STROMA: A UNIFIED BENCHMARK FOR MOTIF SCAFFOLDING ACROSS PROTEIN GENERATIVE PARADIGMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motif scaffolding, the design of a protein that preserves a predefined motif while generating the surrounding scaffold, is central to protein engineering. Models that perform this task have advanced rapidly across three generative paradigms: backbone, all-atom and sequence generation. Existing benchmarks were built alongside particular model families, assume incompatible input representations, and cover a narrow range of motifs, so models from different paradigms cannot be compared on equal terms. We introduce **Stroma**, a unified, extensible benchmark of 100 motif scaffolding tasks, each specified at up to five motif resolutions at once, from residue identity to catalytic side-chain geometry, so that every paradigm receives the same task rather than a translation of it. Nearly a quarter are new, covering protein families no prior benchmark included, among them membrane proteins, nanobodies and nucleic-acid-binding motifs. Where an architecture supports them, we evaluate two conditioning modes (indexed and unindexed) and three motif representations (all-atom, tip-atom and theozyme). Success is defined by explicit thresholds that a single predicted structure must satisfy jointly; we test whether rankings survive re-folding under different structure predictors. We benchmark 12 models (15 checkpoints) spanning all three paradigms. Even though all-atom and backbone models perform similarly on average, only all-atom models can take on the more demanding tasks, such as theozyme scaffolding, which backbone models cannot represent. Under minimal-harness conditioning, protein language models underperform both. Thirteen tasks are solved by no model, pointing to opportunities for further work in motif scaffolding.

1 INTRODUCTION

Protein generative models have advanced quickly on motif scaffolding, the task of generating protein sequences or structures that preserve a predefined motif while designing the surrounding scaffold. It is central to enzyme design, interface engineering, binder discovery and antibody design. The task is non-trivial because it simultaneously requires (i) faithful preservation of motif geometry and sequence identity at a resolution sufficient to maintain function, and (ii) generation of a foldable, physico-chemically compatible scaffold that accommodates the motif without steric clashes. These two objectives may conflict, making the task a demanding test of model quality.

Despite rapid methodological progress, evaluation of motif scaffolding models remains fragmented: each modality has its own benchmark, tailored to its paradigm. Early benchmarks assess backbone placement only, recent all-atom benchmarks add atomic specifications but cover only enzymes, and sequence models need residue-level specifications that backbone-only benchmarks cannot supply. Models are therefore evaluated on different task subsets under incompatible inputs, and it is unclear whether differences in reported success rates reflect capability or evaluation design.

To address this problem, we introduce **Stroma**, a unified benchmark for motif scaffolding designed to support all major classes of protein generative models (Figure 1a). Stroma makes four central contributions.

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.

Contributions.

1. **A unified benchmark of 100 tasks.** 77 retained from MotifBench (Zheng et al., 2025) (30), the AME benchmark (Ahern et al., 2026) (41) and Genie2 (Lin et al., 2024) (6), plus 23 new, spanning 26 functional categories.
2. **Expanded biological diversity.** The 23 new tasks cover 13 functional categories absent from every prior benchmark, among them membrane proteins, nanobody CDR loops, nucleic-acid-binding zinc fingers and three-finger neurotoxins, each curated from an experimentally determined structure.
3. **Multi-resolution task specification.** Every task is supplied at five motif resolutions: (i) *motif residue identity*, *motif atoms*, which serves both the (ii) *backbone* resolution and the (iii) *full atomic* resolution, (iv) *tip-atoms* and (v) *theozyme atoms*, when applicable.
4. **A cross-paradigm campaign under one criterion.** 12 models across 32 conditioning runs, scored by one paradigm-neutral test: motif RMSD, plDDT and, on ligand tasks, a clash bound, satisfied jointly on a single predicted structure.

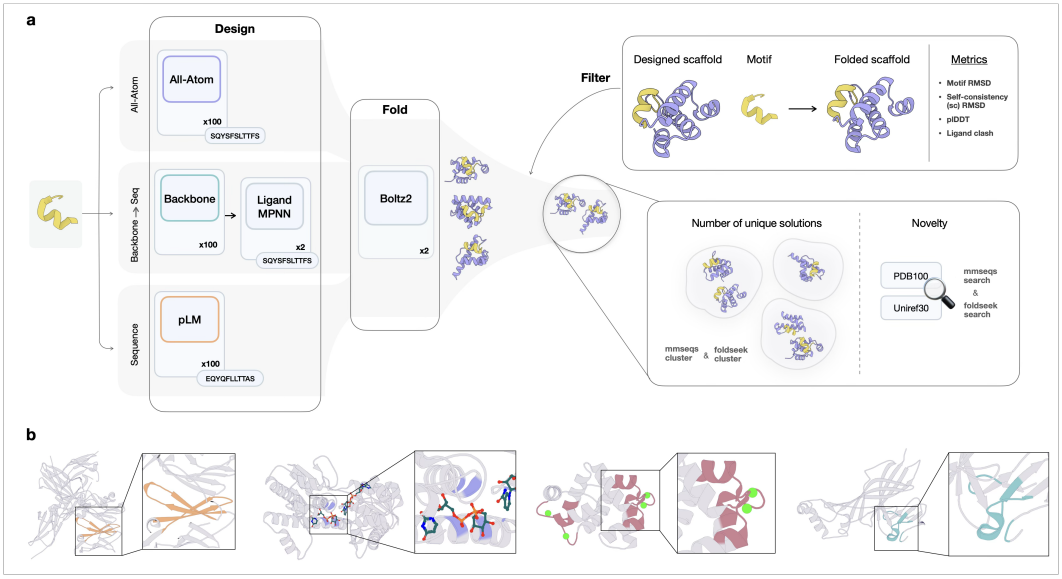


Figure 1: **(a) The Stroma pipeline**. A motif enters on the left and is routed by paradigm: all-atom models and sequence models design directly, backbone models and the all-atom models whose sequence Stroma inverse folds obtain one with LigandMPNN (Dauparas et al., 2025). Every design is then folded by the same structure predictor. A design succeeds only if one predicted structure jointly satisfies every applicable criterion (motif RMSD, self-consistency, confidence and, for ligand-bearing tasks, no clash); successful designs are then clustered to count distinct solutions and searched against sequence and structure databases for novelty. We produce 100 designs per task, two sequences per design for the models that need inverse folding, and two structural predictions per sequence. **(b)** Representative Stroma tasks, one per source, left to right: MB-14 (PDB 5IUS), the β-sheet interface on PD-1 that engages PD-L1; AME-M0092 (PDB 1DLI), the NAD$^+$/UDP-sugar binding active site of UDP-glucose 6-dehydrogenase; G2-1PRW_four (PDB 1PRW), the four canonical helix-loop-helix EF-hand calcium-binding loops of calmodulin, with calcium ions in green; and ST-26 (PDB 2QC1), the three-finger receptor-binding face of α-bungarotoxin, a snake-venom neurotoxin. The motif is coloured against the grey source structure.

2 Related work

Motif-scaffolding benchmarks. The RFdiffusion benchmark (Watson et al., 2023) established the task with 25 problems and became the de facto standard for backbone models; MotifBench (Zheng et al., 2025) filtered and extended it to 30. Both specify motifs at backbone resolution only, so neither

can state the task at the resolution an all-atom model works in. The Atomic Motif Enzyme (AME) benchmark (Ahern et al., 2026) supplies that specification for 41 enzyme active-site tasks but is confined to enzymes, and the Genie2 benchmark (Lin et al., 2024) contributes 6 multi-motif tasks, too few to stand alone. Each is paradigm-specific by construction, so a model can be compared only against others of its own kind. Closest to our aim, Barnett et al. (2026) compare 13 models, the only such comparison to date, but on a single target (TEV protease) with no standardised success criteria. Stroma aggregates the three surviving task sets (77 tasks), adds 23, and specifies every task for all three paradigms at once under one set of thresholds and one pipeline.

Protein generative models. **Backbone models** (RFdiffusion (Watson et al., 2023), Genie2 (Lin et al., 2024) and Genie3 (Lin et al., 2026), Protpardelle-1C (Lu et al., 2025), PPIFlow (Yu et al., 2026)) generate backbone coordinates and obtain sequence by a subsequent inverse-folding step. **All-atom models** generate backbone and side chains jointly, crucial for enzyme active-site design, as inverse folding cannot reliably recover side-chain geometry. They are divided by how their sequence is handled. La Proteina (Geffner et al., 2025) and Protpardelle (Chu et al., 2024) emit a sequence with the structure and are scored on it directly, whereas RFdiffusion2/3 (Ahern et al., 2026; Butcher et al., 2025) and ODesign (Zhang et al., 2025) emit all-atom structure but no sequence the benchmark scores, so theirs is re-derived by inverse folding as for a backbone model. **Sequence models** (PoET-2 (Truong & Bepler, 2025), ESM3 (Hayes et al., 2025), ProGen2 (Nijkamp et al., 2023)) are language models that produce motif-conditioned sequences at high throughput without optimising backbone geometry, imposing motif identity either by constrained decoding (PoET-2, ProGen2) or by conditioning on sequence and structure tracks jointly (ESM3). A list of all checkpoints considered in this work can be found at (Table (S6)). The three families accept different inputs and emit different objects.

3 The Stroma dataset

3.1 Composition: three source benchmarks and 23 new tasks

Stroma consolidates three previously published motif scaffolding benchmarks into one suite, retaining every task from each (Table 1; one task per source is rendered in Figure 1b). **MotifBench** (Zheng et al., 2025) contributes 30 residue-level tasks (prefixed `MB-`), **AME** (Ahern et al., 2026) 41 enzyme active-site tasks with curated atom-level specifications and their ligand annotations, and **Genie2** (Lin et al., 2024) 6 multi-motif tasks whose segments come from different PDB structures and must be scaffolded into one coherent design (prefixed `G2-`).

Table 1: Composition of Stroma by source benchmark.

Source benchmark	Tasks	Primary paradigm	Motif type
MotifBench	30	Backbone design	Structural motifs, interfaces, active sites
AME benchmark	41	All-atom generative	Enzyme active sites
Genie2 benchmark	6	Backbone design	Multi-PDB composite motifs
Stratum (this work)	23	All paradigms	Membrane, nanobody, nucleic acid-binding motifs, toxin, …
Total	**100**	n/a	26 functional categories

After aggregating the three existing benchmarks we performed a gap analysis of biological coverage. Whereas the combined 77 source tasks are heavily weighted toward enzymes (51 tasks, 66% of the combined set), the 23 new tasks substantially broaden the structural, biochemical and interaction inputs presented to the models. They include transmembrane proteins, nanobodies, nucleic-acid-binding proteins, chaperones, protein–protein interfaces, repeat proteins, hormones, antimicrobial peptides, post-translational modification binding sites, structural proteins, photosynthetic proteins, metalloproteins and toxins.

The 23 new tasks are prefixed `ST-`. Each was built from an experimentally determined RCSB PDB structure with its motif identified from the primary literature, selected and verified by a group of protein design researchers, and each motif was curated directly from the deposited coordinates rather

than inherited. Per-category breakdown, curation rationale and all 100 task specifications are in Supplementary B.3 and Table S5.

With the 77 source tasks, Stroma comprises 100 tasks over 26 functional categories (Supplementary Figure S2a), reducing the enzyme fraction from 66% to 51%.

3.2 Multi-resolution task specification

Each task must be evaluable by all three paradigms, which differ in the input format they require. Every task is therefore provided at up to five conceptual resolutions of the same motif: (i) *motif_aa_identity* identity (chain, index and amino acid type) for sequence models; (ii) *motif_atoms* (backbone and side-chain heavy atoms), which serves both the *backbone* resolution, from which backbone models take N/Cα/C/O frames, and the (iii) *all-atom* resolution; and the reduced (iv) *motif_tip_atoms*. The 41 AME enzyme tasks are in addition provided with a fifth resolution *motif_theozyme_atoms*, the catalytic subgraph around each ligand-proximal atom (Section 4.2). Supplementary Table S1 summarises the five resolutions and their supported paradigms.

3.3 Benchmark diversity

Stroma is designed to span a broad range of task difficulties and biological contexts, mitigating benchmark bias toward any single task type and assessing model general capabilities against most possible use-cases.

Functional diversity. The 100 tasks cover 26 functional categories. Enzymes remain the largest at 51 tasks, with viral proteins (7), immune proteins (6) and *de novo* designed proteins and enzymes (6) next, and 18 categories carrying one or two tasks each, so that no single family dominates the evaluation. The full breakdown is in Supplementary Table S4 and Figure S2.

Motif complexity and scaffold length. Motif residue counts range from 1 to 99 (median 10.5, mean 18.5), from single-residue metal sites to large repeat-protein surfaces. For each motif, Stroma reports two complementary views: *motif number* (range 1–4; curated) counts functionally independent substructures, so one enzyme active site is 1 however many loops carry it, while *segment number* (range 1–8) counts the discontiguous sequence stretches a model must scaffold around, directly reflecting the number of independent insertion gaps the model must fill (Figure S1). As per scaffold length, each task specifies a *length range* from which the protocol samples one scaffold length uniformly per design. Ranges follow each source's own convention (Figure S1d). Sampling per design keeps every model on the same target distribution while admitting length-stratified analysis (Supplementary B.2).

Ligands and cofactors. 64 of 100 tasks carry at least one resolved ligand or cofactor, and in every one of them the ligand contacts the motif; pocket residues are recorded at a 5 Å cutoff. The most common are Mg^{2+} (13 tasks), Zn^{2+} (10 tasks), and (Adenosine diphosphate) ADP (5 tasks) (Figure S2b). Tasks with ligands require the generated scaffold to avoid steric clashes with the bound species, an additional constraint not present in apo tasks.

4 Design protocol

4.1 Pipeline architecture and design generation

The pipeline (Figure 1a) routes each model by what it emits. Models that produce structure but no sequence Stroma will score following design $\rightarrow$ inverse fold (LigandMPNN (Dauparas et al., 2025)) $\rightarrow$ predict (Boltz2 (Passaro et al., 2025)) $\rightarrow$ evaluate. To this category belong the five backbone models (RFdiffusion, Genie2, Genie3, PPIFlow and Protpardelle-1C backbone), and all-atom RFdiffusion2, RFdiffusion3 and both ODesign checkpoints. Models that already emit a usable sequence go straight to prediction: La Proteina, Protpardelle-allatom and the three sequence models (PoET-2, ProGen2 and ESM3). Full per-model routing is in Supplementary C.1.

Run counts are parameterised by N (designs per task), M (inverse-folded sequences per design, when applicable) and K (predictions per sequence), yielding $N{\times}M{\times}K$ predicted structures for every model

that inverse folds, and $N \times K$ for those that do not, where M does not apply (Supplementary S6). Every run reported in this paper uses $N = 100$, $M = 2$ and $K = 2$. All models use public checkpoints with fixed seeds and default hyperparameters; full implementation details are in Supplementary C.9.

Coverage constraints. Two practical constraints affect motif coverage. **Single-chain-only models** (Protpardelle-1C single-chain checkpoints, PPIFlow, and the protein language models) cannot natively scaffold multi-chain motifs. Rather than force a quaternary motif into a single polypeptide chain, (which would constitute a distinct task with no single-chain ground truth), these models *skip* multi-chain motifs, which are excluded from the model's denominator. **Multi-PDB motifs**, whose segments originate from distinct source structures and therefore lack a common coordinate frame, are dispatched only when scaffolding is run in unindexed mode or the model declares explicitely the capability. Both scenarios are detailed in Supplementary C.2.

Evaluation metrics. Stroma answers four questions about a design. *Is the motif there?* Motif RMSD at each atom resolution, motif TM-score and motif pLDDT. *Is the scaffold likely to be designable?* Whether a folding model reproduces the designed backbone from its own sequence (self-consistency RMSD and TM-score, designability) and how confidently (pLDDT). *Is it viable and non-degenerate?* Scaffold–ligand clashes, sequence recovery and the compositional complexity of the designed sequence, which catches low-entropy output that folds confidently and would otherwise pass unnoticed. *Is it novel?* Sequence and structure novelty against reference databases. The aggregate metrics turn the rest into per-task success rates and the Stroma score. The full list, with definitions and how each is dispatched, is in Supplementary C.5.

4.2 Scaffolding settings

Stroma runs models under multiple scaffolding settings defined by two dimensions: the *conditioning mode* (indexed vs. unindexed) and, for all-atom models only, the *motif representation*. A setting is a conditioning protocol, so results are reported per setting and never averaged across them.

Conditioning mode: indexed vs. unindexed. **Indexed** gives the model the sequence index of every motif residue, which is the standard mode in prior benchmarks. **Unindexed** withholds them: the model receives only motif geometry and identity and must discover a viable placement as well as generate around it, which tests whether it can find a compatible structural context rather than complete a partially specified design. Every backbone and sequence model in the benchmark declares indexed only; La-Proteina, RFdiffusion2 and RFdiffusion3 run under both.

Motif representation: all-atom, tip-atom, theozyme-atom. For all-atom models we evaluate up to three motif-conditioning variants: the **all-atom** setting (full heavy-atom coordinates), the **tip-atom** setting (amino acid identity plus the distal side-chain atoms beyond the last rotatable bond), and the **theozyme-atom** setting (its catalytic-subgraph selection: the 1–3-hop heavy-atom subgraph around each ligand-proximal atom, AME tasks only; run by RFdiffusion2, RFdiffusion3 and the two ODesign checkpoints). This dimension does not apply to backbone generating or pLM models. The full per-paradigm applicability matrix (Table S6) and the per-model notes are in Supplementary C.3.

5 Evaluation protocol

5.1 Success criteria

Each design is paired with one or more sequences and each sequence is folded into one or more *rollouts*, where a rollout is one predicted structure together with the design and sequence it came from (Supplementary C.5). A design succeeds if a single rollout satisfies every applicable term below on that one structure:

- **The motif is preserved**: motif RMSD below its criterion's threshold: 1.0 Å on backbone atoms or 1.5 Å on the heavy-atom variants.
- **The fold is confident**: global pLDDT above 70.
- **The scaffold is self-consistent**: for models that emit a design backbone, Cα RMSD between design and prediction below 2.0 Å.

- **The ligand fits**: for ligand-bearing tasks, no clash between non-motif scaffold heavy atoms and the cofolded ligand. Motif residues are ligand-adjacent by definition and are excluded; apo tasks skip this term.

Sequence models emit no design backbone, so the third term cannot be computed for them. The headline criterion of Table 2 therefore drops it for *every* model, leaving motif RMSD, pLDDT and the clash bound: the one test all three paradigms take in the same form. Self-consistency is reported separately as designability (Supplementary C.5).

