RETENTION IS NOT PRESERVATION: BUDGET-MATCHED EVALUATION OF ISOFUNCTIONAL PROTEIN REWRITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein engineering often asks for a prescribed sequence change while preserving the properties that make the original sequence useful. Generative models can meet a length or identity command while changing the structure that places annotated sites in context. We formulate **isofunctional rewriting**—compression, expansion or sequence diversification to a commanded budget under a function-preservation constraint—and evaluate it with budget-matched, feasibility-checked controls and component-wise metrics. On 40 held-out proteins at 60% of the wild-type length, ISOFORM retains 96.5% of annotated sites versus 60.0% for a contiguous wild-type fragment, yet its TM-score from the independent AlphaFold2 structural branch is lower (0.487 vs. 0.555). The aggregate proxy ranks the design first ($D_\Phi = 0.347$ vs. 0.371) because the site term offsets losses in structure, embedding and property terms; the aggregate and structural endpoints therefore give different rankings. Across the tested identity-matched controls, the structural ordering is unchanged. All endpoints are computational proxies; biochemical activity is not measured.

1 INTRODUCTION

Protein engineering often changes an existing sequence to satisfy an external constraint while trying to preserve its function. One practical constraint is delivery capacity: viral-vector packaging limits can make a shorter coding sequence desirable even when the original protein is functional (Grieger & Samulski, 2005; Wu et al., 2010). Similar requests arise in enzyme miniaturisation, antibody optimisation, and sequence diversification, where the desired change is local or global but the functional requirement is distributed across the molecule (Hie et al., 2024; Lála et al., 2026). The engineering request therefore has two parts: make a prescribed change, and preserve the properties that make the original sequence useful.

The last few years have produced several complementary ways to generate protein sequences. Protein language models learn sequence regularities at scale and can be used for generation or mutation scoring (Rives et al., 2021; Madani et al., 2023; Verkuil et al., 2022; Meier et al., 2021). Structure-conditioned models instead constrain generation by a backbone or a target motif, as in inverse folding, motif scaffolding, and diffusion-based structure design (Hsu et al., 2022; Dauparas et al., 2022; Wang et al., 2022; Trippe et al., 2023; Watson et al., 2023; Ingraham et al., 2023). Evolutionary and diffusion language models provide another route through sequence space (Alamdari et al., 2023; Wang et al., 2024). More recent systems make the edit itself explicit, including variable-length generation and protein miniaturisation (Gu et al., 2019; Havasi et al., 2025; Baron et al., 2025; Devkota et al., 2026). Together, these methods make it plausible to ask a model for a sequence at a specified length or identity; they do not by themselves establish that the requested rewrite preserves function.

This distinction exposes an evaluation gap. Protein design benchmarks commonly measure fitness, structure, or sequence-level plausibility, but a rewrite request also has a feasibility constraint and a reference sequence that should be compared at the same budget (Notin et al., 2023; Gligorijević et al., 2021). A residue-level motif score can remain favourable while the three-dimensional arrangement of those residues changes; a structure score can improve because a control copies native material while losing the annotated sites that motivated the edit. Aggregate objectives introduce a second ambiguity when their terms use related representations or when their weights are chosen for optimisation rather

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.

than calibrated for evaluation. These issues are familiar in controlled generation, where satisfying a length or edit constraint is a separate problem from preserving content (Kikuchi et al., 2016; Malmi et al., 2022).

We study this question with **isofunctional rewriting**. The input is a wild-type sequence x, an operation o, and a numerical budget β. The operation is one of three user-level requests: *compress* to a target length below $|x|$, *expand* to a target length above $|x|$, or *paraphrase* at a prescribed sequence-identity ceiling. This edit-and-budget interface is related to controlled length generation and sequence editing, but the preservation target is a protein function rather than text likelihood (Gu et al., 2019; Kikuchi et al., 2016; Malmi et al., 2022; Bandel et al., 2022). The output y must meet the budget and should remain close to x under a stated computational surrogate for function. We separate the budget, the matched control, and each preservation metric instead of collapsing them into one score.

We check control feasibility with respect to the annotated-site constraint, whose source annotations come from the UniProt knowledgebase, match the control to the design's budget, and report global structure, site correspondence, embedding, and property terms separately (The UniProt Consortium, 2023; Wang et al., 2022). The conclusions concern predicted structures and learned functional proxies, including language-model and structure-prediction scores, rather than measured biochemical activity (Lin et al., 2023; Meier et al., 2021; Notin et al., 2023; Jumper et al., 2021).

