.

B.7 Cyclization failure cases and accounting of false failures

The model-specific comparison above concerns candidates that pass the topology and geometry checks. A separate failure-case audit is needed because a visual failure can arise at different layers of the pipeline. Figure 8 shows three representative patterns: a region-partition failure in which the red, blue, and yellow regions are merged; a genuine open-chain result in which the termini remain disconnected; and a residue-mapping artifact that creates a false target failure. These patterns are not interchangeable and are assigned to different accounting terminals. The supplied panel is a representative diagnostic and carries no task or model identifier; we therefore do not attribute a pictured row to a particular model without its task-level log.

The middle panel is the failure mode most directly relevant to the requested cyclization primitive: the candidate does not realize a head-to-tail bond and therefore cannot satisfy the $d_{\mathrm{N-C}}$ target. It should be reported separately from the closed-but-scope-violating candidates in the decomposition table. The top panel indicates that the agent or adapter did not preserve the contract's region partition, while the bottom panel demonstrates why mapping validation must precede model attribution. In particular, a strict success rate should not be lowered by a row that is later shown to be a scorer artifact.

Failure pattern	Primary accounting terminal	Interpretation and treatment
Region colors merged	Region partition / scope terminal	The editable, protected, and terminal regions cannot be separated reliably. Do not call this a cyclization target failure; retain it as a compilation or scope-contract failure and inspect the native specification.
Termini not connected	Target topology/geometry failure	The output is an open chain. If valid output, structural integrity, and residue mapping pass, count it as a genuine target failure; do not merge it with unsupported-backend or scope failures.
Residue-mapping artifact	Scorer/mapping audit	The apparent target miss is caused by comparing the wrong residue correspondence. Quarantine the row, repair the official map, and rescore it before assigning a model failure.

Table 6: Failure-case taxonomy for cyclization. The three rows correspond to the panels in Figure 8; only the middle row is a target failure by itself.

Across the five available BoltzGen models, there is no non-empty set of strict cyclization-passing samples to compare: every model has 0/50 task-level joint success. The model-specific differences are instead in the geometry-pass candidate sets (Qwen3.8Max $779/850$, GPT-5.6-Sol $778/850$, Claude Opus 5 $774/850$, DeepSeek V4 Pro $773/850$, and Kimi K3 $766/850$). The 50 scope-pass candidates are shared after canonicalization (49/50 CIFs are also byte-identical), so they provide no evidence for a model-specific protected context advantage. Gemini is not included in this comparison because its BoltzGen score file was not available at the appendix freeze.

For the failure-case figure, the correct conclusion is therefore layered: models almost always compile the explicit bond primitive and usually produce local closure geometry, but strict cyclization remains blocked either by an open-chain target miss, by a region/scope violation, or by an evaluation-map artifact. These terminals should remain separate in the final score table.

B.8 Gate evaluation, in pseudocode

Algorithm 1 gives the evaluator at the level of the frozen scorer. Algorithm 2 spells out the per-edit-type target criteria behind GateTarget, and Algorithm 3 records the three-gate mirror applied to the imported motif-scaffolding cases. All thresholds are the frozen profile values; card-carried overrides from earlier profiles are ignored.

Open-design run (status at freeze). The open-scaffold caliber above—canonical gaps withheld from the instruction, the declared placement audited only for arithmetic self-consistency—was evaluated with the same six language models and three backends as Section 4, on all 30 imported cases per configuration. In the frozen 18-configuration run, grounding, the declared-placement assembly gate, and the motif-preservation gate saturate at $30/30$ everywhere: the two deterministic gates no longer separate models on the imported family, and the discriminative burden rests entirely on the sidecar refold gate, whose scoring was pending at the freeze. Consistent with Algorithm 3, no motif-scaffolding task is counted as a success in the absence of sidecar evidence. Two configuration caveats apply: the RFdiffusion3 adapter compiles with fully fixed motif sequences (an upstream defect), so its redesign capability is not exercised; and Claude Opus 5 requests omit the temperature parameter, which its provider rejects. The one-sentence instruction quoted in Table 4 is the condensed deployment of the same contract; under it, grounding completes on $178/180$ model–task pairs, with each of two models failing a single task.

B.9 Qualitative Case Studies: Spec-Space Planning Disparity

To understand how high-level biological reasoning translates into low-level backend parameterizations, Figure 9 illustrates two representative compilation episodes executed on the ODesign backend.