AME enzyme tasks: theozyme success rate. For the AME enzyme tasks we additionally report a *theozyme criterion*, which replaces the motif RMSD term with one computed over the catalytic atoms alone, at the 1.5 Å heavy-atom cut. It asks a different question from the backbone criterion: a design can place backbone positions correctly while leaving the side chains it must orient wrong. Like every other variant, its terms must pass jointly on one structure. Results accompany the global AME rate in Table S13.

The criteria do not look at the sequence. Motif RMSD, pLDDT and the clash bound are geometric, so a degenerate sequence that folds confidently can satisfy all of them. Across the indexed rows, 11.2% of successful designs contain one amino acid repeated six or more times in the scaffold, a run longer than any chain in the 100 natural proteins these motifs are taken from. Disqualifying those designs leaves the top of the ranking unchanged and lowers every rate. We report this rather than gate on it, because a filter would hide which models produce such sequences. Supplementary D.5 scores the same property continuously, over every design rather than over the repeat-bearing tail.

Sensitivity to the structure predictor. Boltz2 is the predictor for every row of the benchmark, so the ranking could in principle be an artefact of that one choice. To test it, we refolded a fraction of the designs coming from a subset of 7 checkpoints over 20 tasks, with AlphaFold 3 and OpenFold3 as well. While the success rates (SR) vary by up to 3.9 points, the ranking stays unchanged: between-predictor rank correlation is 0.991 to 1.000, while mean success falls from 13.0% under Boltz2 to 10.1% under AlphaFold 3 and 9.1% under OpenFold3.

We keep Boltz2 for three reasons: it co-folds ligands, which 64 of the 100 tasks require; it needs no MSA, which is the regime a *de novo* design is in; and in that regime it reproduces experimental structures far better than the alternatives, with a median Cα RMSD of 2.57 Å against 9.73 and 12.88 Å over the 60 native sequences. Full results are in Supplementary D.3.

5.2 Scoring and novelty

The primary per-task metric is the success rate $\mathrm{SR} = k/N$, where k is the number of passing designs and N the number generated. We report rates with Wilson score 95% confidence intervals. To rank models across the full suite we use the Stroma score, adapted from the MotifBench score (Zheng et al., 2025), which rewards breadth of task coverage over depth. Each successful design is reduced to one representative structure, its highest-pLDDT prediction, and the representatives are clustered with Foldseek-Cluster (van Kempen et al., 2024) at 50% sequence identity and 80% coverage (Supplementary C.6). The number of unique clusters c_i for task i enters:

$$\text{Stroma score} = \frac{1}{T}\sum_{i=1}^{T}\frac{(100+\alpha)\,c_i}{\alpha + c_i}, \quad \alpha = 5. \tag{1}$$

The score ranges from 0 to 100; $\alpha = 5$ places high weight on obtaining a first solution per task. T is the number of tasks scored, so it is the size of whichever task set the column reports over: 89 for the common-subset comparison of Table 2 and the row's own eligible count in the full-coverage table. A Stroma score is therefore only comparable across models scored over the same T, which is the reason the headline table fixes the value at 89. Since both α and the success thresholds are hyperparameters, we sweep each rather than assume the ranking survives them, and we verify it does (Supplementary C.6 and D.4).

We additionally report per-model *scaffold* novelty along two axes: a **structure** axis (Foldseek-Search of the motif-stripped scaffold against PDB100, reported as the mean $(1 - \mathrm{TM\text{-}score})$ across solution

clusters, with TM-score normalised by the design) and a **sequence** axis (MMseqs2 (Hauser et al., 2016) search of the motif-masked scaffold sequence against UniRef30). Novelty is computed only over designs that pass the success criteria, and reported separately from the Stroma score. Database provenance and the algorithm are in Supplementary C.8.

6 BENCHMARKING GENERATIVE MODELS

We evaluate 15 model checkpoints spanning all three paradigms. Table 2 reports one row per checkpoint at its default setting (indexed scaffolding; all-atom conditioning where the checkpoint declares an atom-resolution axis), scored over the 89-task *common subset*. Figure a represents this data colored by source of benchmark. A checkpoint that supports unindexed scaffolding or tip-atom and theozyme-atom conditioning is run under every setting it declares, which expands the 15 checkpoints to the 32-row matrix reported in full in Supplementary D.2, Table S11. Settings are never averaged together: each is a distinct conditioning protocol, so a checkpoint has one row per setting rather than one aggregate number.

RFdiffusion3 ranks first on the 89-task common subset at 25.6% success, followed by La Proteina at 23.1% and PPIFlow at 22.4%; the same three lead by Stroma score (51.6, 47.4 and 43.2, respectively).

Table 2: Cross-model performance on the Stroma *common subset*: the 89 common eligible tasks (single-chain, single-PDB). One row per checkpoint at its default setting, indexed scaffolding and all-atom conditioning; the 15 rows are 12 models, since Protpardelle-1C, ODesign and ESM3 each contribute two checkpoints; the other 16 settings are in Table S11. SR is the mean per-task success rate under the headline criterion of Section 5.1: backbone motif RMSD below 1.0 Å, pLDDT above 70 and, on ligand tasks, no scaffold–ligand clash. 95% CI are Wilson score intervals averaged over tasks. RMSD is the motif RMSD and pLDDT the confidence, both averaged over successful designs only. Novelty is scaffold-only, with the motif excluded before the search: *str* is $1 - \mathrm{TM}$ against PDB100 and *seq* is $1 - (\mathrm{identity} \times \mathrm{coverage})$ against UniRef30, each averaged over solution clusters (Section 5.2). Bold marks the best value in that column.

				Means over successes		Novelty	
Model	SR (%)	95% CI	Stroma	RMSD (Å)	pLDDT	str	seq
Backbone							
PPIFlow	22.4	18.6–27.8	43.2	0.74	89.7	0.39	0.76
Protpardelle-1C backbone	20.4	17.1–26.0	38.5	0.71	87.3	0.35	0.75
RFdiffusion	17.0	14.9–21.4	26.4	0.68	91.3	0.37	0.74
Genie3	16.1	13.4–21.2	36.8	0.74	92.4	0.44	0.75
Genie2	12.1	10.2–16.8	22.0	0.74	88.6	0.42	0.73
All-atom							
RFdiffusion3	**25.6**	21.1–31.9	**51.6**	0.69	87.9	0.42	0.75
La Proteina	23.1	19.1–29.2	47.4	0.70	89.3	0.40	0.76
ODesign-Rigid	17.0	14.0–22.4	29.3	0.68	85.4	0.39	0.74
ODesign-Flex	16.7	14.5–21.4	26.1	0.71	89.2	0.36	0.71
Protpardelle-1C all-atom	14.2	11.6–19.6	31.4	0.72	85.7	0.35	0.76
RFdiffusion2	9.6	7.5–14.6	18.6	0.73	84.3	0.40	0.74
Sequence (pLM)							
ESM3 (seq+struct)	8.0	6.7–12.5	16.7	0.65	81.1	0.30	0.60
ESM3 (seq-only)	5.9	5.0–9.9	9.8	0.61	84.2	0.29	0.56
PoET-2	2.6	2.1–6.6	5.7	0.62	80.1	0.30	0.61
ProGen2 base	1.8	1.5–5.7	4.5	0.54	80.8	0.29	0.64

The model determines the performance. Averaged over the common subset, backbone and all-atom checkpoints are indistinguishable, providing an average of 17.6% SR each, while the sequence models sit at roughly four times lower. The spread inside each family is far larger than the gap between families: for example, RFdiffusion3 and RFdiffusion2 are both all-atom and differ by 16 points. Whether an all-atom model writes its own sequence or receives one by inverse folding makes no measurable difference either (Supplementary D.7). The useful division is therefore not between

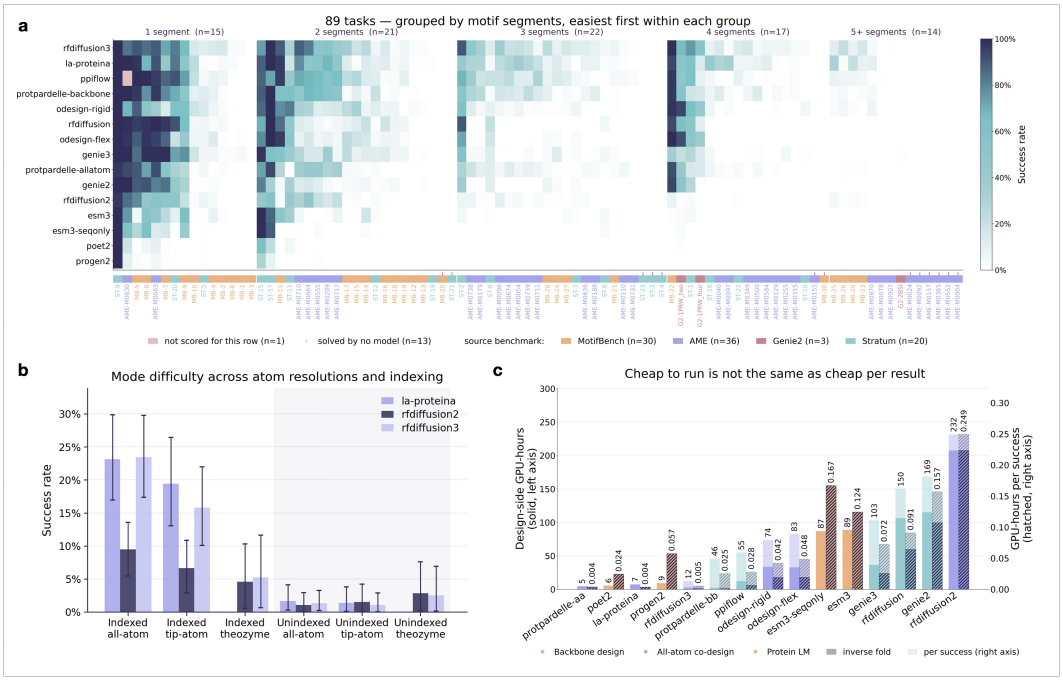


Figure 2: **(a)** Success rate for every task and checkpoint over the 89 common tasks, with columns grouped by motif segment count and sorted easiest-first within each group. Tasks tend to be solved by most models or by none; 13 are solved by none. **(b)** Effect of conditioning mode on the three checkpoints that run all six, with 95% Wilson intervals. Indexing costs more than atom resolution: La Proteina falls from 23.1% to 1.6% between indexed and unindexed all-atom. Each bar uses that row's own eligible tasks (41 to 100), so modes compare within a model, less so across models. **(c)** Design-side GPU-hours, meaning design plus inverse folding (left axis, solid), against GPU-hours per successful design (right axis, hatched); prediction is excluded. The axes are scaled independently, so heights compare within a series only. ProGen2 and PoET-2 are among the cheapest to run and the most expensive per usable design.

architectures but between the two structure paradigms and the sequence one, which trails on every measure we report.

The tasks are the most decisive factors in determining success. What separates results is the task rather than the model (Figure 2a): a task is usually solved by most checkpoints or by none, and 13 of the 89 are solved by none. Those 13 fail for two opposite reasons: seven being enzyme active sites of a few isolated residues spread over four or more segments, with nothing contiguous to build on, and the other six motifs for being so large that little scaffold is left to design. The results highlight two directions in which models can improve - small, non-contiguous scaffolding of many motifs (useful in catalytic-site engineering) as well as large motifs that currently saturate the model's generation length.

Unindexing decreases performance for all compatible models. Three all-atom models run both scaffolding modes: RFdiffusion3, La Proteina and RFdiffusion2. Motif indexing pins the motif residues to specific positions in the chain, telling the model where the motif should sit in the final design. Withdrawing those positions (unindexed scaffolding) lowers the success rate by about an order of magnitude for all three checkpoints (23.5% to 1.3% for RFdiffusion3, Figure 2b). Narrowing the representation of the catalytic-site residues does not lead to such a drastic effect, but it reorders the models: with tip atoms, the side-chain ends involved in the catalytic site, La Proteina leads, while on the catalytic subgraph RFdiffusion3 does (Table S11).

Multi-chain and multi-PDB tasks are more challenging. To investigate task-specific behaviours that might not be fully reflected in the benchmark's final score, we looked more carefully at the

multi-chain and multi-pdb tasks that are expected to be particularly hard to solve. Both ODesign checkpoints do better on the nine multi-chain tasks than on their own native tasks, while RFdiffusion3 falls from 25.6% SR to 2.2% on multi-chain tasks. La Proteina is excluded from this regime as it doesn't allow multi-chain generation. The only two multi-PDB tasks that bring together motifs from different PDB structures are attempted by both RFdiffusion checkpoints (2 and 3) in unindexed mode and none of the 700 designs succeed, emphasizing the complexity of this type of task.

Generation time per successful design. Regarding generation time, Figure 2c shows that while pLMs and some backbone models are cheap to run, they result in designs with lower success rates than more computationally expensive models. This bridges the gap between GPU hours and GPU hours per success. More expensive models are surprisingly less expensive overall as their higher success rate makes up for their slower generation time.

Some models generate the motif mirror images. We also decided to take a closer look at the integrity and validity of the generated designs. We flagged a model-dependent behaviour that is characterized b the presence of mirror images of the motifs in some of the designs. More precisely, 1.35% of all designs and 16.0% of all Genie2 designs harbour motifs made up of D-amino acids (inverted chirality). We hypothesized that this was due to the conditioning function of genie2 that looks to minimize the distance between training examples and generated motifs, not taking chirality into account as RMSD or TM score cannot tell L from D amino acids (Supplementary C.7). Furthermore, two features of the generated population are that (i) Scaffold sequences are systematically less complex than the natural chains the motifs are taken from. Disqualifying the low-complexity designs leaves the ranking essentially unchanged (Supplementary D.5). (ii) Novelty does not rise as success falls: the two structure paradigms (backbone, all-atom) are indistinguishable on both the structural and the sequence axis, and the sequence models are the least novel of the three on both. (Supplementary C.8)

7 Discussion

With this work, our contribution is a ranking of models from different generation paradigms against the same objectives, to help the community determine which model is the best fit for a given task and decide where effort should go next.

The paradigm does not predict the ranking as we expected. The two strongest models are all-atom, yet as category as a whole, all-atom models only match backbone ones, and two backbone models rank third and fourth overall. Only protein language models stand apart: they are not usually trained with a scaffolding objective, and while they can produce successful designs for one- and two-segment motifs, they fail at more challenging tasks. We generated from them with minimal engineering, to avoid biasing them in any form. Our conclusion is that structured-based paradigms are better-suited for the motif scaffolding protein design problem.

What separates the two structural paradigms is capability rather than accuracy. Backbone-only models cannot take atom-level conditioning, therefore not being eligible for the top-atom and theozyme tasks, and they still depend on inverse folding. Advances in latent generative modelling, as in La-Proteina, free all-atom models from both constraints, although that independence does not yet show as higher success.

Stroma is deliberately a complex benchmark, and we show that the most demanding tasks expose the limits of what can be tackled by current models. Thirteen tasks remain unsolved solved by any model, the two multi-PDB tasks by none of the designs made for them, and the benchmark leader solves almost none of the multi-chain tasks that two lower-ranked models solve more often. Motif scaffolding is central to protein design and has been proposed to solve a wide range of problems in the fields (Courbet et al., 2026). Our results point to motif fidelity under increasing complexity as an opportunity for further progress: models place a motif well when it is one contiguous piece, but performance declines when motifs are scattered across several segments, such as in catalytic residues. Stroma provides a unified test to benchmark models toward the designing proteins around tailored, complex functional sites.

Ethics statement

Stroma is built entirely from publicly available experimental structures in the RCSB Protein Data Bank; it involves no human subjects, no personally identifiable information and no newly generated experimental data. All source benchmarks it consolidates are published and publicly released, and their tasks are redistributed with attribution to the originating work.

We note the dual-use character of protein design research. Stroma evaluates the geometric fidelity of computationally designed scaffolds and makes no claim about experimental function; it introduces no new generative capability, and every model it benchmarks is already publicly released. Its intended use is methodological comparison. The benchmark contains no tasks selected for toxicity or pathogenicity: motifs derived from viral or toxin structures (e.g. the SARS-CoV-2 spike receptor-binding motif, the HIV-1 gp41 bundle, three-finger neurotoxins) are included as structurally challenging scaffolding targets and are drawn from long-published structures widely used as design benchmarks.

AI use statement

We used generative AI tools (Anthropic Claude, via Claude Code) as a coding assistant while building the benchmark, and as an aid in drafting this paper. It was *not* used for research ideation, experimental design or literature search. AI-assisted code was reviewed by the authors and is released in full; per-task run manifests record the checkpoint, scaffold lengths and motif bridge behind each result, so every reported number traces to the run that produced it. All numerical results come from executing that pipeline and none from a language model. All AI-assisted text was reviewed and edited by the authors, and we take responsibility for the final content of this work.

Reproducibility statement

The benchmark, the evaluation pipeline and the configuration of every run reported here are released together. Each of the 100 tasks is a versioned JSON record (motif coordinates, residue identities, atom-level views, scaffold length range and provenance to its source benchmark) generated by a script from the deposited PDB entry rather than edited by hand, so the task definitions can be rebuilt from the source structures. Each model is described by a JSON declaration of its capabilities, parameters and container environment, and each metric by a JSON task definition, so the routing and scoring decisions in Section 4.1 and Section 4.1 are read off the released configuration rather than reimplemented from prose.

The campaign reported in Section 6 is fully specified by its released run configuration: $N = 100$ designs per task, $M = 2$ inverse-folded sequences per design and $K = 2$ predictions per sequence, under the global cap length policy with a 256-residue cap applied identically to every model, with Boltz-2 as the structure predictor. Per-task run manifests record the checkpoint that served each task, the sampled and realised scaffold lengths, and which motif bridge was written, so a reported number can be traced to the run that produced it.

Acknowledgments

[Omitted for anonymous submission.]

References

Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M. Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, Magnus S. Bauer, Regina Barzilay, Tommi S. Jaakkola, Rohith Krishna, and David Baker. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods*, 23(1):96–105, January 2026. ISSN 1548-7105. doi: 10.1038/s41592-025-02975-x.

Mubashara Akhtar, Omar Benjelloun, Costanza Conforti, Pieter Gijsbers, Joan Giner-Miguelez, Nitisha Jain, Michael Kuchnik, Quentin Lhoest, Pierre Marcenac, Manil Maskey, Peter Mattson, Luis Oala, Pierre Ruyssen, Rajat Shinde, Elena Simperl, Goeffry Thomas, Slava Tykhonov, Joaquin Vanschoren, Steffen Vogler, and Carole-Jean Wu. Croissant: A Metadata Format for ML-Ready Datasets. In *Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning (DEEM '24)*, pp. 1–6, 2024. doi: 10.1145/3650203.3663326.