The central observation is a dissociation between site retention and structural preservation, a distinction that motivates motif-scaffolding and structure-aware design in the first place (Wang et al., 2022; Trippe et al., 2023). Under compression at $0.6L$, ISOFORM retains 96.5% of annotated active, binding, and metal-coordinating sites, compared with 60.0% for the contiguous fragment on the same 40 proteins. Its evaluation-oracle TM-score is nevertheless lower (0.487 versus 0.555), where TM-score is a structural similarity measure rather than a direct functional assay (Zhang & Skolnick, 2004; 2005). The aggregate D_Φ favours the design (0.347 versus 0.371) because the site component offsets losses in the other components. These are proxy measurements of site retention and predicted structure; they do not establish biological activity, as benchmark fitness and proxy scores should not be conflated with experimental function (Notin et al., 2023; Dallago et al., 2021).

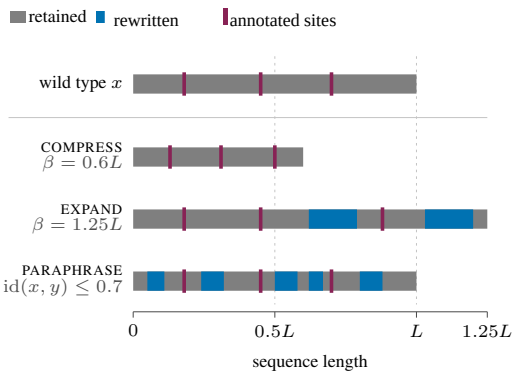
Contributions.

1. **A common task interface.** We place compression, expansion, and sequence diversification under one operation-and-budget interface, and make the preservation surrogate explicit so that its terms can be inspected.
2. **A budget-matched evaluation protocol.** We define operation-specific controls, test their feasibility before using them as references, and report the consequences when a control is structurally informative but not a valid isofunctional solution.
3. **A component-wise diagnostic.** On the held-out compression set, annotated-site retention and predicted structure move in opposite directions. External baselines and a structure-heavy retraining preserve the same ranking under the tested proxy.

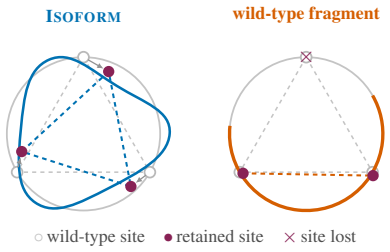
2 Related work

Changing a protein's length. Baron et al. (2025) shrink proteins with discrete diffusion, deletion only; Devkota et al. (2026) encode proteins as fixed-dimensional distributions and decode at a requested length; Cheng et al. (2026) make length a generated quantity through a generalised Poisson process. Wang et al. (2026) predict indels in latent alignment space, Deutschmann et al. (2026) learn homologue trajectories as edit flows, and Zhang et al. (2026) post-train for executable edit trajectories. These works optimize properties or match homologue distributions. Our focus is complementary: we make the rewrite budget explicit and ask how the answer changes when it is compared with a feasible, budget-matched construction.

Keeping function explicitly. Wu et al. (2026) combine language models with Monte Carlo sampling over embedding energy; Lála et al. (2026) miniaturise enzymes under active-site constraints (Lála et al., 2025); Gligorijević et al. (2021); Berenberg et al. (2022) sample variable-length sequences under function-predictor guidance. In structure space the analogous constraint is motif scaffolding,



(a) One interface, three operations



	wild type	ISOFORM	fragment
site retention ↑	1.000	0.965	0.600
TM-score ↑	1.000	0.487	0.555
$\lvert\Delta d\rvert$ (Å) ↓	0.00	0.80	0.27

(b) Compression at $0.6L$: retention vs. structure

Figure 1: Isofunctional rewriting and the dissociation it exposes. Panel (a) shows one operation-and-budget command under a common interface. Grey marks material retained from the wild type, blue marks material written by the model, and dark red marks annotated functional sites; sites are evaluated after alignment. Panel (b) compares compression at $0.6L$: ISOFORM retains more annotated sites than a contiguous wild-type fragment, yet has lower predicted TM-score. The fragment preserves the native order of its retained segment but omits a site, whereas the rewrite keeps the sites and changes their arrangement. The chains are schematic. Site retention and TM-score are the primary $0.6L$ cell ($n = 40$ proteins, Table 2); $|\Delta d|$ is the mean Cα displacement over site pairs retained by both sequences, pooled over the released compression geometry records ($n = 31$, Table 7). All endpoints are computed from predicted structures; biochemical activity is not measured.

where a functional geometry is held fixed while the surrounding chain is generated (Trippe et al., 2023; Wang et al., 2022; Watson et al., 2023); there the motif is preserved by construction and the length of the scaffold is free, whereas our setting commands the length and evaluates how the motif is preserved within it. The function-preserving methods above use per-target search or structure-space constraints. We instead study an amortized conditional model, while treating the evaluation protocol—rather than the model architecture—as the main contribution.