Algorithm 1 Three-gate evaluation of one candidate

Require: task $\tau = (s, X, \mathcal{M}, u)$, source rows, candidate rows

function GATETARGET(τ, src, cand)
  if τ.action = *approach / separation* **then** ▷ per-type criteria: Algorithm 2
    $A \leftarrow$ anchors = protected Cα away from core and patch
    $T \leftarrow \mathrm{KabschFit}(\mathrm{cand}|_A \rightarrow \mathrm{src}|_A)$
    $\hat{D} \leftarrow \mathrm{median}_{i \in \mathcal{M}_{\mathrm{edit}}} \min_{j \in \mathcal{M}_{\mathrm{ref}}} \|T\hat{c}_i - c_j\|$ ▷ patch points from the source
    $D \leftarrow$ same on the source core
    return $(\hat{D} - D \leq -1.0)$ if approach else $(\hat{D} - D \geq +1.0)$
  else if τ.action = *helix / sheet* **then**
    $(f_H, f_E) \leftarrow$ DSSP fractions over cand's editable region
    $b \leftarrow f_H - f_E$ if helix else $f_E - f_H$
    return $b \geq 0.35$ (helix) / $b \geq 0.10$ (sheet) ▷ absolute level of the candidate
  else if τ.action = *mimicry* **then**
    return $\mathrm{KabschRMSD}(\text{mapped N/CA/C/O}) \leq 2.0 \wedge \text{mapped} \geq 3 \wedge \text{coverage} \geq 0.80$
  else if τ.action = *burial* **then**
    $r \leftarrow$ per-residue RSA (Shrake–Rupley, Tien) over cand
    return $\mathrm{median}(r|_{\mathrm{core}}) - \mathrm{median}(r^{\mathrm{src}}|_{\mathrm{core}}) \leq -0.10 \wedge \mathrm{median}(r|_{\mathrm{core}}) \leq 0.40 \wedge R_g$ ratio $\in [0.70, 0.95]$
  else ▷ cyclization
    return exactly one new head-to-tail amide bond $\wedge$ $d_{\mathrm{N-C}} \in [1.20, 1.50]$ $\wedge$ topology evidence complete
  end if
end function

function GATEREFOLD(τ, cand)
  if τ.action = cyclization **then return** N/A
  end if
  motif $\leftarrow$ protected residues of the edit chain, frozen
  $S \leftarrow$ ProteinMPNN redesign of the editable region (k=4)
  for $s \in S$ **do** ▷ ESMFold refold of each sequence
    if $\mathrm{RMSD}_{\mathrm{chain}}(s) < 3.0$ **and** $\mathrm{RMSD}_{\mathrm{motif}}(s) < 2.0$ **then return** PASS
    end if
  end for
  return FAIL ▷ no passing sample found in budget
end function

function GATESCOPE(τ, src, cand)
  return length, chains, order, N/CA/C/O mapping unchanged $\wedge$ identity outside $\mathcal{M}_{\mathrm{edit}}$ unchanged (three-letter comparison) $\wedge$ canonical residues only $\wedge$ protected backbone ≤ 0.1 Å after anchored fit
end function

strict $\leftarrow$ GateTarget $\wedge$ GateRefold $\wedge$ GateScope
terminals (reported separately, never folded into zeros): unsupported-by-backend $>$ backend-failure $>$ invalid-spec $>$ gate failures; missing refold evidence is *pending*, not failure

Both cases expose how subtle differences in the compiled specification reflect fundamentally different levels of protein understanding, directly determining macroscopic task outcomes.

B.10 Execution Bottlenecks and Biophysical Attrition Analysis

Target Attainment vs. Off-Target Scope Leakage. Crucially, reaching the assigned target metric (G_1) does not constitute a valid edit if the surrounding scaffold is deformed. This distinction is most dramatic in RFdiffusion3: while the model successfully drives the editable core toward the requested geometry in 96.0 tasks (27.4%), it maintains scope fidelity in only 51.0 tasks (14.6%). As a direct result, joint geometric compliance collapses to just 26.0 tasks (**7.4%**). Unconstrained

Algorithm 2 GateTarget per edit type (the switch of Algorithm 1 expanded; frozen profile values)