Alexander J. Barnett, Rajendra KC, Pratikshya Pandey, Pamodha Somasiri, Kirsten A. Fairfax, Sandy Hung, and Alex W. Hewitt. Benchmarking generative AI protein models reveals differences between structural and sequence-based approaches. *Genomics, Proteomics & Bioinformatics*, 24(2), February 2026. doi: 10.1093/gpbjnl/qzag014.

Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo Design of All-atom Biomolecular Interactions with RFdiffusion3, September 2025. ISSN 2692-8205.

Alexander E. Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W. Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27):e2311500121, July 2024. doi: 10.1073/pnas.2311500121.

Alexis Courbet, David Juergens, Xinru Wang, Thomas Schlichthaerle, Mohamad Abedi, Robert Ragotte, Linna An, Indrek Kalvet, Sam Pellock, Ljubica Mihaljevic, Cameron Glasscock, Arvind Pillai, Adam Broerman, Nathan Ennist, Ella Haefner, Nora McNamara-Bordewick, Ian Haydon, Lance Stewart, Gaurav Bhardwaj, and David Baker. The past, present and future of de novo protein design. *Nature*, 652(8112):1139–1152, April 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10328-7.

Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, 22(4): 717–723, April 2025. doi: 10.1038/s41592-025-02626-1.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, November 2021. doi: 10.1145/3458723.

Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang, Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching, July 2025.

Maria Hauser, Martin Steinegger, and Johannes Söding. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, 32(9):1323–1330, May 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw006.

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, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew 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, February 2025. doi: 10.1126/science.ads0018.

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2, May 2024.

Yeqing Lin, Minji Lee, Aakarsh Vermani, Ellena Jiang, Sebastiaan De Cooman, Matej Špeťko, and Mohammed AlQuraishi. Fast and ultra-capable protein design: Advancing the frontier through atomistic SE(3)-equivariance with Genie 3. *bioRxiv*, May 2026. doi: 10.64898/2026.05.01.722168.

Tianyu Lu, Richard Shuai, Petr Kouba, Zhaoyang Li, Yilin Chen, Akio Shirali, Jinho Kim, and Po-Ssu Huang. Conditional Protein Structure Generation with Protpardelle-1C, August 2025. ISSN 2692-8205.

Erik Nijkamp, Jeffrey A. Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. ProGen2: Exploring the boundaries of protein language models. *Cell Systems*, 14(11):968–978, November 2023. doi: 10.1016/j.cels.2023.10.002.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction, June 2025.

Timothy Fei Truong, Jr. and Tristan Bepler. Understanding protein function with a multimodal retrieval-augmented foundation model. arXiv:2508.04724, August 2025.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, February 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0.

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, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8.

Qilin Yu, Liangyue Guo, Xiayan Qin, Xikun Huang, Baihui Tian, Hongzhun Wang, Yu Liu, Yunzhi Lang, Di Wang, Zhouhanyu Shen, Jie Lin, and Mingchen Chen. High-affinity protein binder design via flow matching and in silico maturation. bioRxiv 2026.01.19.700484, January 2026.

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, October 2025.

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, February 2025.

A RELEASE AND LICENSING

What is released. Four artifacts are published together: (i) **the Stroma task set**, one versioned JSON record per task carrying motif coordinates, residue identities, the atom-level views of Section 3.2, the scaffold length range and provenance to the source benchmark, plus the `data/tasks.csv` table of Table S2; (ii) **the evaluation code**, comprising the orchestrator, the model and metric descriptors, the adapters and the analysis scripts, which is what makes the routing and scoring decisions in this paper readable off configuration rather than reimplemented from prose; (iii) **a Datasheet for Datasets** (Gebru et al., 2021), covering motivation, composition, collection, preprocessing, uses, distribution and maintenance; and (iv) **a Croissant metadata file** (Akhtar et al., 2024), so the task set is machine-readable by dataset tooling and its fields are discoverable without reading the schema documentation.

Licensing. The task records, the derived motif views and the accompanying documentation are released under **CC BY 4.0**. The evaluation code is released under the **MIT** licence. Both permit redistribution and derivative work with attribution, which is the condition a benchmark needs if later papers are to extend the task set. The underlying experimental structures are not redistributed as primary data: every task names its RCSB PDB entry and is regenerated from that entry by script, so the coordinates carry the PDB's own terms of use and the release contains derived annotations rather than a mirror of the archive. The task definitions inherit their terms from the benchmarks they come from, all of which permit redistribution with attribution: MotifBench's specification is published under **CC BY 4.0**, the AME active sites are drawn from M-CSA (**CC BY 4.0**) as assembled in the RFdiffusion2 repository, and the Genie2 multi-motif cases come from that project's **Apache-2.0** repository. Each task record names its source benchmark and source identifier, and the release carries the upstream copyright notices and citations. Model weights are likewise not redistributed; each model descriptor records the upstream source and that model's own licence, several of which are non-commercial or gated and cannot be mirrored.

Access. The code and the task dataset can be found at `https://github.com/plainwater27/stroma`. The dataset, and its datasheet and Croissant file, will be hosted on HuggingFace on the release is public.

B DATASET DOCUMENTATION

Table S1: Multi-resolution task specifications in Stroma, their relationship to benchmark columns, and what each one withholds. Amino acid identity and, in indexed scaffolding, sequence position accompany every row; unindexed scaffolding withholds the position. Coordinates are copied verbatim from the deposited structure, with no superposition, renumbering or idealisation at specification time, so what distinguishes the rows is the atom selection rather than any transformation of the coordinates.

Representation	Atoms included	Source column	Withheld	Supports
Residue-level	Identity only	`motif_aa_identity`	All coordinates	Sequence language models
Backbone-level	N, Cα, C, O	`motif_atoms`	Side-chain coordinates	Backbone design models
All-atom	All heavy atoms	`motif_atoms`	Hydrogens, unresolved in the source	All-atom generative models[a]
Tip-atom	Distal side-chain	`motif_tip_atoms`	The residue's other atoms, backbone included	All-atom models (reduced)
Theozyme	Catalytic subgraph	`motif_theozyme_atoms`	The residue's other atoms	All-atom models (41 AME tasks)

B.1 TASK SCHEMA DEFINITION

Each task in Stroma is described by 22 fields (Table S2). The schema is designed to be self-contained: a single TSV row provides all information needed to condition any of the three generative paradigms, evaluate the resulting designs, and trace the task back to its experimental source.

Stroma provides formal JSON Schema definitions for all three schema types in `schemas/`: `motif.schema.json` (required motif geometry, provenance, and constraint fields), `model.schema.json` (model capabilities, pipeline stages, and environment), and `task.schema.json` (metric definitions, dispatch types, and output specifications).

Table S2: Stroma task schema (22 columns). Columns are grouped by function. Atom-level columns (marked $\star$) are computed automatically by the annotation pipeline. `num_segments` and `segment_residues` (marked §) are derived from `motif_residues`. `length_range` and `length_rationale` (marked ‡) are derived deterministically from per-source rules (Section B.2). All other columns are specified in the input task definitions.

Column	Description
Task identification	
`task_id`	Unique identifier (`MB-`, `AME-`, `G2-`, `ST-` prefixes)
`source_benchmark`	MotifBench, AME, Genie2, or Stratum
`source_id`	Original identifier in the source benchmark
`pdb_id`	4-character RCSB PDB accession (semicolon-separated for multi-PDB tasks)
Motif specification	
`motif_residues`	Motif residue ranges: `ChainStart-End;...`
`segment_residues`§	Contiguous-segment form: adjacent ranges on the same chain merged
`num_motifs`	Number of functionally independent substructures (range: 1–4; curated)
`num_segments`§	Number of truly discontiguous sequence stretches (range: 1–8; derived from `motif_residues`)
`motif_residue_count`	Total residues in the motif (range: 1–99)
`motif_atoms`*	All heavy atoms per motif residue
`motif_tip_atoms`*	Distal side-chain atoms past the last rotatable bond (PDB-derived for all tasks)
`motif_theozyme_atoms`*	Theozyme-style 1–3-hop atom subgraph around the ligand-proximal atom (AME tasks only; blank for all others)
`motif_aa_identity`*	Three-letter amino acid identity per motif residue
Design parameters	
`source_length`	Scaffold length as reported in the source benchmark (scalar for MotifBench, `"95-185"` for AME, range for Genie2, blank for Stratum)
`length_range`‡	Total scaffold length range `"min-max"` (inclusive); the downstream protocol samples one length uniformly per design
`length_rationale`‡	Short text explaining which source-rule produced the range; for Stratum, a per-chain breakdown
`redesign_positions`	Motif residues eligible for sequence redesign
`ligands`*	Cofactors/ligands within 4.5 Å of motif
Metadata	
`description`	Short protein/motif name
`protein_type`	Functional category (26 categories)
`justification`	Biological rationale for inclusion
`reference`	Primary literature citation

B.2 ANNOTATION PIPELINE

A design goal of Stroma is that each task is fully specified at multiple resolutions (residue-identity-level, backbone-atomistic-level, and all-atomistic). This has been accomplished in the following way.

Atom-level annotation. Starting from a task definition file that specifies each task at the residue level (task identifier, PDB accession, motif residue ranges, protein type, and metadata), we compute four atom-level columns per task:

1. **`motif_atoms`**: all heavy atoms (non-hydrogen) for each motif residue, extracted from ATOM records in the PDB file. Atoms are ordered backbone-first (N, Cα, C, O) then alphabetically by atom name. Format: `ChainResnum:[N,CA,C,O,CB,...];...`
2. **`motif_tip_atoms`**: the subset of side-chain atoms distal to the final rotatable bond for each motif residue, determined by a static lookup table over the 19 non-glycine standard amino acid types, taken from La Proteina (Geffner et al., 2025) and reproduced in Table S3. Glycine is excluded, having no side chain. This column is derived directly from PDB structures for *all* tasks, including AME, ensuring a single, unambiguous meaning across the benchmark. It enables tip-atom motif conditioning (Section 4.2), where only the functionally most informative atoms are provided to the model.
3. **`motif_theozyme_atoms`**: a theozyme-style atom subgraph from the AME source benchmark (AME tasks only; blank for all non-AME tasks). For each motif residue the selection is the 1–3-hop subgraph of heavy atoms around the atom closest to the ligand, preserving the curated AME atom selections verbatim. This column captures the original benchmark's focus on functionally critical side-chain positioning in enzyme active sites.
4. **`motif_aa_identity`**: three-letter amino acid codes for each observed motif residue, resolved from the PDB structure for all tasks. Format: `ChainResnum:AAA;...`
5. **`ligands`**: PDB HET codes for small-molecule ligands and cofactors within 4.5 Å of any motif heavy atom. Common crystallisation artifacts (water molecules, buffer salts, cryoprotectants, glycosylation sugars, and modified amino acids are filtered automatically. This column supports the ligand clash evaluation criterion (Section 5.1): 64 of 100 tasks include at least one ligand.

Table S3: The tip-atom lookup: for each amino acid, the side-chain atoms after the final rotatable bond. Reproduced from Appendix F of La Proteina (Geffner et al., 2025), which follows the task definition of Protpardelle (Chu et al., 2024), and matches their released implementation atom for atom. These are the atoms handed to the model under tip-atom conditioning (Section 4.2) and the atoms over which `motif_rmsd_tip` is computed. GLY has no entry, so a glycine motif residue contributes nothing to a tip-atom selection. Four residues (ALA, CYS, SER, and THR) carry their Cα; PRO includes in addition its backbone N because its side chain forms a ring with the backbone nitrogen.

Residue	Tip atoms	Residue	Tip atoms
ALA	CA, CB	MET	CG, SD, CE
ARG	CD, NE, CZ, NH1, NH2	PHE	CB, CG, CD1, CD2, CE1, CE2, CZ
ASN	CB, CG, OD1, ND2	PRO	N, CA, CB, CG, CD
ASP	CB, CG, OD1, OD2	SER	CA, CB, OG
CYS	CA, CB, SG	THR	CA, CB, OG1, CG2
GLN	CG, CD, OE1, NE2	TRP	CB, CG, CD1, CD2, NE1, CE2, CE3, CZ2, CZ3, CH2
GLU	CG, CD, OE1, OE2		
GLY	—	TYR	CB, CG, CD1, CD2, CE1, CE2, CZ, OH
HIS	CB, CG, ND1, CD2, CE1, NE2	VAL	CB, CG1, CG2
ILE	CB, CG1, CG2, CD1		
LEU	CB, CG, CD1, CD2		
LYS	CE, NZ		

AME task handling. For AME tasks, `motif_atoms` and `motif_tip_atoms` are computed identically to all other tasks: all heavy atoms are extracted from PDB ATOM records, and tip atoms are derived from the standard residue-type lookup applied to those PDB atoms. This gives every column a single, unambiguous meaning across the full benchmark.

The original AME curated atom selections, a 1–3-hop subgraph around the atom closest to the ligand, reflecting the benchmark's focus on functionally critical side-chain geometry, are preserved verbatim

in the dedicated `motif_theozyme_atoms` column, which is populated for AME tasks only and blank for all non-AME tasks. Ligand annotations for AME tasks are likewise preserved from the source data. Amino acid identities are resolved from PDB structures for all tasks, including AME.

Scaffold generation length range. Rather than committing each task to a single scalar generation length, Stroma emits a *range* `length_range` (`"min-max"`) from which the downstream evaluation protocol samples one target length uniformly per design. Sampling per design (rather than fixing one value per task) keeps every model on the same target distribution.

The ranges are produced deterministically for each source benchmark, with a single rule per source:

- **AME** (41 tasks): the fixed range `95-185` for every task, matching the RFdiffusion design range from the AME source benchmark (Ahern et al., 2026). Every prior paper evaluates AME against this range.
- **MotifBench** (30 tasks): `[`$\lfloor 0.8 \cdot s \rfloor$`, `$\lceil 1.2 \cdot s \rceil$`]` where s is the published `source_length` (scaffold length plus motif residues).
- **Genie2** (6 tasks): the configuration-specified range from the source benchmark (Genie2 supplies a per-task range, not a scalar).
- **Stratum** (23 tasks): `[`$\lfloor 0.8 \cdot c \rfloor$`, `$\lceil 1.2 \cdot c \rceil$`]` where the centre c is the sum across every motif-bearing chain of that chain's resolved residue count in the reference PDB (its natural-fold context), *excluding crystallization-fusion partners*.

A universal floor raises `length_min` to `motif_residue_count + 10` when the rule would otherwise leave fewer than ten scaffold residues for linkers and termini; this case is recorded explicitly in `length_rationale` and applies to tasks whose motif is large relative to the reference chain (e.g. ST-5 Zif268, ST-14 collagen, MB-25 Kemp eliminase). Across the 100 tasks, range midpoints span 34–299 residues (Figure S1d).

B.3 New task curation

This section expands Section 3.1: the per-category breakdown of the 23 new `ST-` tasks, the motif curation procedure, and the full category distribution of the assembled 100-task benchmark.

The 23 new tasks span the following categories (number of tasks in parentheses):

- **Membrane proteins** (2): β_2-adrenergic receptor orthosteric pocket (2RH1) and A_{2A} adenosine receptor ligand-binding site (3EML). These tasks present unique scaffolding challenges due to hydrophobic transmembrane segments and are relevant to drug design.
- **Nanobodies** (2): the camel VHH single-domain antibody cAb-Lys3 (1MEL) and the anti-TAFI nanobody VHH-a204 (5HVG), both classified as nanobodies (VHH single-domain antibodies). CDR loop scaffolding is a common design objective in therapeutic protein engineering.
- **Nucleic acid-binding proteins** (1): Zif268 three-finger zinc finger DNA-binding domain (1AAY). Motif geometry must accommodate nucleic acid backbone contacts.
- **Chaperones** (1): Hsp90 N-terminal ATP-binding domain (1YER).
- **Signaling proteins** (2): H-Ras GTPase nucleotide-binding switch (P-loop, switch I, switch II; 5P21) and CDK2–cyclin A catalytic loops (1FIN).
- **Repeat proteins** (1): designed ankyrin repeat protein NI3C (2XEE); the motif is the three internal ankyrin repeats, testing the model's ability to scaffold modular, repetitive structural elements.
- **Viral proteins** (2): HIV-1 gp41 six-helix bundle hydrophobic pocket (1AIK) and SARS-CoV-2 spike RBD receptor-binding motif (6M0J).
- **Other categories** (12): PPI hub (1YCR), hormone (3HHR), antimicrobial peptide (1DFN), PTM reader (2LVM), de novo designed proteins (3V1C, 4BS0), structural protein (1K6F), photosynthetic protein (1JB0), metalloprotein (1E5Y), electron carrier (1FHH), immune protein (CTLA-4 checkpoint receptor, 3OSK), and toxin (α-bungarotoxin three-finger neurotoxin, 2QC1).

Literature-grounded motif curation. The `ST-` motifs were re-curated directly from the deposited coordinates and primary literature for each entry: catalytic residues, metal-coordination spheres, CDR loops and binding-interface residues were verified against the structure (e.g. contact distances to ligands, metals or partner chains) and the original publication.

Table S4: Stroma task categories. Categories with ≤ 2 tasks are grouped under "Other" for conciseness.

Category	Tasks	Representative motif types
Enzyme	51	Catalytic triads, metal coordination, cofactor binding
Viral protein	7	Receptor-binding motifs, fusion helices
Immune protein	6	Checkpoint receptors (PD-1, CTLA-4), cytokine-receptor interfaces
De novo designed protein	3	Symmetric interfaces, designed active sites
De novo designed enzyme	3	Kemp eliminase, retro-aldolase active sites
Electron carrier	3	[Fe-S] clusters, cytochrome heme sites
Fungal effector	3	Secreted virulence motifs
Signaling protein	3	GTPase switch regions, kinase activation loops
Calcium-binding protein	2	EF-hand motifs
Membrane protein	2	GPCR ligand-binding pockets
Chaperone	1	N-terminal ATP-binding pocket
Nucleic acid-binding	1	Zinc finger DNA-recognition helices
Repeat protein	1	Ankyrin repeats
Other (13 categories)	14	Nanobody CDR, hormone, AMP, toxin, metalloprotein, ...
Total	**100**	

B.4 Full task list

Table S5 lists all 100 Stroma tasks grouped by source benchmark. Columns are: unique task identifier, PDB accession code, short description, number of discontinuous sequence segments (`num_segments`), segment residue ranges, and the scaffold-length range (`length_range`, format `"min-max"`) from which the downstream protocol samples one target length per design. Multi-PDB tasks (Genie2) show semicolon-separated accession codes in the PDB column and PDB-prefixed ranges in the segment column. Full atom-level specifications and literature references are provided in the accompanying repository.