Diversity as a goal. Within fixed-length inverse folding (Dauparas et al., 2022; Hsu et al., 2022), Zhou et al. (2025) and Silva et al. (2025) pursue diversity, but as a regulariser rather than a commanded target. Among the systems compared in Table 1, none exposes a sequence-identity ceiling as a user-level control. The comparison follows the published capabilities and implementations listed there. In that literature length is also fixed, and what is preserved is structure rather than function.

Edit-based generative modelling. The Levenshtein Transformer (Gu et al., 2019; Gu & Kong, 2021) learns deletion and insertion policies from alignment-derived edit paths, and recent work casts variable-length generation as edit flows (Havasi et al., 2025) or insertion language models (Patel et al., 2025; 2026), with continuous-time formulations reaching biological sequences (Sharafutdinov & André, 2026; Kim et al., 2026). These learn unconditional or property-directed edit processes. We use the same broad edit perspective with an operation, a numeric budget, and an explicit preservation surrogate.

Where the controls come from. Rewriting under an invariant is an old problem in text, where the three operations are studied together (Malmi et al., 2022; Filippova et al., 2015; Kikuchi et al., 2016; Iyyer et al., 2018; Kumar et al., 2020) and identity ceilings are already user-set (Bandel et al., 2022). We borrow two things: scoring preservation and divergence jointly rather than one at a time (Reimers & Gurevych, 2019; Zhang et al., 2020; Zhu et al., 2018), and the extractive baseline, whose protein analogue is the control this field is missing. Protein benchmarks have meanwhile concentrated on fitness prediction and representation quality (Notin et al., 2023; Dallago et al., 2021; Rao et al., 2019); what a *rewrite* should be scored against has not been fixed in the same way.

Table 1: Capability comparison with the closest prior art. ✓ present, ✗ absent, ○ partial. A partial entry denotes support only for a restricted operation, target regime, or interface. The table summarizes the published interfaces used in our comparison. The claimed distinction is the conjunction of an amortized operation-and-budget interface and component-wise evaluation; the matrix is not intended as a universal ranking of prior work.

Method	Operations			Control interface		
	COMPRESS	EXPAND	PARAPHRASE	target length	identity budget	explicit oracle
SCISOR (Baron et al., 2025)	✓	✗	✗	○	✗	✗
Raygun (Devkota et al., 2026)	✓	✓	○	✓	✗	✗
GPFlow (Cheng et al., 2026)	○	○	✗	✗	✗	✗
STRIDE (Zhang et al., 2026)	○	○	✗	✗	✗	✗
Mind the Gap (Wu et al., 2026)	✗	✗	✓	✗	✗	✓
Miniaturising (Lála et al., 2026)	✓	✗	✗	○	✗	✓
DivPro (Zhou et al., 2025)	✗	✗	○	✗	✗	○
ISOFORM (this work)	✓	✓	✓	✓	✓	✓

3 ISOFUNCTIONAL REWRITING

Let $x \in \mathcal{A}^L$ be a wild-type protein sequence and let y be a rewrite. We use *isofunctional* as an operational label for satisfying the stated computational proxies; it is not evidence of biochemical equivalence. The input command is (x, o, β), where $o \in \{\text{COMPRESS}, \text{EXPAND}, \text{PARAPHRASE}\}$ is the operation and β is its budget.

For compression, the target length is $L' = \beta L < L$. For expansion, $L' = \beta L > L$. For paraphrase, $|y| \approx L$ and the global Needleman–Wunsch identity satisfies $\mathrm{id}(x, y) \leq \beta$; the ceiling prevents the model from solving preservation by returning a near-copy, while a separate minimum-change requirement would be needed to demand a particular amount of divergence. All lengths are rounded by the protocol and all identity values are reported on the alignment, so a continuous target is subject to quantisation.

3.1 A COMPUTABLE PRESERVATION SURROGATE

True function is not observed in this study. We therefore define a normalized surrogate D_Φ whose terms are reported separately:

$$D_\Phi(x, y) = \lambda_\mathrm{s} d_\mathrm{struct} + \lambda_\mathrm{m} d_\mathrm{motif} + \lambda_\mathrm{e} d_\mathrm{embed} + \lambda_\mathrm{p} d_\mathrm{prop}, \tag{1}$$

with $(\lambda_\mathrm{s}, \lambda_\mathrm{m}, \lambda_\mathrm{e}, \lambda_\mathrm{p}) = (0.40, 0.35, 0.15, 0.10)$. Each d uses the fixed calibration defined for that component, is oriented so that lower is better, and is on a comparable scale across proteins; the aggregate is a diagnostic score, not a calibrated probability of functional equivalence. We do not use a universal numerical threshold ϵ to claim biological success; when a binary decision is needed, the protocol reports the pre-specified budget and site-retention checks alongside D_Φ.