Require: candidate aligned to the source by the anchored fit T of Algorithm 1; masks $\mathcal{M}_{\text{edit}}, \mathcal{M}_{\text{ref}}$; D = the source value of the case metric

if τ.action = *inter-domain approach / separation* **then**
  metric: median over edit-core residues of the nearest pseudo-CB distance to reference-patch points ($\hat{D}$, computed as in Algorithm 1)
  return $\hat{D} - D \le -1.00$ Å (approach) / $\hat{D} - D \ge +1.00$ Å (separation)
else if τ.action = *helix / sheet enrichment* **then**
  $(f_H, f_E) \leftarrow$ DSSP residue fractions over the candidate's editable region (mkdssp when available; the deterministic ϕ/ψ heuristic serves only as per-file fallback)
  return $f_H - f_E \ge 0.35$ (helix) / $f_E - f_H \ge 0.10$ (sheet)
else if τ.action = *reference-fragment mimicry* **then**
  $R \leftarrow$ backbone Kabsch RMSD of the mapped edit-core residues (N/CA/C/O) against the reference fragment
  return $R \le 2.00$ Å $\wedge$ mapped ≥ 3 $\wedge$ coverage of the editable region ≥ 0.80
else if τ.action = *shape compaction (v3)* **then**
  $\delta_{\text{SASA}} \leftarrow 1 - \sum_{i \in \mathcal{M}_{\text{edit}}} \text{SASA}_i^{\text{cand}} \,/\, \sum_{i \in \mathcal{M}_{\text{edit}}} \text{SASA}_i^{\text{src}}$ ▷ like-for-like atoms: backbone+CB when either side lacks side chains
  return $\delta_{\text{SASA}} \ge 0.10 \wedge R_g^{\text{cand}}/R_g^{\text{src}} \in [0.70, 0.95]$ ▷ anti-collapse guard; the v2 asphericity leg is diagnostic only
else if τ.action = *core burial (v3)* **then**
  $r \leftarrow$ per-residue relative solvent accessibility over the candidate (Shrake–Rupley, Tien)
  return $\text{median}(r|_{\mathcal{M}_{\text{edit}}}) - \text{median}(r^{\text{src}}|_{\mathcal{M}_{\text{edit}}}) \le -0.10 \wedge \text{median}(r|_{\mathcal{M}_{\text{edit}}}) \le 0.40 \wedge R_g^{\text{cand}}/R_g^{\text{src}} \in [0.70, 0.95]$
else ▷ head-to-tail cyclization
  return exactly one new head-to-tail amide bond $\wedge$ $d_{\text{N-C}} \in [1.20, 1.50]$ Å $\wedge$ adapter topology evidence complete
end if

no target case exists for motif scaffolding inside the agent lineage: the imported contract is evaluated by Algorithm 3

Algorithm 3 Motif-scaffolding evaluation (imported MotifBench output contract; the three-gate mirror of Algorithm 1)

Gate 1, assembly: format valid (single-chain mmCIF, exactly one model) $\wedge$ placement valid (declared total length and segment order; scaffold gaps are the model's own design, audited only for arithmetic self-consistency with the public total length)
Gate 2, sidecar refold: any of 8 ProteinMPNN $\rightarrow$ ESMFold designed sequences refolds with whole-chain backbone RMSD < 2.0 Å and motif RMSD < 1.0 Å
Gate 3, motif preservation: motif identity valid $\wedge$ per-segment motif backbone RMSD ≤ 2.0 Å against the source motif (Kabsch over N/CA/C/O) $\wedge$ sequence changes confined to scaffold positions
return success = Gate 1 $\wedge$ Gate 2 $\wedge$ Gate 3; missing sidecar evidence never passes and is never imputed

diffusion backbones routinely achieve local structural objectives by dispersing structural strain into invariant motifs, proving that strict coordinate-level boundary auditing (C_α–RMSD ≤ 0.5 Å) is indispensable to prevent global conformational drift. Conversely, BoltzGen and ODesign provide substantially tighter contextual conditioning, retaining 94.5% and 80.6% scope fidelity, respectively.

B.11 Special Setting in motif scaffolding

Budget Scaling Dynamics ($s@1 \rightarrow s@16$) and Kinetic Traps. Tracking the scaling trajectory from $s@1$ to $s@16$ underscores the role of structural diversity in overcoming conformational energy barriers. Across all models, scaling the candidate budget yields substantial performance amplification (e.g., a $> 3\times$ gain for Claude-Opus-5). Scaffolding isolated motifs requires navigating highly

(A) CASE 1 — SEPARATION: HOTSPOT CURATION

Task. *"Move the red edit core away from the purple reference; substitutions only inside the edit core."*
Context. Edit core: `B:67--74` (∼ 310 kcal interface); reference patch: `A:{62, 64, 95--97, 101--102, 141, 143, 146--148}`.