Table S5: Complete Stroma task list (100 tasks). Columns: task identifier, PDB accession, description, number of discontinuous segments, segment residues, and `length_range` (`"min-max"`, residues; the downstream protocol samples one target length uniformly from this range per design).

Task	PDB	Description	N. Seg.	Segment residues	Length range
MotifBench					
MB-1	1LDB	Lactate dehydrogenase	1	A186-206	116-176
MB-2	1ITU	Renal dipeptidase	1	A124-147	139-209
MB-3	2CGA	Strained chymotrypsinogen loop that undergoes conformation change	1	A184-194	108-164
MB-4	5WN9	RSV G-protein 2D10 site	1	A170-189	76-114
MB-5	5ZE9	P-loop	1	A229-243	92-138
MB-6	6E6R	Ferredoxin Protein	1	A25-35	68-104
MB-7	6E6R	Ferredoxin Protein	1	A25-35	168-254
MB-8	7AD5	Orphan protein	1	A99-113	112-168

(continued on next page)

(continued from previous page)

Task	PDB	Description	Seg.	Segment residues	Length range
MB-9	7CG5	Orphan protein	1	A6-20	112-168
MB-10	7WRK	Orphan protein	1	A80-94	112-168
MB-11	3TQB	Parallel beta strand and loop	2	A37-51;A65-79	124-186
MB-12	4JHW	RSV F-protein Site 0	2	F63-69;F196-212	99-149
MB-13	4JHW	RSV F-protein Site 0	2	F63-69;F196-212	179-269
MB-14	5IUS	PD-L1 binding interface on PD-1	2	A63-82;A119-140	113-171
MB-15	7A8S	Orphan protein	2	A41-55;A72-86	104-156
MB-16	7BNY	Orphan protein	2	A83-97;A111-125	124-186
MB-17	7DGW	Orphan protein	2	A22-36;A70-84	124-186
MB-18	7MQQ	Orphan protein	2	A80-94;A115-129	104-156
MB-19	7MQQ	Orphan protein	2	A80-94;A115-129	184-276
MB-20	7UWL	IL17-RA interface to IL17-RB	2	E63-73;E101-111	157-237
MB-21	1B73	Glutamate racemase active site	3	A7-8;A70;A178-180	104-158
MB-22	1BCF	Di-iron binding motif	4	A18-25;A47-54;A92-99;A123-130	125-189
MB-23	1MPY	Catechol dioxygenase active site	6	A153;A199;A214;A246;A255;A261	104-158
MB-24	1QY3	GFP pre-cyclized state	3	A58-71;A96;A222	192-290
MB-25	2RKX	De novo designed Kemp eliminase	8	A9-11;A48-50;A101;A128;A169;A176;A201;A222-224	191-287
MB-26	3HOJ	De novo designed retroaldol enzyme	8	A51-53;A81;A110;A131;A159;A180-184;A210-211;A231-233	173-261
MB-27	4XOJ	Trypsin catalytic triad and oxyanion hole	3	A55;A99;A190-192	124-186
MB-28	5YUI	Carbonic anhydrase active site	3	A93-97;A118-120;A198-200	68-104
MB-29	6CPA	Carboxypeptidase active site	5	A69-72;A127;A196;A248;A270	166-250
MB-30	7UWL	IL17-RA interface to IL17-RB	4	E63-73;E101-111;E132-142;E165-174	174-262
AME					
AME-M0024	1NZY	4-chlorobenzoyl-CoA dehalogenase	6	A64;A86;A90;A114;A137;A145	95-185
AME-M0040	13PK	phosphoglycerate kinase	4	A39;A219;A376;A399	95-185
AME-M0050	1DBT	orotidine-5'-phosphate decarboxylase	6	A33;A60;A62;A123;A215;B65	95-185
AME-M0054	1QFE	3-dehydroquinate dehydratase (type 1)	3	A86;A143;A170	95-185
AME-M0058	1CJU	adenylate cyclase	4	A396-397;A440;B1029;B1065	95-185

(continued on next page)

(continued from previous page)

Task	PDB	Description	Seg.	Segment residues	Length range
AME-M0078	1AL6	citrate (Si)-synthase	5	A244;A274;A320;A329;A375	95-185
AME-M0092	1DLI	UDP-glucose 6-dehydrogenase	6	A118;A145;A204;A208;A260;A264	95-185
AME-M0093	1DQA	hydroxymethylglutaryl-CoA reductase (NADPH)	4	A691;A767;B559;B866	95-185
AME-M0096	1CHM	creatinase	3	A232;A262;A358	95-185
AME-M0097	1CTT	cytidine deaminase	4	A102;A104;A129;A132	95-185
AME-M0110	1C0P	D-amino-acid oxidase	3	A1054;A1335;A1339	95-185
AME-M0129	1OS7	taurine dioxygenase	4	A99;A101;A255;A270	95-185
AME-M0151	1Q0N	2-amino-4-hydroxy-6-hydroxymethyldihydropteridinediphosphokinase	4	A82;A92;A95;A97	95-185
AME-M0157	1QH5	hydroxyacylglutathione hydrolase	6	A54;A56;A58-59;A110;A134;A173	95-185
AME-M0179	1Q3S	chaperonin ATPase	3	A64;A97-98;A393	95-185
AME-M0188	1XEL	UDP-glucose 4-epimerase	3	A124;A149;A153	95-185
AME-M0209	1LIJ	adenosine kinase	2	A136;A318	95-185
AME-M0255	1MG5	alcohol dehydrogenase (SDRtype)	4	A108;A139;A152;A156	95-185
AME-M0315	1EY3	enoyl-CoA hydratase	4	A98;A141;A144;A164	95-185
AME-M0349	1E3V	delta 5-3-ketosteroid isomerase	4	A16;A40;A100;A103	95-185
AME-M0365	1PFK	Phosphofructokinase I	6	A11;A72;A125;A127;A129;A171	95-185
AME-M0375	4TS9	Purine nucleoside phosphorylaseDeoD-type	7	A20;A24;A87;A90;A204;A217;C43	95-185
AME-M0500	1E3I	alcohol dehydrogenase (class II)	4	A46-48;A67;A178;A182	95-185
AME-M0552	1FGH	aconitate hydratase	5	A100-101;A165;A447;A642;A644	95-185
AME-M0555	1F8R	L-amino-acid oxidase	2	A223;A326	95-185
AME-M0584	1LDM	L-lactate dehydrogenase	4	A106;A166;A169;A193	95-185
AME-M0630	1J79	dihydroorotase	1	A250	95-185
AME-M0636	1UAQ	cytosine deaminase (yeast)	3	A64;A89;A91	95-185
AME-M0663	1RK2	ribokinase	1	A252-255	95-185
AME-M0664	2DHN	dihydroneopterin aldolase	2	A22;A100	95-185
AME-M0674	1UF7	N-carbamoyl-D-amino-acid hy-drolase	3	A46;A126;A171	95-185
AME-M0710	1RA0	cytosine deaminase (bacterial)	2	A156;A217	95-185
AME-M0711	2ESD	glyceraldehyde-3-phosphate dehydrogenase (NADP+)	3	A154;A250;A284	95-185
AME-M0717	1X7D	ornithine cyclodeaminase	2	A56;A228	95-185
AME-M0731	1MT5	fatty acid amide hydrolase	3	A142;A217-218;A238-241	95-185
AME-M0732	1XS1	dCTP deaminase	5	A124;A126;A138;C111;C115	95-185

(continued on next page)

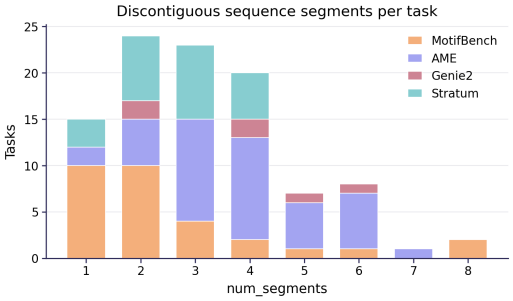
(continued from previous page)

Task	PDB	Description	Seg.	Segment residues	Length range
AME-M0738	1O98	phosphoglycerate mutase(2,3-bisphosphoglycerate-independent)	3	A62;A154;A261	95-185
AME-M0739	1KNP	L-aspartate oxidase	3	A290;A351;A386	95-185
AME-M0870	1OH9	acetylglutamate kinase	5	A8;A11;A45;A162;A217	95-185
AME-M0904	1QGX	3'(2'),5'-bisphosphate nucleotidase	6	A49;A72;A142;A144-145;A147;A294	95-185
AME-M0907	1RBL	ribulose-bisphosphate carboxylase (type I)	5	A175;A201;A203-204;A294;A334	95-185
Genie2					
G2-4JHW+5WN9	4JHW;5WN9	Two epitopes	2	4JHW/F254-278;5WN9/A170-189	85-175
G2-1PRW_two	1PRW	Two 4-helix bundles	4	A16-35;A52-71;A89-108;A125-144	120-200
G2-1PRW_four	1PRW	Four EF-hands	4	A21-32;A57-68;A94-105;A125-144	88-163
G2-3BIK+3BP5	3BIK;3BP5	Two PD-1 binding motifs	2	3BIK/A121-125;3BP5/B110-114	30-60
G2-3NTN	3NTN	Two 3-helix bundles	6	A342-348;A367-372;B342-348;B367-372;C342-348;C367-372	89-109
G2-2B5I	2B5I	Two binding sites	5	A11-23;A35-45;A61-72;A81-95;A119-133	116-166
Stratum					
ST-1	2RH1	Beta-2 adrenergic receptor orthosteric pocket (TM3/TM5/TM6/TM7 segments)	4	A109-118;A193-207;A286-293;A308-316	225-339
ST-2	3EML	A2A adenosine receptor orthosteric pocket (TM5/TM6/TM7 segments)	3	A168-177;A246-253;A264-274	230-346
ST-3	1MEL	Camel VHH single-domain antibody (cAb-Lys3) CDR1/CDR2/CDR3 paratope	3	A26-35;A50-58;A95-107	105-159
ST-4	5HVG	Anti-TAFI nanobody VHH-a204 CDR1/CDR2/CDR3 paratope	3	B26-35;B50-58;B96-112	97-147
ST-5	1AAY	Zif268 (EGR1) three-finger Cys2His2 zinc-finger DNA-binding domain	1	A103-187	95-102
ST-6	1YER	Hsp90 N-terminal domain ATP/nucleotide-binding pocket (Bergerat fold)	3	A51-58;A93-98;A136-140	165-249
ST-7	1YCR	MDM2 p53-binding hydrophobic cleft	3	A50-62;A72-75;A93-104	68-102

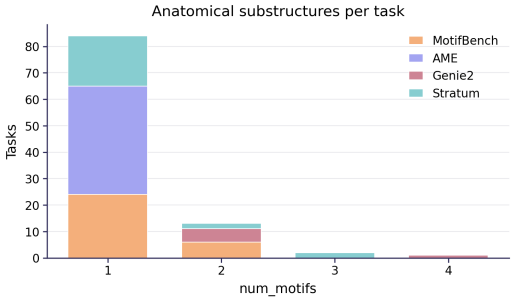
(continued on next page)

(continued from previous page)

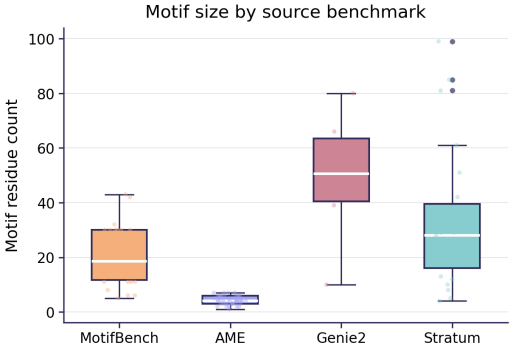
Task	PDB	Description	Seg.	Segment residues	Length range
ST-8	5P21	H-Ras GTPase nucleotide-binding switch (P-loop, switch I, switch II)	3	A10-17;A30-38;A57-67	132-200
ST-9	2XEE	DARPin NI3C internal repeat binding modules (three internal ankyrin repeats)	1	A46-144	125-189
ST-10	3HHR	Human growth hormone Site 1 receptor-binding epitope (helix 1, loops, helix 4)	4	A18-25;A41-48;A61-68;A164-190	147-221
ST-11	1DFN	Human alpha-defensin HNP-3 disulfide-stabilized amphiphilic beta-sheet	2	A4-10;A18-31	31-36
ST-12	2LVM	53BP1 tandem Tudor methyl-lysine aromatic cage (H4K20me2 reader)	2	A1495-1502;A1519-1523	98-148
ST-13	3V1C	De novo C2-symmetric Zn-binding homodimer MID1 (interfacial His3 sites)	4	A12-16;A39;B12-16;B39	74-112
ST-14	1K6F	Collagen-like (Pro-Pro-Gly)10 triple-helix (three staggered chains)	3	A1-27;B1-27;C1-27	91-105
ST-15	1JB0	Photosystem I PsaC two [4Fe-4S] clusters FA and FB (eight Cys ligands)	2	C10-20;C47-57	64-96
ST-16	1E5Y	Azurin type-1 (blue) copper site (His46/Cys112/His117/Met121 + Gly45)	4	A45-46;A112;A117;A121	102-154
ST-17	1FHH	Rubredoxin [Fe(S-Cys)4] iron site (Cys6/Cys9/Cys39/Cys42)	2	A6-9;A39-42	42-64
ST-18	1AIK	HIV-1 gp41 six-helix bundle hydrophobic pocket (fusion-inhibitor target)	2	C628-635;N565-581	56-84
ST-19	6M0J	SARS-CoV-2 spike RBD receptor-binding motif (ACE2 interface)	2	E417;E446-505	155-233
ST-20	3OSK	CTLA-4 immune-checkpoint receptor B7-binding MYPPPY/FG-loop epitope	1	A97-106	96-146
ST-21	1FIN	CDK2 ATP-binding Gly-rich (P-)loop and activation (T-)loop	2	A10-18;A145-172	238-358
ST-22	4BS0	De novo Kemp eliminase HG3.17 catalytic tetrad (Asp127/Gln50/Cys84/Trp44)	4	A44;A50;A84;A127	239-359
ST-23	2QC1	Alpha-bungarotoxin three-finger neurotoxin nAChR-binding face	3	A6-11;A27-42;A68-71	59-89



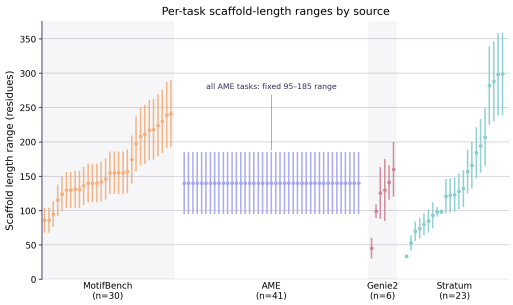
(a) Segment complexity: distribution of discontinuous segments (`num_segments`) per task, coloured by source. Range: 1–8.



(b) Motif complexity: distribution of functionally independent substructures (`num_motifs`) per task, coloured by source. Range: 1–4.



(c) Motif size by source: residue counts range from 1 to 99. New tasks extend the range with large repeat-protein motifs.



(d) Scaffold-length range per task, ordered within each source. Every AME task carries the same fixed 95–185 range.

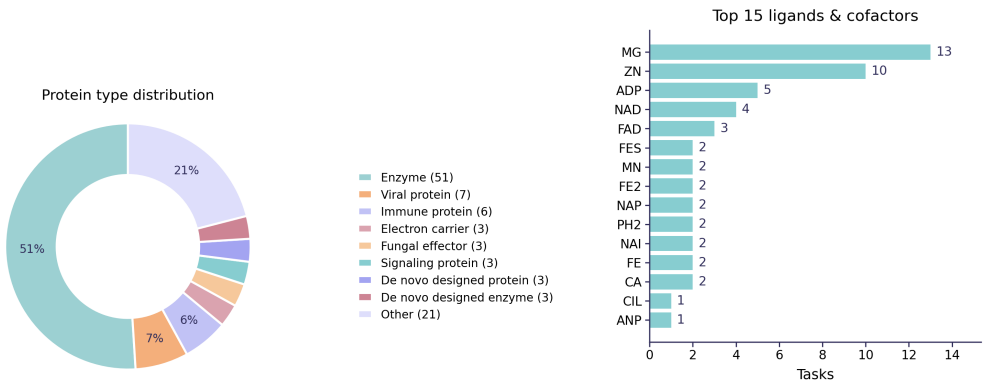
Figure S1: Structural properties of Stroma tasks. **(a)** Discontiguous-segment count per task (`num_segments`; range 1–8) measures the number of independent insertion gaps the model must fill. **(b)** Functionally independent substructure count (`num_motifs`; range 1–4); a single enzyme active site spread across many segments counts as one substructure. **(c)** Motif residue count distribution across source benchmarks (range 1–99, median 11, mean 19). **(d)** Per-task scaffold-length range. Each interval shows the permitted range of total design lengths for one task, with its midpoint indicated. They are grouped by source: each source keeps its own convention, from AME's single fixed range to the per-task ranges of the other three.

C Protocol Details

C.1 Pipeline Routing Detail

This section expands the pipeline routing summarised in Section 4.1.

The pipeline classifies each model by output modality and routes it through the appropriate stages: **backbone models** (RFdiffusion (Watson et al., 2023), Protpardelle-1C cc58 (Lu et al., 2025), PPIFlow (Yu et al., 2026), Genie2 (Lin et al., 2024), Genie3 (Lin et al., 2026)) follow design → inverse fold (LigandMPNN (Dauparas et al., 2025)) → predict (Boltz2 (Passaro et al., 2025)) → evaluate; **all-atom models** (RFdiffusion2 (Ahern et al., 2026), RFdiffusion3 (Butcher et al., 2025), La Proteina (Geffner et al., 2025), Protpardelle-1C cc91 (Lu et al., 2025)) and **protein language models** (PoET-2 (Truong & Bepler, 2025), ESM3 (Hayes et al., 2025) in its sequence-and-structure and sequence-only settings, and ProGen2 (Nijkamp et al., 2023)) skip inverse folding and proceed directly from design to prediction. Models may declare a `pipeline.stages` override in their model.v1 descriptor to depart from the default type-based routing; RFdiffusion3 and both ODesign checkpoints use this mechanism to insert LigandMPNN inverse folding despite being all-atom models. Backbone models that emit only a Cα trace (Genie2 and Genie3) cannot use the all-atom LigandMPNN and instead route inverse folding through a Cα-only ProteinMPNN variant. Genie3 conditions on the



(a) Protein types

(b) Top ligands and cofactors

Figure S2: Benchmark diversity. **(a)** Protein-type distribution across the 100 tasks. Enzymes account for 51%; the remaining 49% span 25 functional categories. **(b)** Top ligands and cofactors across the 64 ligand-bearing tasks (64% of the benchmark); Mg^{2+} (13 tasks), Zn^{2+} (10), and ADP (5) dominate.

motif's full side chains but deliberately declares no atom-resolution axis: declaring one would rename the run directory and change nothing else, since the atomistic variants are gated on model type, and tip-atom and theozyme conditioning are in any case unavailable to it: Genie3 tokenises a motif residue as its Cα plus side-chain atoms and never represents N, C or O, which 36 of the 100 tip-atom selections and 21 of the 41 theozyme selections name. ODesign ships an integrated OInvFold head that emits a sequence alongside the structure, but Stroma routes it through LigandMPNN instead, as ODesign's own protein-motif evaluations do (OInvFold is used there for nucleic-acid tasks). The reason is that OInvFold relabels motif positions: it reads only backbone atoms, so it can name a residue whose side chain it cannot see, and LigandMPNN pins a motif residue by trusting the label in the input structure. The design stage therefore restores the motif residue names before inverse folding, renaming a position only where the atoms present prove the label wrong. Two checkpoints are evaluated separately: **ODesign-Rigid** holds the receptor fixed during diffusion, while **ODesign-Flex** allows receptor side chains to flex.