Structure. $d_\mathrm{struct} = 1 - \mathrm{TM}$, where TM is computed by predicting both structures with the stated oracle and aligning them with TM-align. This is a similarity between predicted structures, not a measurement of the experimentally realized fold.

Annotated sites. d_motif is the alignment-restricted loss over UniProt active, binding, and metal-coordinating sites. It charges lost sites and displacement in relative sequence position. The three-dimensional arrangement of sites is evaluated separately with Cα distances because sequence-position drift is only one-dimensional.

Embedding. d_embed is cosine distance in an ESM2-650M function-aligned embedding space. This term is useful for training but is shared with some controls and therefore is not treated as an independent biological measurement.

Properties. d_prop is the normalized disagreement between the DeepFRI GO predictor, the ProtT5 enzyme-family classifier, and a geometric metal-site predictor applied to x and y. These predictors are also shared between training and evaluation; their contribution is reported separately from the independent structural branch.

For COMPRESS, the KEEP mask sets deletion logits to zero at annotated positions. This is a generation constraint on source tokens, not a guarantee that a later global alignment will count every site as retained. We therefore report generation-level masking and alignment-level retention separately. For PARAPHRASE, in-place substitution is used because a delete-then-insert policy can otherwise move the compensating token to a different position.

Naturalness is kept outside D_Φ. The fluency term $\mathcal{F}(y)$ is an ESM-1v pseudo-log-likelihood used by the training objective and is not interpreted as function preservation.

4 CONTROLS AND EVALUATION PROTOCOL

A control is useful only if it satisfies the same command as the design and is admissible under the stated task. We therefore treat feasibility and budget matching as part of the evaluation rather than as post-hoc filters.

Metrics and estimands. For each protein we report target-length or identity compliance, alignment-level retention of annotated active, binding, and metal-coordinating sites, predicted-structure TM-score, site-geometry displacement, embedding distance, property disagreement, and sequence naturalness. TM is higher-is-better; all distances are lower-is-better. The primary compression estimand is the paired difference between ISOFORM and the contiguous fragment at $0.6L$, with the protein as the unit of analysis. Other budgets and operations are reported as stress tests and are not pooled with this primary endpoint unless stated explicitly.

Operation-specific controls. For compression, the control is a contiguous fragment of the wild type at the commanded length. For expansion, it is the wild type with a duplicated native segment inserted at a random interior position. For paraphrase, it is an ESM2 masked-fill substituter calibrated to the same identity ceiling. A reference homologue at the matched budget is reported as a descriptive ceiling, not as a competing method.

Feasibility. A fragment that discards annotated sites quantifies the structural benefit of copying native coordinates. At $0.6L$ it retains 0.600 of annotated sites, so we label it a *structural reference* and use it to measure the preservation–structure gap.

Budget matching. A baseline that changes the sequence by a different amount can win by rewriting less or lose by rewriting more. We therefore use identity-matched masked fill for paraphrase. Length targets use the protocol's integer rounding rule; identity is measured on the global Needleman–Wunsch alignment and reported to three decimals.

Expansion as a specification stress test. The expansion command requires a longer sequence but does not require inserted material to be novel or to carry a specified function. Consequently, a padded wild type is feasible under the stated proxy and can obtain a high TM-score without demonstrating a meaningful rewrite. We report expansion results as an audit of this specification and do not use them as evidence of successful functional rewriting.

Candidate selection and oracle provenance. For each prompt we sample eight candidates at temperature 1.0. Candidates are ranked by the complete normalized training-oracle distance, and only the selected candidate is sent once to the frozen evaluation pipeline. The structural evaluation branch uses AlphaFold2 in single-sequence mode with no MSA and three recycles, and is independent of the training structural branch (ESMFold). ESM2 embeddings and the property predictors are shared between training and evaluation; their terms are therefore diagnostic and are reported separately. Every TM-score in the paper is from the AlphaFold2 structural branch.

Statistics. Proteins are the paired units. We use two-sided exact sign tests, report Cliff's delta for paired direction, and use 10^4 per-protein bootstrap resamples for uncertainty. Budgets from the same protein are not treated as independent replicates. For each prompt we record the candidate identifier, training-oracle score, and selected flag; the primary table reports the selected evaluation values.

5 ISOFORM: CONDITIONAL EDITING UNDER THE TASK INTERFACE

ISOFORM is a conditional edit model that instantiates the task. The primary endpoint is whether the protocol distinguishes site retention from structural agreement.