Model	Compiled Output (JSON)
Claude-Opus-5	`{"hotspot": "A/62, A/141, A/147", "partial_diff": "B/59--82"}`
Kimi-K3	`{"hotspot": "A/62, A/64, A/95, A/96, A/97, A/101, A/102, A/141, A/143, A/146, A/147, A/148", "partial_diff": "B/59--82"}`
Qwen3.8Max	`{"hotspot": "A/62", "partial_diff": "B/67--74"}`

Biophysical Mechanism. ODesign's `hotspot` primitive defines the spatial conditioning vector toward which the redesigned segment is steered. Claude-Opus-5 intentionally curates **3 of the 12** candidate patch residues, selecting a coherent spatial cluster on the distal interface that creates an effective pulling vector. In contrast, Kimi-K3 verbatim transcribes **all 12** residues without geometric selection; because the landmarks envelop the interface, their isotropic conditioning vectors cancel out. Conversely, Qwen3.8Max over-prunes the specification down to an isolated **single residue**, providing an insufficient conditioning signal. While all three models satisfy the syntax format, only Claude-Opus-5 encodes an authentic *directional comprehension* of interface mechanics.

(B) CASE 2 — APPROACH: BOUNDARY FIELD AUDITING

Task. *"Move the red edit core closer to the purple reference."*
Context. Edit core: `B:99--110`; relax shell: `B:91--98, 111--118`; reference patch: `A:{7--12, 193, 196--197, 201}`.

Model	Compiled Output (JSON)
Claude-Opus-5	`{"design": "B/1--98, 12--12, B/111--128", "hotspot": "A/8..A/201 (9 res)", "partial_diff": "B/99--110"}` *(Core only)*
DSV/Kimi/Qwen	`{"design": "B/1--98, 12--12, B/111--128", "hotspot": "A/7..A/201 (9 res)", "partial_diff": "B/91--118"}` *(Core + shell)*

Biophysical Mechanism. All four models construct structurally viable backbone design spans and correct hotspot selections. The sole divergence lies in the `partial_diff` field, which defines the continuous coordinates subject to generative denoising. Claude-Opus-5 confines diffusion strictly to the editable core (`B:99--110`). In contrast, DeepSeek-V4-Pro, Kimi-K3, and Qwen3.8Max conflate the editable core with the surrounding *relaxation shell*, expanding diffusion across `B:91--118`. Diffusing through protected residues inevitably alters invariant coordinates and triggers sequence redesign outside the legal window (`substitution_outside_edit_core`). Consequently, **Claude-Opus-5 uniquely solves the task**, whereas the other models fail the G_3 scope contract completely.

Figure 9: Two real compilation episodes on the ODesign backend illustrating how biomolecular spatial reasoning manifests in specification space. **Top (Case 1):** Hotspot curation reflects directional planning across complex interfaces, where uncurated landmarks cancel out directional guidance. **Bottom (Case 2):** Boundary auditing serves as the sole discriminating factor, where confusing the editable core with the relaxation shell results in catastrophic G_3 scope failure.

rugged, non-convex thermodynamic landscapes, where single-candidate sampling ($s@1$) frequently falls into kinetically trapped local minima. Expanding the budget supplies the conformational entropy necessary to sample viable alternative topologies, markedly boosting the recovery of physically self-consistent carriers.

Diagnostic Bottlenecks and the Sim-to-Real Horizon. Fine-grained error diagnostics across contextual editing (Table 1) and de novo scaffolding (Table 2) reveal that failure modes stem not from high-level intent parsing, but from coupled spatial planning (58.7–61.6% failure) and the co-optimization of local deformation with contextual preservation (73.0–82.1% failure; see Appendix B.10 for per-tier failure taxonomies). While baseline agents (such as Kimi-K3) stumble early on specification compilation, frontier models (Claude-Opus-5 and GPT-5.6-Sol) predominantly fail by over-deforming conserved residues to fulfill target geometries. Crucially, syntactic validity and

isolated metric attainment repeatedly conceal these physical defects. Together, while our findings establish that frontier LLMs possess genuine 3D spatial planning competencies, translating these blueprints into physical molecules remains bottlenecked by generator scope leakage and thermodynamic refolding barriers (detailed downstream biophysical analyses in Appendix B.10). Bridging this simulation-to-reality gap via physics-informed sequence–structure co-design defines the primary horizon for autonomous biomolecular agents.