Scaffold-length policy. Each task carries a curator-defined total-length range (Section B.2). At run time the orchestrator clips this range under one of two policies before sampling, so that lengths are comparable in the way each analysis requires. `model_cap` clips each model to its own `capabilities.max_residues` (used for single-model and debugging runs); `global_cap` clips every model to $\min(L_{\text{cap}}, \texttt{max_residues})$ for the cross-model tables, with $L_{\text{cap}} = 256$, La-Proteina's architectural maximum, so all paradigms are compared over a common length budget. Tasks whose minimum length exceeds the resulting cap are skipped before design (logged with the reason). For each (task, design index) the orchestrator then draws one target length from the clipped range with a deterministic RNG seeded by the task identifier, so the same design index sees the same target length across models and across reruns. How a model consumes that target is recorded per design as `length_source`: `orchestrator_fixed` when the model builds exactly the sampled length, or `model_internal` when the model samples its own length (range-native designers such as RFdiffusion2/3, La-Proteina, Protpardelle, and Genie2/3), in which case the target is advisory and the realized length is measured post hoc. The actual generated length is always logged alongside the target, so length-stratified analysis is valid regardless of which policy or source applies.

Ligand-pocket conditioning for Boltz2. For ligand-bearing tasks, Boltz2 left unconstrained places the cofactor freely, and it can drift off the scaffolded active site, rendering the clash and motif-geometry criteria uninformative. During motif specification we record, per task, the motif residues within 5 Å of any ligand heavy atom (`chemistry.ligand_pocket`). At prediction time the Boltz2 adapter maps these to the prediction's 1-based chain indexing and emits a *soft* `pocket` constraint, i.e. pocket conditioning rather than a hard steering potential (the `force` flag is disabled; distance tolerance 6 Å). The constraint is applied in indexed mode only and degrades to a no-op when

no pocket or no matching residues are found. It is deliberately soft: a genuinely incompatible design can still fail the ligand-clash criterion, so the steering improves pose realism without masking true failures.

C.2 Constraints: multi-chain and multi-PDB handling

This section expands the two coverage constraints summarised in Section 4.1.

Single-chain-only models (Protpardelle-1C single-chain checkpoints, PPIFlow, and the protein language models PoET-2 and ESM3) cannot natively scaffold multi-chain and multi-PDB motifs (totalling 11). An earlier version of the benchmark *chain-collapsed* (monomerized) such motifs, relabelling all chains to chain A, renumbering sequentially from 1 while preserving exact 3D coordinates, and asking the model to bridge the former inter-chain gaps with *de novo* linkers. We dropped this design: collapsing a quaternary or inter-chain motif into a single polypeptide is a scientifically distinct task with no single-chain ground truth, so a collapsed result is not comparable to native multi-chain scaffolding. Instead, single-chain-only models *skip* multi-chain motifs. The skip is centralized in the orchestrator (mirroring the multi-PDB skip below) and keyed on whether the motif's segments span more than one chain. Skipped tasks are marked not-applicable and excluded from the model's denominator rather than counted as failures; the multi-chain tasks form a dedicated `multi_chain` reporting stratum, evaluated by the models that handle multiple chains natively. The nine tasks that contain motifs spanning multiple chains are: AME-M0050, AME-M0058, AME-M0093, AME-M0375, AME-M0732, G2-3NTN, ST-13, ST-14 and ST-18.

Multi-PDB motifs are tasks whose segments originate from two or more distinct source PDB structures (e.g. task G2-4JHW+5WN9 combines a segment from PDB 4JHW with a segment from PDB 5WN9). Unlike multi-chain motifs, where all segments share a common crystal frame and their relative spatial arrangement is physically meaningful, multi-PDB segments exist in unrelated coordinate systems. Feeding these coordinates directly to a model that interprets input positions as absolute geometry (indexed scaffolding) would impose a spurious inter-segment distance dictated by the arbitrary crystal frames, forcing an unrealistic scaffold. The two tasks that contain multi-PDB motifs are: G2-4JHW+5WN9 and G2-3BIK+3BP5.

To prevent this, the pipeline checks each motif at runtime: a multi-PDB motif is dispatched only when (i) the *active* run is in unindexed mode, which treats segment positions as flexible constraints rather than fixed coordinates, or (ii) the model explicitly declares multi-PDB capability via the `supports_multi_pdb` flag in its model descriptor. Crucially the gate keys off the *active scaffolding mode*, not merely whether the model lists unindexed among its capabilities, so an indexed run never emits a meaningless joined-frame motif RMSD even for a model that also supports unindexed. Models that satisfy neither condition skip the task automatically. In practice, RFdiffusion2 and La-Proteina handle multi-PDB motifs through their unindexed scaffolding mode, while Genie2 supports them natively in indexed mode via group-letter annotations in its REMARK 999 input format, which signal that segments belong to distinct structural contexts. RFdiffusion, Protpardelle-1C, and PPIFlow, all indexed-only without multi-PDB support, skip these tasks.

C.3 Evaluation settings: full specification

This section expands the evaluation settings summarised in Section 4.2, giving per-paradigm applicability of the indexed/unindexed and all-atom/tip-atom/theozyme-atom dimensions, the motif-representation definitions, and the prediction-based RMSD computation (including Cα matching under unindexed evaluation).

Conditioning mode applicability per paradigm.

- **Backbone design models.** Indexed evaluation is the natural mode, providing motif residues as fixed-position backbone constraints at known sequence positions. Unindexed evaluation, where only motif geometry is provided without positional indices, applies only to models that explicitly declare support for free-placement conditioning. All backbone models in the current benchmark (RFdiffusion v1, Protpardelle-1C cc58, PPIFlow, Genie2, Genie3) support indexed mode only and are evaluated exclusively under that setting.

- **All-atom generative models.** La-Proteina, RFdiffusion2 and RFdiffusion3 declare both scaffolding modes and are evaluated under the indexed/unindexed $\times$ atom-resolution factorial, following the framework of La Proteina (Geffner et al., 2025). Both ODesign checkpoints declare indexed mode only, across all three atom resolutions. **Theozyme-atom** conditioning is declared by RFdiffusion2, RFdiffusion3 and both ODesign checkpoints, and is evaluated on the 41 AME tasks only; La-Proteina declares tip-atom but not theozyme. Protpardelle-1C cc91 declares only indexed mode and all-atom conditioning and is evaluated under that single setting.
- **Protein language models.** Indexed evaluation provides motif residues as fixed tokens at known sequence positions in the inpainting input. PoET-2 enforces motif residues via logit-clamped segmented autoregressive decoding; ESM3 pins motif positions on the sequence track (and optionally the structure track) natively, without logit clamping; ProGen2 is autoregressive with no inpainting capability, so it pins motif residues by the same constrained decoding as PoET-2. All four sequence rows are single-chain only, so multi-chain motifs are skipped (not chain-collapsed) and reported in the `multi_chain` stratum, and none of them declares unindexed support, so the mode axis does not apply to this paradigm.

Motif representation: all-atom, tip-atom, and theozyme-atom (all-atom models only). For all-atom generative models we evaluate three motif-conditioning variants. The **all-atom** setting provides full heavy-atom coordinates for each motif residue. The **tip-atom** setting provides only amino acid identity and the distal side-chain atoms beyond the final rotatable bond, from the per-residue lookup of Table S3. The **theozyme-atom** setting provides only the catalytic-subgraph atom selection from the AME source benchmark: for each motif residue, the 1–3-hop subgraph of heavy atoms around the atom closest to the ligand (stored in `motif_theozyme_atoms`; AME tasks only). Theozyme-atom conditioning is restricted to AME tasks by construction; tasks from other source benchmarks (MotifBench, Genie2, Stratum) have no theozyme-atom annotation and are automatically skipped when this setting is requested. This dimension is not applicable to backbone or pLM models and only to all-atom models that support atomistic conditioning.

Motif RMSD is measured on the prediction; matching under unindexed evaluation. For all paradigms, the motif-comparison family (motif RMSD and its atom-resolution variants, motif TM-score) is computed on the *Boltz2-predicted (refolded)* structure, not the raw design. This is field-standard and is what makes motif RMSD well-defined per rollout for backbone generators: the motif sequence is pinned by inverse folding, so Boltz2 reconstructs the motif side chains. Reference residues are remapped to the prediction's $1..L$ numbering through a per-design bridge (`motif_position_map.json` for sequence-only models; the contig `renumber_map.json`/`.trb` for structure models); when no bridge is available the matching falls back to the Cα assignment below. In the unindexed setting, motif positions in the generated protein are unknown. We use an exact minimum-cost matching adapted from La Proteina (Geffner et al., 2025): pairwise Cα distances between reference motif residues and all design residues form a cost matrix, and the globally optimal assignment is found via the Hungarian algorithm ($\mathcal{O}(k^3)$ in the motif residue count k). RMSD is then computed over the matched set after Kabsch alignment.

The cost combines distance with **residue identity**: a candidate of the wrong amino acid is charged a large penalty on top of its distance. Distance alone is not sufficient, and the reason is worth stating because it is easy to get wrong. The reference motif sits in its source structure's frame while the candidates come from the refold, which the predictor writes in a frame of its own, and nothing superposes the two before the assignment is solved; a purely geometric cost therefore decides the match by an accident of framing rather than by correspondence. Identity is known for every motif residue even when its position is not, and it is part of what unindexed conditioning already supplies to the model, so using it costs no information. The penalty is a strong preference rather than a hard constraint, which keeps the assignment feasible for a model that legitimately redesigns the motif; such a rollout still receives a best-effort match and is then judged by the atom-coverage gate.

C.4 Success-criterion configuration

This section expands the success criteria of Section 5.1. Success is evaluated *per rollout*: the dependency metrics (motif RMSD, pLDDT, scaffold self-consistency RMSD, ligand clash) are joined on the rollout key `{design}/{seq}/{prediction}`, and a design passes iff some

Table S6: Evaluation settings for every checkpoint, grouped by paradigm and ordered as in Table 2. ✓: evaluated; ×: not applicable. [a] Backbone and pLM models have no atom-resolution axis; motif conditioning uses backbone coordinates or sequence tokens respectively. The "All-atom" column indicates the single conditioning variant available; tip-atom and theozyme-atom conditioning do not apply to these paradigms. [b] Protpardelle-1C all-atom (checkpoint cc91) supports indexed all-atom only: no tip-atom, no unindexed. [c] Theozyme-atom conditioning is restricted to AME benchmark tasks (populated `motif_theozyme_atoms`); non-AME tasks are automatically skipped. Declared by RFdiffusion2, RFdiffusion3 and both ODesign checkpoints; La-Proteina declares tip-atom but not theozyme. [d] ESM3 seq-only is an ablation evaluated in indexed mode only; motif residues are pinned on the sequence track but the structure track is fully masked. [e] ODesign declares indexed conditioning only, at all three atom resolutions. It handles multi-chain motifs natively; multi-PDB motifs are skipped in indexed mode like every other model, since it does not declare `supports_multi_pdb`.

Checkpoint	Indexed			Unindexed		
	All-atom	Tip-atom	Theozyme[c]	All-atom	Tip-atom	Theozyme[c]
Backbone[a]						
PPIFlow	✓	×	×	×	×	×
Protpardelle-1C backbone	✓	×	×	×	×	×
RFdiffusion	✓	×	×	×	×	×
Genie3	✓	×	×	×	×	×
Genie2	✓	×	×	×	×	×
All-atom						
RFdiffusion3	✓	✓	✓	✓	✓	✓
La Proteina	✓	✓	×	✓	✓	×
ODesign-Rigid[e]	✓	✓	✓	×	×	×
ODesign-Flex[e]	✓	✓	✓	×	×	×
Protpardelle-1C all-atom[b]	✓	×	×	×	×	×
RFdiffusion2	✓	✓	✓	✓	✓	✓
Sequence (pLM)[a]						
ESM3 (seq+struct)	✓	×	×	×	×	×
ESM3 (seq-only)[d]	✓	×	×	×	×	×
PoET-2	✓	×	×	×	×	×
ProGen2 base	✓	×	×	×	×	×

single rollout satisfies all applicable criteria jointly on that one predicted structure. This replaces an earlier best-of-each-metric rule that could combine the best motif RMSD and the best pLDDT from *different* predictions into a chimeric success. For the atom-resolution variants, `success_rate_{tip,theozyme,all_atom}` additionally require the refold to preserve the motif's named side-chain atoms (the `motif_preserved` gate, defined with the motif-RMSD metrics below); the backbone-only `success_rate` baseline is ungated. The ligand-clash criterion is configurable via `clash_threshold` and is applied only when ligand atoms (HETATM records) are present in the Boltz2 output; it is silently skipped for apo tasks. Motif residues are excluded from the scaffold side of the clash check because active-site residues are by definition in close proximity to the ligand. Because Boltz2 co-folds the designed sequence with ligand CCD codes, severe clashes should be rare for well-designed scaffolds; the criterion flags structurally incompatible designs that would otherwise pass the RMSD and pLDDT filters. All thresholds in Table S7 are runtime-configurable defaults. The full ligand-clash metric definition is given below in Section C.5.

C.5 Detailed evaluation metrics

Stroma computes twenty-five evaluation metrics organised into six dispatch types, the categories by which `run_eval.py` calls a metric module (Table S8). Each metric is defined by a JSON descriptor in `pipeline/evaluate/tasks/` that specifies its inputs, parameters and outputs, and is validated against `task.schema.json`.

Comparison metrics compare the generated design against the reference motif geometry:

Table S7: Stroma success and designability criteria. Thresholds are defaults; all are configurable at runtime.

Criterion	Metric	Threshold	Applies to	Purpose
Motif RMSD (backbone)	`motif_rmsd`	≤ 1.0 Å	All paradigms	Motif geometry preservation
Motif RMSD (heavy-atom)	`motif_rmsd_{all_atom, tip, theozyme}`	≤ 1.5 Å	All-atom conditioning	Side-chain geometry preservation
Global pLDDT	`plddt`	≥ 70	All paradigms	Scaffold foldability
Ligand clash	`ligand_clash`	$= 0$ clashes	Tasks with ligands (64/100)	Protein–ligand compatibility
Scaffold Cα RMSD	`scaffold_rmsd`	≤ 2.0 Å	Structure models	Self-consistency (designability)

- **Motif RMSD.** Backbone (N, Cα, C, O) and Cα-only RMSD between motif residues in the Boltz2-predicted structure and the reference crystal structure, after optimal Kabsch superposition. Why the prediction rather than the design, and how reference residues are matched to it, are covered in Supplementary C.3.
- **Motif RMSD (all-atom).** Heavy-atom and all-atom RMSD of motif residues, assessing side-chain geometry recovery beyond backbone.
- **Motif RMSD (theozyme).** RMSD computed exclusively over the catalytic-subgraph atom selection from `motif_theozyme_atoms` (the 1–3-hop subgraph of heavy atoms around the ligand-proximal atom per residue). Applicable to AME tasks only; provides a focused measure of functional-geometry recovery for enzyme active sites. The atomistic variants (all-atom, tip-atom, theozyme) compare *named* side-chain atoms, so they additionally report `atom_coverage` (matched/total reference atoms) and a `motif_preserved` flag (coverage $\geq$ a configurable threshold). When fewer than two named atoms correspond between the reference and the refold, the metric returns an explicit `motif_not_preserved` status rather than an opaque shape mismatch, and the matching `success_rate` variant gates on `motif_preserved` so a refold that did not keep the motif's side chains counts as a true failure in the denominator.
- **TM-score.** Template modelling score (Zhang & Skolnick, 2004) between the reference motif Cα atoms and the corresponding residues in each generated design, providing a length-normalised measure of motif fold similarity (0–1 scale; > 0.5 indicates the same fold). Residue matching follows the same strategy as motif RMSD (by residue number for indexed evaluation, exact minimum-cost distance-and-identity matching for unindexed).

Design-vs-prediction metrics compare the generated design against its Boltz2-predicted structure:

- **Scaffold RMSD.** Backbone (N, Cα, C, O) and Cα-only RMSD between the designed structure and its Boltz2-predicted structure computed over *all* residues after optimal Kabsch superposition, measuring structural self-consistency (designability).

Single-structure metrics assess properties of individual predicted structures:

- **pLDDT.** Mean predicted local distance difference test score extracted from the Boltz2 confidence JSON (per-residue pLDDT array, together with PAE, iPTM, and pTM when available); falls back to the B-factor column when no confidence file is present. Reports the global mean pLDDT across all residues.
- **Motif pLDDT.** Mean pLDDT restricted to motif residues, measuring whether the structure predictor is confident specifically about the scaffolded functional region. Motif residue keys may be remapped before averaging when the model renumbers output residues: sequence-only models (PoET-2, ESM3) use `motif_position_map.json`; structure models that renumber residues sequentially (e.g. RFdiffusion v1) use `renumber_map.json`, which

Table S8: The twenty-five Stroma evaluation metrics by dispatch type. The dispatch type is how `run_eval.py` calls the metric module, not a statement about what the metric means: *comparison* is reference vs. candidate, *single structure* is per-prediction, *design vs. prediction* compares a design to its own refold, *sequence comparison* works on sequences, *aggregate* runs over all designs of a task, and *cross predictor* is the one type that reads outside its own run, comparing a rollout against sibling predictor runs. *`tm_score` is a retained legacy identifier that dispatches to `motif_tm_score`, so the twenty-five registry entries are twenty-four distinct measurements.