Backbone and conditioning. The model uses a non-autoregressive parallel edit policy in the Levenshtein Transformer family (Gu et al., 2019; Gu & Kong, 2021). A deletion policy chooses source residues to remove and an insertion policy fills the resulting placeholders. Operation, scalar budget, and a wild-type functional embedding are injected into the protein-language-model trunk. For compression, the annotated-site deletion mask is applied to source positions. For paraphrase, each deleted position is remasked and refilled in place with the original residue forbidden; this avoids treating a substitution as an unrelated insertion elsewhere in the sequence.

Training objective. Natural rewrite pairs are mined from homologous families and converted to edit trajectories by global alignment. Synthetic pairs cover minimal-domain compression and structure-guided expansion. The identity filter prevents low-identity targets from turning into unconditional generation. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{edit}} + \alpha_1 D_\Phi(x, y) - \alpha_2 \mathcal{F}(y) + \alpha_3 \mathcal{L}_{\text{budget}} + \alpha_4 \mathcal{L}_{\text{PARAPHRASE}}, \tag{2}$$

where $\mathcal{F}$ is ESM-1v pseudo-log-likelihood and $\mathcal{L}_{\text{PARAPHRASE}} = \max(0, \text{id}(x, y) - \beta)$ is used only for paraphrase. Structural rewards require folding and are estimated with policy-gradient updates; differentiable terms are used directly. The complete normalized training-oracle score ranks the eight candidates at inference. The selected candidate is evaluated once with the frozen evaluation pipeline; the evaluation-oracle score is not used for selection.

The distinction between these two scores matters. The training oracle uses ESMFold for its structural term, whereas the evaluation structural branch uses AlphaFold2 without an MSA. ESM2 embeddings and property predictors are shared and are not independent evidence; they are retained to diagnose which components determine the aggregate ranking.

6 EXPERIMENTS

Setup and primary estimand. The evaluation set contains 40 held-out proteins with UniProt functional-site annotations. The primary estimand is the paired difference between ISOFORM and the contiguous-fragment control for compression at $0.6L$. We report the protein-level mean, sample standard deviation, paired wins, and an exact two-sided sign test.

Training pairs are mined from clustered sequence databases and converted to edit trajectories by global alignment. The model is evaluated at three compression budgets, three expansion budgets, and three paraphrase ceilings; these additional cells are stress tests and are not treated as independent replicates of the primary endpoint. Raygun and SCISOR are evaluated at their available target lengths under the same structural scoring pipeline. Claims about external systems are restricted to the systems and checkpoints that were actually run.

The budget frontier is consistent with the primary dissociation. For compression, the design/control TM pairs are $0.834/0.862$ at $0.90L$, $0.671/0.714$ at $0.75L$, and $0.487/0.555$ at $0.60L$. Expansion gives $0.900/0.932$, $0.862/0.915$, and $0.642/0.874$ at $1.10L$, $1.25L$, and $1.50L$, respectively; these are specification stress tests because a padded wild type can satisfy the length command. The corresponding paraphrase pairs are $0.957/0.967$, $0.918/0.952$, and $0.933/0.939$ at identity ceilings 0.90, 0.80, and 0.70. These secondary cells are descriptive summaries rather than independent replicates; their per-protein vectors are not part of the release. The complete operation–budget table is in Appendix C.

Question 1: do the preservation criteria agree? At $0.6L$, the commanded target length is reached for all 40 proteins. The primary table gives macro-average alignment-level site retention of 0.965 for ISOFORM and 0.600 for the fragment. The corresponding d_{motif} means are 0.083 and 0.421. The design has the lower motif distance on all 40 paired proteins (exact two-sided sign test $p = 1.8 \times 10^{-12}$); this is a site-correspondence result, not a biochemical activity measurement.

Table 2: Primary endpoint for COMPRESS at $0.6L$ ($n = 40$ proteins). Standard deviations are across proteins; retention is the macro-average of per-protein alignment-level fractions. Bracketed intervals are percentile 95% bootstrap CIs from 10^4 resamples with seed 42.

metric	n	ISOFORM	fragment
TM (mean $\pm$ SD)	40	0.487 ± 0.083	0.555 ± 0.075
TM (95% CI)	40	$[0.462, 0.513]$	$[0.532, 0.579]$
D_Φ (mean $\pm$ SD)	40	0.347 ± 0.036	0.371 ± 0.046
D_Φ (95% CI)	40	$[0.336, 0.358]$	$[0.357, 0.385]$
site retention (macro mean)	40	0.965	0.600
site retention (95% CI)	40	$[0.951, 0.978]$	$[0.563, 0.634]$
paired TM wins	40	$7/40$	—
paired TM gap (95% CI)	40	$[-0.082, -0.053]$	—
Cliff's δ	40	-0.65	—
sign-test p	40	4.23×10^{-5}	—