Metric	Summary
Comparison (6)	
`motif_rmsd`	Backbone / Cα motif RMSD, Kabsch-aligned
`motif_rmsd_all_atom`	Motif RMSD over all heavy atoms
`motif_rmsd_tip`	Motif RMSD over backbone + side-chain tip atoms
`motif_rmsd_theozyme`	Catalytic-subgraph RMSD (AME tasks only)
`motif_tm_score`	TM-score over motif residues
`tm_score`*	Legacy identifier, dispatched to `motif_tm_score`
Design vs. prediction (2)	
`sc_rmsd`	Scaffold self-consistency: design vs. refold
`sc_tm_score`	Full-structure TM-score, design vs. refold
Single structure (6)	
`plddt`	Global fold confidence
`motif_plddt`	Confidence restricted to motif residues
`ligand_clash`	Scaffold–ligand steric overlaps
`novelty_sequence`	MMseqs2 vs. UniRef30; motif masked out
`novelty_structure`	Foldseek vs. PDB100; motif stripped out
`sequence_complexity`	Scaffold-only compositional complexity; never gated
Sequence comparison (1)	
`sequence_recovery`	Motif sequence identity fraction
Cross predictor (2)	
`predictor_agreement_structure`	TM-score / Cα-RMSD between two predictors' folds
`predictor_agreement_confidence`	Rank correlation and signed bias of confidence
Aggregate (8)	
`success_rate`	Joint per-rollout pass rate (backbone criterion)
`success_rate_no_sc`	The headline criterion: as above, without the self-consistency term
`success_rate_all_atom`	As above, gated on the all-atom motif RMSD
`success_rate_tip`	As above, gated on the tip-atom motif RMSD
`success_rate_theozyme`	As above, gated on the theozyme motif RMSD
`success_rate_no_clash`	As above, without the ligand-clash criterion
`designability`	Scaffold self-consistency pass rate (structure models)
`designability_plm`	Single-sequence foldability (sequence models)

maps original chain/residue identifiers to design-local positions. Feeds into the success-rate aggregator alongside global pLDDT.

- **Ligand clash.** Severe steric clashes between non-motif scaffold heavy atoms and co-folded ligand atoms in Boltz2 predictions. Boltz2 co-folds the designed sequence with ligand CCD codes; this metric checks whether the resulting structure contains geometrically impossible atom overlaps at the scaffold–ligand interface. A clash is defined as a heavy-atom pair within 1.5 Å. Motif residues are excluded from the scaffold side because they are expected to be ligand-adjacent. For tasks without ligands (HETATM records absent), the metric returns `has_ligands=false` and a neutral `n_clashes=0`. Reported summary statistics: `n_clashes`, `min_distance` (closest scaffold–ligand distance), `n_scaffold_atoms`, `n_ligand_atoms`, and `clashing_residues` (unique scaffold residues involved in clashes).
- **Sequence novelty.** MMseqs2 search against UniRef30 (ColabFold 2302 release), with the motif residues masked to `X` in the query so that only scaffold residues drive the alignment (Supplementary C.8). Requires `mmseqs` on PATH and a pre-indexed database; skipped gracefully otherwise.

- **Structural novelty.** Foldseek search against PDB100, the structural analogue of that masking: every motif residue is *stripped* from the query structure, so the query is scaffold-only and coverage is measured over the scaffold (Supplementary C.8). Requires `foldseek` on PATH and a pre-indexed database; skipped gracefully otherwise. Both novelty metrics are *success-gated*: they run only over the prediction structures whose design passed the success criteria, so novelty characterises solutions rather than arbitrary samples.

Sequence comparison metrics compare designed and reference sequences:

- **Sequence recovery.** Fraction of motif residues whose amino acid identity matches the reference crystal structure.

Aggregate metrics combine outputs from upstream metrics:

- **Success rate.** Fraction of designs with at least one rollout that jointly satisfies every applicable threshold of Supplementary C.4 on that one predicted structure. A rollout whose motif could not be bridged to the reference numbering is recorded as *unevaluable* rather than scored: scoring it would report a number nobody measured, and failing it would report a genuine zero. A task with no bridged design at all is reported N/A, exactly as a degenerate single-residue motif is; a partly bridged task keeps a rate over its measured designs, which the strict denominator makes a floor.
- **Designability.** Fraction of designs whose scaffold Cα RMSD to the Boltz2 prediction falls below a threshold (default 2.0 Å), indicating a recoverable backbone. Applies to structure-generating models (backbone design, all-atom).
- **Designability (pLM).** For sequence-only models (PoET-2, ESM3, ProGen2) there is no designed backbone to compare against the prediction; scaffold RMSD is therefore not defined. Instead, *designability* for pLMs is operationalised as single-sequence foldability: the fraction of designs whose sampled sequence folds to a confident Boltz2 structure (pLDDT $\geq$ 70). Motif preservation is deliberately *not* folded into it, being reported by the motif-RMSD family instead, so the two properties stay separable for models that have no backbone of their own.

Designability is deliberately *not* subject to the unevaluable rule above: its criterion is scaffold RMSD or pLDDT, neither of which reads the motif bridge.

Metrics are computed in dependency order: aggregate metrics (success rate, designability) depend on upstream metrics (motif RMSD, pLDDT, scaffold RMSD) and are automatically scheduled after them. Sequence and structural novelty require external tools (`mmseqs` and `foldseek` on PATH) and pre-indexed databases; they are skipped gracefully when not configured.

Cross-predictor metrics are the one dispatch type that reads outside its own run. Both compare a rollout against the sibling runs that folded the *same* designs under a different structure predictor, resolved from the predictor level of the run directory:

- **Predictor agreement (structure).** TM-score and Cα-RMSD between the two predictors' structures for one rollout. Both folded the same sequence, so Cα atoms correspond one-to-one by index and no alignment is involved.
- **Predictor agreement (confidence).** Spearman and Pearson correlation plus median signed bias on pLDDT, motif pLDDT, pTM and ipTM. Correlation says whether the two predictors *order* rollouts alike; bias says whether one is uniformly more confident, which is what moves a fixed pLDDT threshold and is invisible to correlation.

Both are a graceful no-op for a single-predictor run, and both require that the compared runs share a design stage: running the pipeline twice under different predictors regenerates the designs, which would confound predictor disagreement with design sampling noise. Rollouts that cannot be compared (shape mismatch, missing structure) are excluded from the summary rather than scored as disagreement.

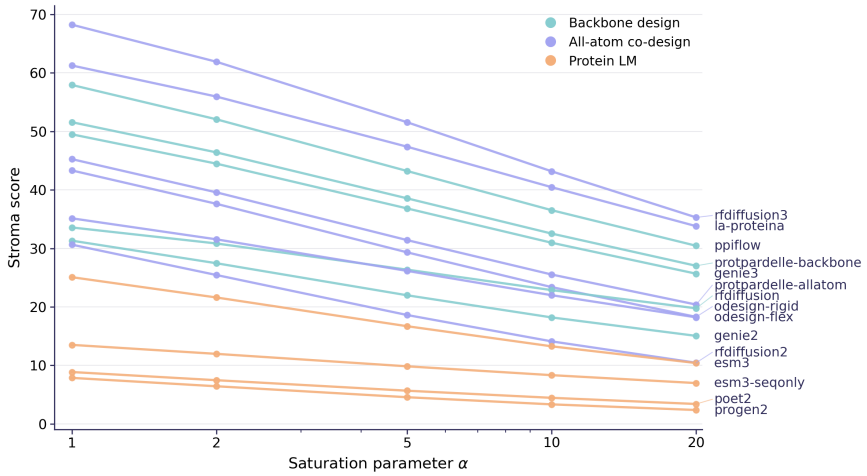


Figure S3: Stroma score against the saturation parameter α, one line per checkpoint, over the 89-task common subset. Raising α lowers every score, because it slows the rate at which the per-task term saturates, but the lines do not cross: the ordering is unchanged ($\rho \geq 0.996$ against $\alpha = 5$, Table S9) and RFdiffusion3 leads throughout.

C.6 The Stroma score: cluster definition and sensitivity to α

The Stroma score counts unique solutions as Foldseek clusters at 50% sequence identity and 80% coverage, with no TM-score cut. MotifBench and Genie3 instead define a unique solution by a TM threshold, 0.5 and 0.6 respectively, so cluster counts are not interchangeable between the two conventions and ours are the more permissive. Re-clustering the campaign at TM 0.5 and scoring it under MotifBench's own four-check criterion lowers every score, by 0.9 to 6.2 points, and leaves the leading four models in the same order. We report the Stroma score under our definition and this comparison alongside it, rather than silently adopting either.

The parameter α in Eq. 1 controls the trade-off between breadth (obtaining *any* solution per task, rewarded heavily for small α) and depth (accumulating diverse solutions per task, rewarded for large α). To verify that the choice $\alpha = 5$ does not drive our conclusions, we recompute the Stroma score for each model under five values $\alpha \in \{1, 2, 5, 10, 20\}$ and measure the Spearman rank correlation ρ between each resulting ranking and the reference ranking at $\alpha = 5$ (Table S9). The sweep is what decides whether the benchmark outcome depends on this hyperparameter; we report it either way rather than treating stability as given.

The ranking is insensitive to it. Over $\alpha \in \{1, 2, 5, 10, 20\}$ the mean Stroma score moves from 37.5 to 18.5, because α sets how quickly the per-task term saturates, but the induced ordering barely moves and the leading model is RFdiffusion3 at every value.

Table S9: Spearman rank correlation ρ between Stroma score rankings at different values of α and the reference ranking at $\alpha = 5$. Computed via `plot_stroma_alpha_sweep` in `analysis/plots.py`.

Reference ($\alpha = 5$) vs.	$\alpha = 1$	$\alpha = 2$	$\alpha = 5$	$\alpha = 10$	$\alpha = 20$
Spearman ρ	0.996	0.996	1.00	1.000	0.996

C.7 Design-side motif placement: procedure and control

For each design that carries a structure, the checker reads the per-design motif bridge the model wrote (Supplementary C.3), pairs each reference motif residue with the design residue the bridge names, and computes a Kabsch-aligned Cα RMSD over those pairs. Nothing is refolded, and no correspondence is inferred: the measurement uses the model's own account of where it put the motif. Designs with

fewer than three bridged residues, or with non-finite coordinates, are recorded as unmeasured rather than scored, since a superposition over two points is not determined. Protein language models are excluded because they emit no structure to superimpose.

Thresholds. A placement beyond 2.0 Å is flagged and one beyond 5.0 Å is treated as a failure to place, matching the scale at which a motif ceases to be present in any useful sense. These thresholds classify the diagnostic only; they do not enter the success criterion of Section 5.1, which remains defined on the refold.

Chirality. A reflection leaves every interatomic distance unchanged but cannot be undone by a proper rotation. A mirror-image motif therefore reads as a large placement RMSD with no other symptom, and no distance-based metric (RMSD, TM-score, or a distance-map objective) can separate it from a motif the model failed to place. Whenever a placement exceeds the flag threshold we refit the same pairs against the reflected design coordinates; if the reflected fit is both better and within the threshold, the design is reported as a chirality failure. This distinction is not cosmetic: a model whose motif conditioning is distance-based is reflection-invariant by construction, so a mirrored motif satisfies its objective exactly as well as the correct one does, and the failure is a property of the conditioning rather than of a bad sample. The test is applied under backbone and all-atom conditioning, where Cα geometry is the conditioned quantity.

Result. Over the 28 rows that emit a structure, 3,108 of 230,253 designs (1.35%) are mirrored, and the distribution is far from uniform (Figure S4). Genie2 is the outlier at 15.96%, four times the next row, and mechanistically so: its motif conditioning is distance-based, and distances are reflection-invariant, so a mirrored motif satisfies its objective exactly as well as the correct one. Eight rows are at zero, including both Protpardelle checkpoints, PPIFlow and every all-atom row of RFdiffusion2 and RFdiffusion3. Those same two models mirror at 1.1 to 2.3% under tip-atom and theozyme conditioning: with only the tip atoms pinned, the rest of the residue is free to settle into a reflected arrangement. The five protein language model rows are excluded, having no structure to reflect.

None of these designs reach the success criterion. Of 2,032 rollouts from severely mirrored designs, zero pass, against a 9.8% baseline for the same models, and their median refold motif RMSD is 7.4 to 7.7 Å against a 1.0 Å cut. The mechanism matters more than the rate: they are excluded by the refold, because the predictor folds an L-amino-acid chain and will not reproduce a reflected backbone, and not by any criterion we impose. Were motif RMSD scored on the design instead of the prediction, that protection would disappear without any metric changing.

Control. A placement number is only evidence if a deliberately wrong correspondence scores worse. Every row is therefore measured twice: once through the model's bridge, and once through the same bridge with the design-side residue index shifted by a fixed offset. If the shifted mapping does not score worse, the check is not measuring placement on that row and its numbers are reported as unmeasured. This matters in a specific, non-hypothetical way: a motif whose Cα atoms are close to collinear and evenly spaced superimposes well onto a shifted copy of itself, so short motifs can produce a confident-looking number that the control shows to be meaningless. Running the control per row and automatically, rather than as a separate pass someone has to remember, is what keeps an unmeasurable row from being reported as a clean one.

Scope. The diagnostic runs under indexed conditioning only. Under unindexed conditioning the motif's location is recovered by the evaluator's matching rather than declared by the model, so there is no model-authored correspondence to audit and a design-side number would measure the matcher. For tip-atom and theozyme conditioning, Cα placement is not the conditioned quantity, so per-design findings are suppressed and the row's placement is reported as informational rather than as a gate.

C.8 Novelty computation

This is the *structure* novelty axis (the sequence axis is the MMseqs2 metric of Section C.5). For each successful design the motif residues are first stripped from the query structure (scaffold-only novelty), then Foldseek-Search approximates the highest TM-score to any PDB100 structure. Successful

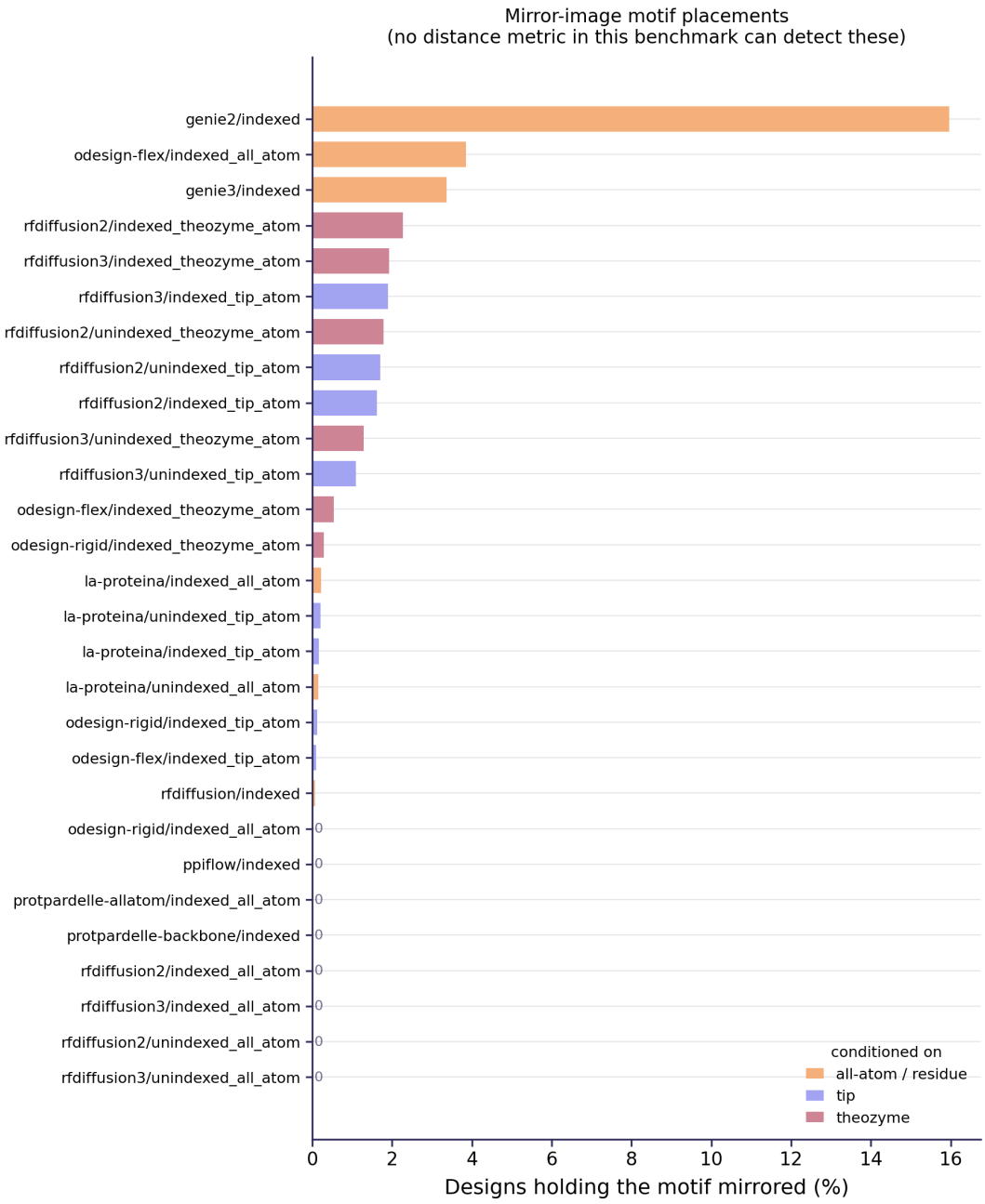


Figure S4: Mirror-image motif placement per row, over the 28 rows that emit a structure. Genie2's distance-based motif conditioning is reflection-invariant by construction and it mirrors on 16.0% of designs, four times the next row; eight rows never do. The five sequence-model rows are absent, having no structure to reflect.

designs are clustered with Foldseek-Cluster; within each cluster, mean $(1 - \text{TM-score})$ is computed. Novelty is the mean across clusters (Algorithm 1). The candidate set S being successful designs is what makes novelty success-gated.

Result. Averaged over the headline rows, the two structure paradigms are indistinguishable on both axes: structural novelty 0.391 for backbone and 0.388 for all-atom, sequence novelty 0.745 and 0.741. Sequence models are lower on both, at 0.295 and 0.602, and they are also the models that succeed least (Figure S5). This is worth stating because the opposite is the natural expectation. A model that

Algorithm 1 Novelty computation for a single task

Require: Set of successful designs $S = \{s_1, \ldots, s_k\}$; PDB100 database $\mathcal{D}$
Ensure: Novelty score $\nu \in [0, 1]$
if $S = \emptyset$ **then**
  return $\nu \leftarrow 0$
end if
Cluster S via Foldseek-Cluster $\rightarrow$ clusters $\{C_1, \ldots, C_m\}$
for each design $s \in S$ **do**
  $\mathrm{TM}_s \leftarrow \max_{d \in \mathcal{D}} \mathrm{TMscore}(s, d)$
end for
for $\ell = 1$ **to** m **do**
  $\nu_\ell \leftarrow \frac{1}{|C_\ell|} \sum_{s \in C_\ell} (1 - \mathrm{TM}_s)$
end for
$\nu \leftarrow \frac{1}{m} \sum_{\ell=1}^{m} \nu_\ell$
return ν

succeeds less often might be doing so because it explores further from known structures; these do not. Per-model values are in Table 2. Note also that novelty is computed only over successful designs, so a row with few successes rests on few clusters.