Table 3: Auditable component-wise paired differences at COMPRESS $0.60L$ ($n = 40$). Δ is the per-protein design-minus-control distance; negative values favour ISOFORM. The final column is the exact two-sided sign test, with ties excluded.

term	mean Δ	ISOFORM lower / decisive	p
d_{struct}	$+0.0680$	$7/40$	4.23×10^{-5}
d_{motif}	-0.3384	$40/40$	1.82×10^{-12}
d_{embed}	$+0.3504$	$0/40$	1.82×10^{-12}
d_{prop}	$+0.1462$	$6/40$	8.36×10^{-6}
D_Φ	-0.0241	$26/40$	0.081

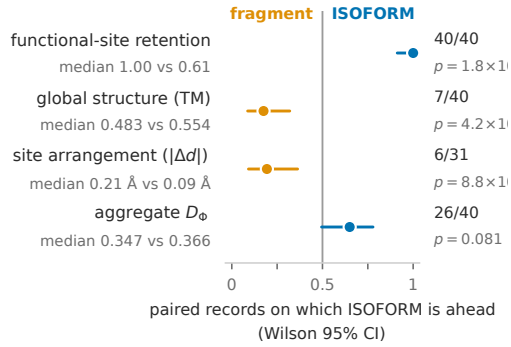
The structural result has the opposite direction. Mean evaluation-oracle TM is 0.487 ± 0.083 for ISOFORM and 0.555 ± 0.075 for the fragment. The design wins on $7/40$ pairs, giving Cliff's $\delta = -0.65$ and an exact two-sided sign-test $p = 4.23 \times 10^{-5}$. The protein is the unit of all three statistics, and the standard deviations are across proteins.

The component-wise means are $d_{\text{embed}} = 0.466$ versus 0.116 and $d_{\text{prop}} = 0.428$ versus 0.282 for the design and fragment, respectively. With the fixed weights in Equation 1, the weighted contributions (structure, motif, embedding, property) are $(0.205, 0.029, 0.070, 0.043)$ for ISOFORM and $(0.178, 0.147, 0.017, 0.028)$ for the fragment, summing to $D_\Phi = 0.347$ and 0.371. The aggregate and the structural endpoint answer different questions: the aggregate ranks the design because of the site term, while the independent structural branch ranks the fragment higher. Figure 2 reports the paired directions and the full geometry distribution; Figure 3 makes the term contributions and weight sensitivity explicit; no pooled residual term is used.

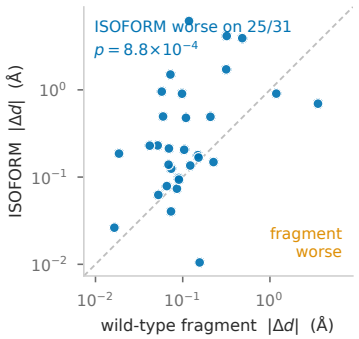
The paired component directions make this trade-off explicit. The design has lower structure distance on $7/40$ proteins, lower motif distance on $40/40$, lower embedding distance on $0/40$, and lower property distance on $6/40$; the aggregate distance is lower on $26/40$. These are paired sign tests over the same primary records, rather than independent tests over the four terms.

Question 2: what does site geometry add? The alignment-level site score is one-dimensional, so we also measure Cα displacement for matched annotated-site pairs. The released file contains 31 pooled, pre-filtered records without protein or budget identifiers; Figure 2 shows their paired values, medians, and empirical CDFs, while the numerical summary is in Appendix E. This is a secondary predicted-structure diagnostic, not a protein-level replicate or activity assay.

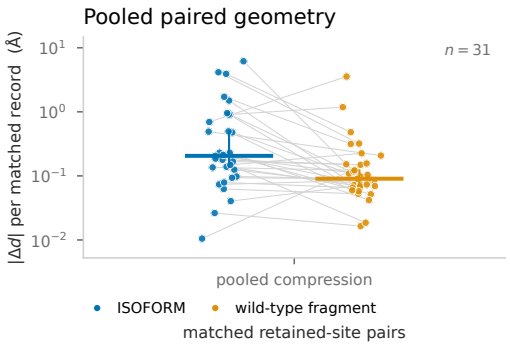
Question 3: what do simple and learned controls establish? The fragment is a strong structural reference but loses annotated sites at $0.6L$; Raygun, SCISOR, and the training-free editor underperform the corresponding simple constructions in the available runs. The $n = 40$ pooled coverage summary uses the same AlphaFold2 branch and is descriptive evidence in Appendix B, not an independent endpoint.



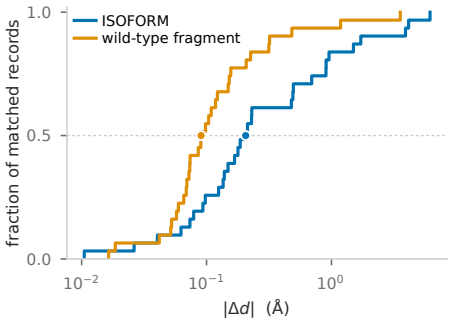
(a) Paired win rates with Wilson 95% intervals at the primary $0.60L$ cell; ties are excluded.



(b) Paired site-geometry scatter on equal log axes; points above the diagonal have larger ISOFORM displacement.



(c) Pooled matched geometry records with medians and marginal bootstrap intervals.



(d) Empirical CDF of the site-pair displacement, retaining the tail behaviour.

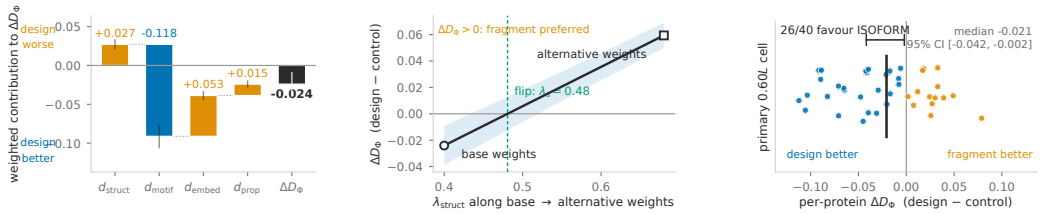
Figure 2: Compression dissociation at $0.60L$. The four views compare ISOFORM with the contiguous fragment without mixing units. Retention, TM, and D_Φ use the 40 primary proteins; geometry uses the 31 released matched records and is a secondary predicted-structure diagnostic.

Expansion and paraphrase. Expansion is a feasibility stress test because padding can preserve the wild type without a meaningful rewrite. In paraphrase, in-place substitution meets the identity ceilings, while masked-fill and BLOSUM controls show no consistent structural advantage; identity accounting and the embedding-free correction are in Appendices I and F.

Ablation scope. A structure-heavy retrain uses the same oracle-stage budget as the base. Across three seeds its paired TM win rate falls from 0.183 ± 0.014 to 0.158 ± 0.014 (seed 42: 0.175 to 0.150), so the ranking persists under this retraining protocol.

7 Conclusion

Meeting a rewrite budget and preserving a functional proxy are separable objectives. At COMPRESS $0.60L$, ISOFORM retains more annotated sites than the fragment (96.5% vs. 60.0%) but has lower independent-branch TM (0.487 vs. 0.555); D_Φ ranks it higher only because the site term offsets losses elsewhere. Reweighting the same records reverses this aggregate ranking near $\lambda_{\text{struct}} = 0.48$. Thus the paired protocol exposes a proxy trade-off, while expansion shows that length alone can be met by padding. The evidence is computational (40 pairs), and biochemical activity and orthogonal validation remain open.



(a) Weighted term contributions at the primary $0.60L$ cell; bars show means and bootstrap 95% intervals.

(b) Weight-only interpolation from the base to the alternative weights; the aggregate changes sign near $\lambda_{\text{struct}} = 0.48$.

(c) Per-protein ΔD_Φ distribution at the primary cell; negative values favour ISOFORM.

Figure 3: The favourable aggregate is a weight-sensitive proxy. All four oracle terms are distances and use the common sign convention Δ = design − control. The decomposition, weight-only sensitivity, and per-protein distribution are computed from the released 40 primary records. The alternative-weight curve is not a separately retrained model; no cross-budget component claim is made because term-level vectors for $0.75L$ and $0.90L$ are not released.

Reproducibility Statement

Appendix I specifies the model, data filters, controls, candidate selection, and oracle provenance. The release includes the evaluation CSV, run-log joins, and figure scripts; pooled or restricted subsets are identified in captions.

Ethics Statement

Low-identity protein rewriting may have dual-use implications. We therefore report identity-versus-proxy frontiers and restrict released examples to benign benchmark families. The computational oracle should not be interpreted as a substitute for experimental biosafety review or for activity assays.

AI use statement

Language tools were used for copy editing and consistency checks. The authors are responsible for the scientific claims, data, code, and final manuscript.

References

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X. Lu, Nicolo Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, 2023. doi: 10.1101/2023.09.11.556673.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. Quality controlled paraphrase generation. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2203.10940.

Ethan Baron, Alan N. Amin, Ruben Weitzman, Debora Marks, and Andrew Gordon Wilson. A diffusion model to shrink proteins while maintaining their function. *arXiv preprint arXiv:2511.07390*, 2025.