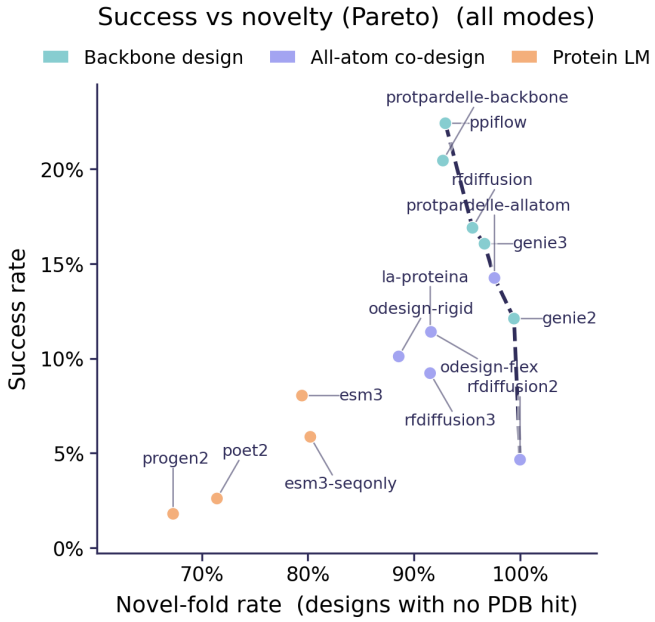


Figure S5: Success against the share of designs with no PDB hit, one point per checkpoint, averaged over every mode that checkpoint ran rather than the indexed rows alone, so the rates are lower than Table 2's. The sequence models sit at the bottom left: they neither succeed more nor explore further. The dashed line is the Pareto front.

C.9 Implementation details

Run counts. Every row of the campaign uses $N = 100$ designs per task, $M = 2$ inverse-folded sequences per design (backbone and inverse-folding models only) and $K = 2$ predictions per sequence. A *rollout* is one (design, sequence, prediction) triple, so a backbone model contributes $N \times M \times K$ rollouts per task and an all-atom or pLM model $N \times K$.

Scaffold length. Every row runs under the `global_cap` policy of Supplementary C.1 at a 256-residue cap, the smallest `max_residues` among the benchmarked models, so all paradigms are compared over one length budget.

Structure prediction. Boltz-2 is the predictor for every row reported in the main table, co-folding the designed sequence with the ligand on ligand-bearing tasks under the soft pocket constraint of Supplementary C.1.

Checkpoints and environment. Each model is run from its public released checkpoint inside a pinned container image, with the parameters, image and entry point recorded in the model's JSON declaration. Where a model exposes several checkpoints selected per task, the checkpoint that served each task is recorded in that task's run manifest rather than inferred afterwards.

Result aggregation. Per-design evaluation results (motif RMSD, pLDDT, scaffold RMSD, success pass/fail) are aggregated across all models and settings into flat CSV files by `aggregate_results.py`. Each row of `per_task_results.csv` is one (model, setting, task) triple carrying the success rate, mean RMSD, mean pLDDT and the task-level metadata (category, source benchmark, ligand status, segment count). Every downstream analysis in this paper (cross-model ranking, per-category breakdowns, threshold sensitivity, failure-mode classification and figure generation) is a `groupby` over that single denormalised table.

D Additional results

D.1 Results by task category

Results are reported separately over the three mutually exclusive structural strata defined in Section 6: *native* (89 tasks), *multi-chain* (9 tasks) and *multi-PDB* (2 tasks). Single-chain-only models skip the multi-chain and multi-PDB strata rather than collapsing them into one polypeptide, so those tasks are excluded from the model's denominator rather than scored as failures.

Table S10: Success rate per structural stratum, headline criterion. A dash means the checkpoint declares itself single-chain-only and skips the stratum, so those tasks leave its denominator rather than counting as failures. No checkpoint is scored on multi-PDB in indexed mode: joining segments from unrelated source structures yields a fictional inter-segment geometry, which the orchestrator refuses unless the run is unindexed.

Model	Paradigm	Native (89)	Multi-chain (9)
RFdiffusion3	all-atom	25.6	2.2
La Proteina	all-atom	23.1	–
PPIFlow	backbone	22.4	–
Protpardelle-1C backbone	backbone	20.4	–
ODesign-Rigid	all-atom	17.0	27.4
RFdiffusion	backbone	17.0	16.3
ODesign-Flex	all-atom	16.7	24.8
Genie3	backbone	16.1	–
Protpardelle-1C all-atom	all-atom	14.2	–
Genie2	backbone	12.1	–
RFdiffusion2	all-atom	9.6	8.7
ESM3 (seq+struct)	pLM	8.0	–
ESM3 (seq-only)	pLM	5.9	–
PoET-2	pLM	2.6	–
ProGen2 base	pLM	1.8	–

The multi-PDB stratum has no column because nothing is scored in it: the two tasks are reachable only in unindexed mode, only RFdiffusion2 and RFdiffusion3 attempt them there, and none of the 700 designs they produced succeeds. That is the stratum's result, not an omission.

Only five checkpoints attempt the multi-chain stratum, and the ordering there is not the native one. Both ODesign checkpoints score *higher* on multi-chain than on native tasks (27.4% and 24.8% against 17.0% and 16.7%), while RFdiffusion3 falls from 25.6% to 2.2%. With nine tasks these are wide intervals and we do not read a ranking from them, but they are the reason the strata are reported separately: averaging them into the headline would move models by several points on nine tasks that most checkpoints never attempt.

D.2 The complete run matrix

Table S11 gives the complete run matrix: every checkpoint under every setting it declares. One denominator differs from the common subset: PPIFlow reports 88 tasks rather than 89 because AME-M0630 is unscored for it, its motif bridge having failed, so the task is reported N/A rather than as a zero (Supplementary C.5). Its 100 designs exist; they cannot be measured against the reference numbering. The ESM3 sequence-only ablation and the ODesign-Flex checkpoint are first-class rows of Table 2 rather than supplementary entries, since each is a distinct checkpoint rather than an alternative setting of one.

D.3 Predictor dependence: full results

Design. Seven design models $\times$ 20 tasks $\times$ 10 designs = 1,400 designs, each folded at $K = 2$ by Boltz2, AlphaFold 3 and OpenFold3, plus a fourth Boltz2 arm with the ligand pocket hint enabled. Nine models were run; both ODesign checkpoints are excluded from every figure below, because their sequences came from ODesign's OInvFold head, which Stroma has since replaced with LigandMPNN. Tasks were drawn from the design-stage validation campaign, which covers all nine models, under three criteria: every model has at least 10 designs for the task, the motif has at least 2 residues (backbone motif RMSD is degenerate for a single-residue motif, so such a task can only contribute ties), and the draw is stratified by source $\times$ ligand. The result is 13 ligand-bearing and 7 ligand-free tasks with motif sizes from 2 to 99 residues across all four sources. Arms are comparable because they share a design stage: re-running the pipeline under a different predictor would regenerate the designs and confound predictor differences with design sampling noise. The arms were verified to share their inputs by extracting the sequence back out of 2,799 folded structures and comparing: zero mismatches.

Ranking transfers, rates do not. The ordering is stable across all three predictors (Table S14) while the absolute rates are not, falling by about four points from Boltz2 to OpenFold3. The correlations drop only on the ligand-free stratum, where 7 tasks and a compressed spread of rates let small differences reshuffle ranks; that is a small-sample artefact rather than a disagreement between predictors, and the per-stratum figures come from the earlier nine-model pass, which could not be recomputed after the two ODesign checkpoints were dropped because that study's run tree is no longer on disk.

Threshold sweeps. Sweeping the backbone motif-RMSD cut over $\{0.75, 1.0, 1.5, 2.0, 3.0\}$ Å moves the mean success rate from 11.6% to 30.0% under Boltz2 while leaving between-predictor correlation at 0.93–1.00, and the predictor ordering (Boltz2 $>$ AlphaFold 3 $>$ OpenFold3) holds at every cut, so it is a uniform offset rather than an artefact of where the line is drawn. The leaderboard's top entry is stable to 1.5 Å and changes at 2.0 Å. Sweeping the heavy-atom cut over the four all-atom checkpoints gives OpenFold3 rates of 0.23, 0.67, 0.70, 0.73 and 0.77 of Boltz2's at 1.0, 1.5, 2.0, 3.0 and 5.0 Å, over mean atomistic success rates of 4.88, 9.12, 11.12, 16.00 and 19.25% under Boltz2. The 0.23 at the tightest cut is a tail effect rather than disagreement about side chains: it is confined to that one threshold and disappears as soon as the cut moves off the edge of the distribution. The benchmark's heavy-atom criterion sits at 1.5 Å (Section 5.1), inside the flat region, so the criterion Stroma actually reports is not in the fragile regime.

Two criteria do little work. Moving the pLDDT cut from 50 to 80 changes the mean success rate by about a point, and the ligand-clash test binds on 3.0% of the 2,860 ligand-bearing rollouts (85 of them). The joint criterion is carried almost entirely by motif RMSD and scaffold self-consistency. We report this because it bounds what the other two criteria can be doing: they exclude a small number of designs that the geometric criteria would otherwise pass, which is their purpose, but they do not drive the ranking.

The ligand pocket hint moves the ligand, not the verdict. Enabling the hint cuts the scaffold–ligand clash rate from 3.0% to 2.2% of ligand-bearing rollouts, so it demonstrably changes where the predictor puts the ligand, while success moves from 12.7% to 12.4%. Ligand placement therefore reaches the verdict only through the clash test, which binds rarely, and the hint-off arms above can be read against the campaign's hint-on protocol.

Native-sequence control. Folding the benchmark's own native sequences separates the predictors far more than folding designs does (Table S14), and 27 of the 60 shorter tasks come within 2 Å under Boltz2 against 7 and 3 for the other two. Above 300 residues all three degrade, which is outside the benchmark's regime. All folding was single-sequence, which is the benchmark's regime but not AlphaFold 3's design regime for natural proteins, so the defensible claim is that *in the regime the benchmark runs in* Boltz2 reproduces experimental structures far better, not that it is a better predictor in general. It also explains the agreement above: a de novo design has no meaningful MSA for any of the three, so the predictors agree on designs and diverge on natural sequences.

Coverage is not a differentiator. All 72 CCD ligand codes in the benchmark resolve in both AlphaFold 3's and OpenFold3's chemistry dictionaries, and both fold the benchmark's extremes (three chains, three ligands, and the 814-residue native) correctly. ESMFold2 was excluded on coverage grounds instead: its packaged build featurises one protein chain with no MSA and no ligands, so it cannot represent 69 of the 100 tasks, and all 41 AME tasks carry ligands, so choosing it would delete the theozyme axis entirely.

Limits. The comparison covers seven of the fifteen benchmarked checkpoints: those that emit a structure, had validation-campaign designs available, and are still configured as they were when folded. It excludes PPIFlow, Protpardelle-backbone, both ODesign checkpoints and all five sequence models, so predictor dependence is unmeasured for the sequence paradigm. The subset also contains no multi-chain tasks, because the pool shared by all nine models excludes them, so it says nothing about the multi-chain stratum. Absolute success rates are not transferable between predictors and are not comparable to the campaign's, since the designs are a different and earlier set; the claim is about ranking, not levels. Among natives the multi-chain stratum ($n = 9$) is the one place the ordering is inconclusive: AlphaFold 3 has the better median (4.62 Å) and the worse mean (12.61 Å), so it succeeds on a few and misses badly on others, which at $n = 9$ is not a result.

D.4 Threshold sensitivity

The success criteria in Section 5.1 fix a motif RMSD threshold per criterion, 1.0 Å on backbone Cα atoms and 1.5 Å on the heavy-atom variants, together with a pLDDT threshold of 70 and a scaffold self-consistency threshold of 2.0 Å. To show that the reported ranking is not an artifact of these choices, we recompute every model's success rate over a grid of thresholds and report the rank correlation of the resulting orderings against the headline setting.

The same number is not the same stringency across atom sets. Section 5.1 takes each threshold from the benchmark that published it, which leaves the backbone and heavy-atom cuts on different footings. Measured over the 323,327 rollouts of the 41 AME tasks for which all three atom sets are defined, a backbone cut of 1.0 Å passes 6.1% of rollouts; passing the same fraction needs 1.63 Å on heavy atoms and 1.07 Å on the catalytic subgraph. The inherited cuts therefore sit either side of equivalence: at 1.5 Å the heavy-atom criterion is slightly stricter than backbone-at-1.0 (5.5% pass), while the theozyme criterion is more permissive (10.1%). We report the published cuts rather than an equalised one, because a cut chosen to equalise pass rates would no longer be the number either source benchmark validated, but rows scored under different atom sets should be read with this in mind.

Loosening the cut raises every rate, as it must, but the ordering is stable across the range a reader would consider: $\rho \geq 0.96$ from 0.5 to 1.5 Å, degrading to 0.83 only at 2.5 Å, by which point a 2.5 Å motif is not meaningfully preserved. The exception is the top of the table. At 1.5 Å and above La Proteina overtakes RFdiffusion3, which is the concrete reason we take each threshold from the benchmark that published it rather than choosing a round number: the choice is visible precisely where it matters most. Figure S6 shows the per-model curves, which separate by paradigm rather than crossing.

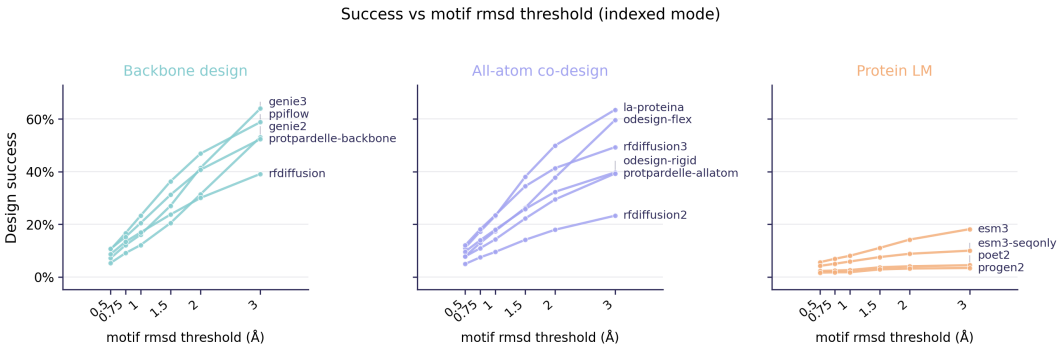


Figure S6: Per-task success against the motif RMSD threshold, one panel per paradigm, indexed rows. Curves rise together and rarely cross, which is why the rank correlation in Table S15 stays high while the absolute rates more than quadruple.

D.5 Scaffold sequence complexity

The success criteria are geometric, so nothing in them constrains the sequence a model writes around the motif. We therefore score every design for compositional complexity and report the result as a diagnostic. No design is filtered on it.

The score. For a scaffold of length L with Shannon entropy H over its amino-acid composition and longest single-residue run r,

$$q = \frac{H}{\log_2 \min(20, L)} - \min\left(0.5,\ \frac{\max(0,\ r-3)}{10}\right), \qquad q \in [0, 1]. \tag{2}$$

The first term is entropy normalised by the most a sequence of that length could carry, so a short scaffold is not penalised for being short. The second separates 40 glutamates in one block from the same 40 scattered through the chain, which entropy alone cannot do. The motif is excised before scoring, matched by residue number through the same bridges the evaluation uses, so q measures what the model invented rather than the natural fragment it was handed.

Designs are less complex than the proteins the motifs come from. Figure S7a compares the 444,282 scored rollouts of the indexed rows against the 97 motif-bearing chains of the source structures, which are the fairest available reference because they are the proteins these motifs were cut from. The designed median is 0.81 against 0.94, and only 5 of the 97 natural chains fall below the median design. The gap is not an artefact of the length normalisation: residues drawn independently at UniProt background frequencies also score a median of 0.94. It is a shift of the whole population rather than a tail of bad designs, and every checkpoint sits below the natural median (Figure S7b), the ESM3 variants furthest (medians 0.59 and 0.70) and PoET-2 and ODesign-Rigid least (0.85).

The criteria do not see it. Rollouts below 0.50 have a median pLDDT of 73 against 72 for those above, and 55.1% of them clear the pLDDT gate against 53.0%: a low-complexity scaffold is not merely tolerated by the confidence term, it passes marginally more often. Figure S7d restricts to the rollouts that already clear that gate, the population the criterion has accepted, and 9.0% of them sit below 0.50 at the same median pLDDT as the rest.

Where a cut would fall, and what it would cost. Low complexity is sometimes the correct answer, so any cut has to be conservative. The lowest-scoring tasks are repetitive by nature (Figure S7c), collagen (ST-14, median 0.22) most of all, and the variance of q across tasks (0.0060) exceeds its variance across models (0.0048), so a high cut penalises a model for which tasks it happens to solve. At 0.50 exactly one of the 98 tasks has a median below the cut and no natural chain does, which is why we name that value. Disqualifying every rollout below it removes 6.9% of all successes and leaves the ranking essentially unchanged: Spearman $\rho = 0.986$ against the reported order, the leader and the top four unchanged, five of fifteen rows moving by one or two places, RFdiffusion falling two and Protpardelle-1C all-atom losing the most at 22.2% of its successes. The ordering survives to

0.60 and breaks at 0.80, where half of all successes are flagged and the cut is deciding the ranking rather than describing the sequences (Table S16). We report the finding and the defensible cut rather than applying one, since filtering would hide exactly the models whose sequences it describes.

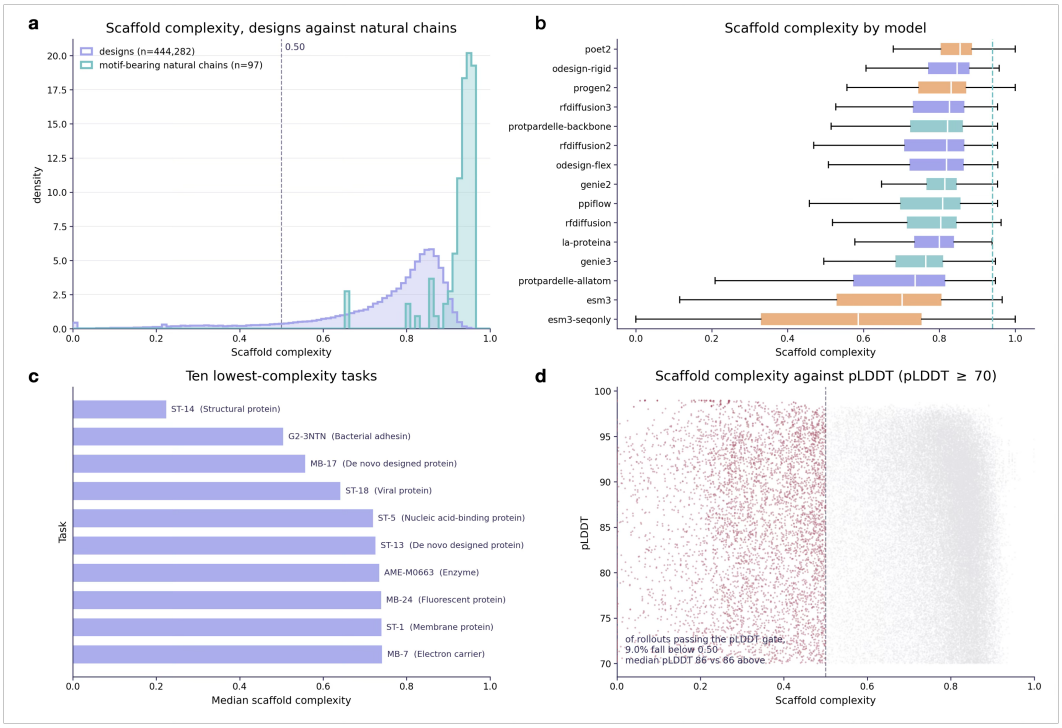


Figure S7: Scaffold sequence complexity, Equation 2, over the indexed rows. **(a)** The designed population against the 97 motif-bearing chains of the source structures; the dashed line is the 0.50 cut. **(b)** Per checkpoint, coloured by paradigm, with the natural median dashed. **(c)** The ten lowest-scoring tasks with their protein type: these are repetitive by nature, which is the reason to keep any cut low. **(d)** Complexity against pLDDT for the rollouts that already clear the pLDDT gate, showing that the confidence term does not separate them.

D.6 What the models build

Success says whether a design satisfies the criteria, not what kind of protein it is. Assigning secondary structure with DSSP to every *successful* design of the campaign, on the predicted structure that everything else is scored on (20,122 designs across the 15 checkpoints), shows that all three paradigms build predominantly helical scaffolds (Figure S8) and differ in degree rather than in kind: the median design is 61% helix for backbone models, 53% for all-atom and 44% for sequence models, with strand medians of 5%, 13% and 16%. Half of the backbone designs contain essentially no strand at all (50% below a 5% strand fraction, against 41% and 38% for the other two paradigms). Whether that reflects the models, the motifs they are given or the confidence gate favouring helical folds is not separable from these data, but it bounds how varied the solutions on this benchmark currently are.

D.7 Inductive bias analysis: full specification

Per-task success rates are compared across paradigms with a Kruskal–Wallis test followed by pairwise Mann–Whitney U tests with Bonferroni correction. To separate the contribution of task difficulty from that of the model, we fit a linear mixed-effects model over per-task, per-model success rates, with a random intercept per task:

$$y_{mt} = \beta_0 + \beta_{\mathrm{par}(m)} + \boldsymbol{\beta}^{\top}\mathbf{x}_t + u_t + \varepsilon_{mt}, \qquad u_t \sim \mathcal{N}(0, \sigma^2_{\mathrm{task}}), \quad \varepsilon_{mt} \sim \mathcal{N}(0, \sigma^2), \quad (3)$$

where y_{mt} is the success rate of model m on task t, $\beta_{\mathrm{par}(m)}$ a fixed effect for the model's paradigm, $\mathbf{x}_t$ the task's structural covariates (motif residue count, number of discontiguous segments, and the

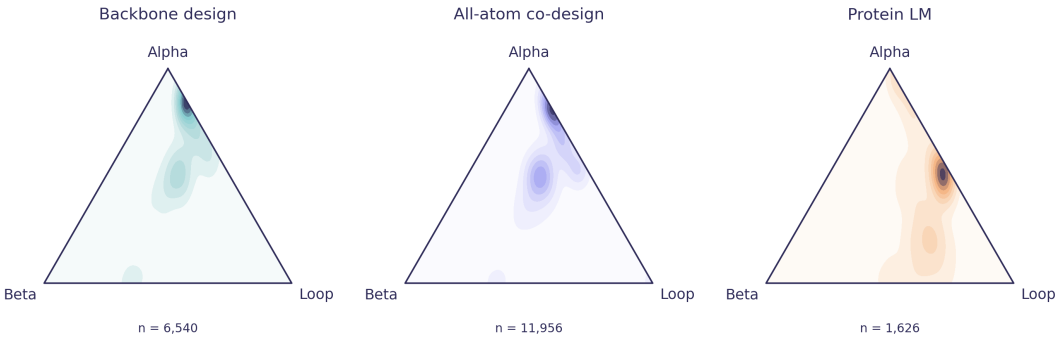


Figure S8: Secondary-structure composition of every successful design, by paradigm: helix, strand and loop fractions on a ternary density, assigned by DSSP on the predicted structure. All three paradigms concentrate towards the helix vertex, the sequence models least tightly.

spatial extent of the motif), and u_t a per-task random intercept. The variance ratio $\sigma^2_{\text{task}}/(\sigma^2_{\text{task}} + \sigma^2)$ gives the share of variation attributable to the task rather than the model.

Variance decomposition. Task identity accounts for 81.6% of the variance in per-task success, model identity for 3.4%. A benchmark in which these were reversed would be measuring the models; this one is measuring the tasks, and the models are separated mainly by how far up a shared difficulty ordering they get.

Paradigm. Mean per-task success is 17.6% for backbone checkpoints (95% CI 11.9–23.9), 17.6% for all-atom (12.7–23.0) and 4.6% for sequence models (1.8–8.0). The first two intervals overlap almost completely. The sequence paradigm is separated from both, and the separation survives every stratification we report. The three paradigms also build similar kinds of protein, which Supplementary D.6 quantifies.

Inverse folding versus co-design. All-atom checkpoints split into those that emit their own sequence and those that receive one from LigandMPNN. Grouped that way, co-designed rows average 18.7% (task-bootstrap CI 13.2–24.8, $n = 2$ models) and inverse-folded rows 17.2% (11.9–22.9, $n = 4$). The 1.5-point gap sits inside a 6.4-point within-group spread, and the group label explains 0.06% of per-task variance against 3.4% for model identity.

Two per-task paired comparisons did reach small p-values, self-consistency RMSD worse by 0.63 Å and pLDDT worse by 3.9 for the inverse-folded group ($p \leq 0.0001$). Both are artefacts of group size. With two models on one side and four on the other, "group" and "which model" are nearly the same variable, and on all four metrics tested, removing a single model reverses the sign of the effect. What the pooled test detects is which individual models fall on each side, not an architectural difference. We therefore report all-atom as one paradigm rather than splitting it.

Failure-mode classification. Thirteen of the 89 common tasks are solved by no checkpoint, and success falls steeply with fragmentation for every paradigm (Figure S9). Segment count is the strongest single discriminator of unsolvability (AUC 0.747, $p = 0.004$): tasks with six segments are unsolved 83% of the time and average 0.4% success, against 0% unsolved and 34.8% mean success for single-segment tasks. But the unsolved set is bimodal, and a test run over all 89 tasks at once detects nothing (residues per segment, AUC 0.45, $p = 0.56$).

Mode A is seven enzyme active sites, all from AME, carrying four to seven residues across four to six segments: roughly one residue per segment, with no contiguous fragment to build from, and the correct geometry defined only between them. **Mode B** is the opposite, motifs of 22 to 43 residues in two to four segments that occupy a median 28% of the shortest permitted design against 9% for solved tasks (AUC 0.787, $p = 0.020$; on motif size alone AUC 0.833, $p = 0.007$). It contains both nanobody tasks, two immune proteins, a toxin and a signalling protein. There is too little scaffold left for the model to have any freedom in satisfying the constraint.

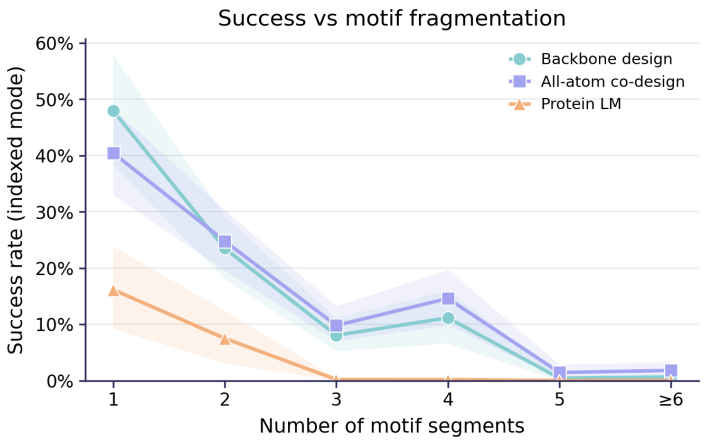


Figure S9: Success against motif fragmentation, by paradigm, indexed rows, with 95% intervals. The two structure paradigms track each other across the whole range; the sequence models fall to zero beyond two segments and do not recover.

Table S11: The complete Stroma run matrix: 32 rows over the 15 model checkpoints of Table 2, one row per checkpoint × scaffolding mode × atom-conditioning resolution. The row that appears in Table 2 is each checkpoint's first: indexed scaffolding at all-atom conditioning. Only models declaring `capabilities.atom_resolutions` carry a conditioning axis, and only models declaring unindexed support carry the mode axis, so the matrix is ragged rather than a full cross-product. n is the eligible-task count (of 100) derived from the pipeline's skip gates, as in Table 2; it varies within a checkpoint because theozyme conditioning is defined only on the 41 AME enzyme tasks, and because indexed mode skips the multi-PDB tasks whose joined-frame motif RMSD would be meaningless while unindexed mode (geometric matching) can attempt them. Two further reductions are model-declared limits rather than task properties. RFdiffusion2 unindexed tip-atom is capped at motifs of at most 28 residues, which makes 25 of the 100 tasks N/A for that row and that row alone: the mode splits the motif into chain-like components and labels them from a 26-letter alphabet, so a larger motif exhausts the alphabet. The boundary was measured over the full benchmark and is exact (largest success 28 residues, smallest failure 29); all 41 AME tasks pass, and every excluded task is a large structural motif of 29–99 residues. We declare the limit rather than work around it, so the row reports the regime the method was built and validated for. SR here is over each row's own n, not the 89-task common subset, so a row's value differs from Table 2 wherever its n does. RMSD is the mean motif RMSD over successful designs. La Proteina carries no theozyme rows at all, because its catalytic-atom placement was measured and found unusable in both modes, so the capability is not declared.

Model	Mode	Conditioning	n	SR (%)	Stroma	RMSD (Å)
Backbone						
RFdiffusion	indexed	n/a	98	16.9	26.4	0.67
Genie2	indexed	n/a	89	12.1	22.0	0.74
Genie3	indexed	n/a (all motif atoms)	89	16.1	36.8	0.74
PPIFlow	indexed	n/a	88	22.4	43.2	0.74
Protpardelle-1C backbone	indexed	n/a	89	20.4	38.5	0.71
All-atom						
RFdiffusion2	indexed	all-atom	98	9.5	17.6	0.72
	indexed	tip-atom	98	6.7	9.4	0.68
	indexed	theozyme	41	4.6	10.2	0.77
	unindexed	all-atom	100	1.1	1.7	0.74
	unindexed	tip-atom	75	1.5	3.4	0.68
	unindexed	theozyme	41	2.9	5.5	0.79
RFdiffusion3	indexed	all-atom	98	23.5	47.6	0.70
	indexed	tip-atom	98	15.8	28.6	0.69
	indexed	theozyme	41	5.2	11.8	0.76
	unindexed	all-atom	100	1.3	3.8	0.74
	unindexed	tip-atom	100	1.1	2.5	0.73
	unindexed	theozyme	41	2.5	5.1	0.78
ODesign-Rigid	indexed	all-atom	98	17.9	28.5	0.67
	indexed	tip-atom	98	5.3	5.0	0.64
	indexed	theozyme	41	2.8	1.6	0.71
ODesign-Flex	indexed	all-atom	98	17.4	25.8	0.69
	indexed	tip-atom	98	8.3	9.4	0.65
	indexed	theozyme	41	4.3	6.6	0.75
La Proteina	indexed	all-atom	89	23.1	47.4	0.70
	indexed	tip-atom	89	19.4	32.5	0.70
	unindexed	all-atom	89	1.6	5.7	0.65
	unindexed	tip-atom	89	1.4	3.9	0.65
Protpardelle-1C all-atom	indexed	all-atom	89	14.2	31.4	0.72
Sequence (pLM)						
ESM3 (seq+struct)	indexed	n/a	89	8.0	16.7	0.65
ESM3 (seq-only)	indexed	n/a	89	5.9	9.8	0.61
ProGen2 base	indexed	n/a	89	1.8	4.5	0.54
PoET-2	indexed	n/a	89	2.6	5.7	0.62

Table S12: Full-coverage results: each checkpoint scored over *its own* eligible-task set rather than the common subset of Table 2. n is that set's size, derived from the pipeline's skip gates rather than hand-entered, and *extra* is how many tasks beyond the common 89 the checkpoint could attempt. These rates are not comparable across rows, because the denominators differ; they answer a different question, namely how a model does over everything it can represent. The Stroma score is an average over tasks, so it too is tied to n: adding harder tasks lowers it, and a model that attempts more can therefore score below one that attempts fewer while performing identically on the tasks they share. Only the fixed-n column of Table 2 ranks models. A task a model cannot represent is excluded from its denominator rather than counted as a failure, so a larger n is a wider capability claim and not a handicap. That exclusion is deliberately generous to the narrower models, and the generosity is why both tables exist: scoring a skipped task as a failure would conflate *cannot attempt* with *attempted and failed*, while hiding the skips would let a narrow model look equal to one that does strictly more, so Stroma reports the rate and the coverage separately and never combines them into one number. Read alongside the per-stratum results in Supplementary D.1, which give the multi-chain and multi-PDB tasks their own denominators.

Model	n	extra	SR (%)	Stroma	RMSD (Å)
Backbone					
RFdiffusion	98	+9	16.9	26.4	0.67
Genie2	89	+0	12.1	22.0	0.74
Genie3	89	+0	16.1	36.8	0.74
PPIFlow	88	-1	22.4	43.2	0.74
Protpardelle-1C backbone	89	+0	20.4	38.5	0.71
All-atom					
RFdiffusion2	98	+9	9.5	17.6	0.72
RFdiffusion3	98	+9	23.5	47.6	0.70
ODesign-Rigid	98	+9	17.9	28.5	0.67
ODesign-Flex	98	+9	17.4	25.8	0.69
La Proteina	89	+0	23.1	47.4	0.70
Protpardelle-1C all-atom	89	+0	14.2	31.4	0.72
Sequence (pLM)					
ESM3 (seq+struct)	89	+0	8.0	16.7	0.65
ESM3 (seq-only)	89	+0	5.9	9.8	0.61
ProGen2 base	89	+0	1.8	4.5	0.54
PoET-2	89	+0	2.6	5.7	0.62

Table S13: AME enzyme tasks under the backbone criterion and under the theozyme criterion, for the four checkpoints that declare theozyme conditioning. Both columns are scored on the same rollouts of the same theozyme-conditioned rows, over the 36 AME tasks that fall in the common subset. The theozyme criterion is not uniformly the stricter test: at the published cuts, 1.5 Å on catalytic heavy atoms against 1.0 Å on backbone atoms, it passes more often for the two RFdiffusion checkpoints and less often for both ODesign ones, which is a statement about where each model places catalytic side chains rather than about the criterion's severity.

Model	Backbone SR (%)	Theozyme SR (%)	Δ
RFdiffusion3	5.9	25.0	$+19.1$
RFdiffusion2	5.2	12.9	$+7.7$
ODesign-Flex	4.9	4.8	-0.1
ODesign-Rigid	3.2	0.4	-2.8

Table S14: The three predictors over the seven-model basis: mean backbone success rate, and the native-sequence control (median Cα RMSD to the experimental structure, over all 98 single-source tasks and over the 60 at or below 300 residues). The three pairwise rank correlations between predictors are 0.991 to 1.000 and per-design verdict agreement is $\kappa = 0.70$–0.80. All folding is single-sequence, which is the benchmark's regime.

Predictor	Mean SR (%)	Native, all (Å)	Native, $\leq$300 (Å)
Boltz2	13.0	6.27	2.57
AlphaFold 3	10.1	14.27	9.73
OpenFold3	9.1	16.93	12.88

Table S15: Motif RMSD threshold sweep, indexed rows. Mean success is the mean per-task rate across the 15 checkpoints; ρ is the Spearman correlation of the induced model ranking against the headline setting of 1.0 Å.

Motif RMSD cut (Å)	0.50	0.75	**1.00**	1.25	1.50	2.00	2.50
Mean success (%)	7.2	10.8	**14.2**	17.9	21.8	29.0	34.9
Spearman ρ	0.975	0.993	**1.000**	0.986	0.964	0.925	0.832
Leader		RFdiffusion3				La Proteina	

Table S16: Scaffold complexity cut sweep, headline rows over the 89-task common subset. Flagged is the share of rollouts below the cut, lost the share of successes they carry, and ρ the Spearman correlation of the induced ranking against the reported one.

Complexity cut	0.40	0.45	**0.50**	0.55	0.60	0.70	0.80
Rollouts flagged (%)	5.6	6.9	**8.7**	11.0	14.2	24.7	47.9
Successes lost (%)	4.6	5.6	**6.9**	8.7	11.4	21.3	48.8
Mean success (%)	13.5	13.4	**13.2**	12.9	12.6	11.2	7.3
Spearman ρ	0.996	0.986	**0.986**	0.993	0.993	0.982	0.893
Rows moved	2	5	**5**	4	4	7	11
Top four intact	yes	yes	**yes**	yes	yes	no	no

Table S17: The two failure modes among the 13 unsolved tasks, against the 76 solved ones. The two halves are separated by opposite extremes of the same two variables, which is why pooling them finds no effect.

	Mode A: scattered	Mode B: oversized	Solved
Tasks	7	6	76
Median segments	6	3	3
Median motif residues	6	34	11
Motif / shortest allowed design	0.06	0.28	0.09