Daniel Berenberg, Jae Hyeon Lee, Simon Kelow, Ji Won Park, Andrew Watkins, Vladimir Gligorijević, Richard Bonneau, Stephen Ra, and Kyunghyun Cho. Multi-segment preserving sampling for deep manifold sampler. *arXiv preprint arXiv:2205.04259*, 2022.

Chaoran Cheng, Zhanghan Ni, Yanru Qu, Yuxin Chen, Ruihan Guo, Jiajun Fan, and Ge Liu. Variable-length generative protein design via generalized poisson flow. *arXiv preprint arXiv:2607.09039*, 2026.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Sam Pellock, Doug Tischer, Frederick Chan, Bruce Koepnick, Hannah Nguyen, Alex Kang, Banumathi Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Nicolas Deutschmann, Constance Ferragu, Jonathan D. Ziegler, Shayan Aziznejad, and Eli Bixby. EvoFlows: Evolutionary edit-based flow-matching for protein engineering. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026. arXiv:2603.11703.

Kapil Devkota, Daichi Shonai, Joey Mao, Yeon Soo Ko, Wei Wang, Scott Soderling, and Rohit Singh. Miniaturizing and modifying natural proteins with raygun. *Nature*, 2026. doi: 10.1038/s41586-026-10842-8. Online ahead of print, 29 July 2026.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. Sentence compression by deletion with LSTMs. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015. doi: 10.18653/v1/d15-1042.

Vladimir Gligorijević, Daniel Berenberg, Stephen Ra, Andrew Watkins, Simon Kelow, Kyunghyun Cho, and Richard Bonneau. Function-guided protein design by deep manifold sampling. *bioRxiv*, 2021. doi: 10.1101/2021.12.22.473759. Preprint, not journal-published.

Joshua C. Grieger and R. Jude Samulski. Packaging capacity of adeno-associated virus serotypes: impact of larger genomes on infectivity and postentry steps. *Journal of Virology*, 79(15):9933–9944, 2005.

Jiatao Gu and Xiang Kong. Fully non-autoregressive neural machine translation: Tricks of the trade. *Findings of ACL-IJCNLP*, 2021. arXiv:2012.15833.

Jiatao Gu, Changhan Wang, and Jake Zhao. Levenshtein transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1905.11006.

Marton Havasi, Brian Karrer, Itai Gat, and Ricky T. Q. Chen. Edit flows: Variable length discrete flow matching with sequence-level edit operations. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09018.

Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42:275–283, 2024.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

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

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018. doi: 10.18653/v1/n18-1170.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michal Pacholska, Tobias Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. doi: 10.18653/v1/d16-1140.

Jeongchan Kim, Yunkyung Ko, and Jong Chul Ye. LPDP: Inference-time reward control for variable-length DNA generation with edit flows. *arXiv preprint arXiv:2605.11368*, 2026.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8, 2020. doi: 10.1162/tacl_a_00318.

Jakub Lála, Ayham Al-Saffar, and Stefano Angioletti-Uberti. BAGEL: Protein engineering via exploration of an energy landscape. *PLoS Computational Biology*, 2025. doi: 10.1371/journal.pcbi.1013774.

Jakub Lála, Harsh Agrawal, Fanfei Dong, Jude Wells, and Stefano Angioletti-Uberti. An energy landscape approach to miniaturizing enzymes using protein language model embeddings. *bioRxiv*, 2026. doi: 10.64898/2026.03.04.709378. Preprint.

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

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023.

Eric Malmi, Yue Dong, Jonathan Mallinson, and Aleksandr Chuklin. Text generation with text-editing models. *arXiv preprint arXiv:2206.07043*, 2022. NAACL 2022 Tutorial.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.1101/2021.07.09.450648.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.1101/2023.12.07.570727.

Dhruvesh Patel, Aishwarya Sahoo, Avinash Amballa, and Tahira Naseem. Insertion language models: Sequence generation with arbitrary-position insertions. *arXiv preprint arXiv:2505.05755*, 2025.

Dhruvesh Patel, Benjamin Rozonoyer, Soumitra Das, and Tahira Naseem. A continuous-time markov chain framework for insertion language models. *arXiv preprint arXiv:2606.10199*, 2026.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. doi: 10.18653/v1/d19-1410.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Emil Sharafutdinov and Ingemar André. Tree-conditioned edit flows for ancestral sequence reconstruction. *arXiv preprint arXiv:2605.04119*, 2026.

Luca Alessandro Silva, Barthelemy Meynard-Piganeau, Carlo Lucibello, and Christoph Feinauer. Fast uncovering of protein sequence diversity from structure. *International Conference on Learning Representations (ICLR)*, 2025. Spotlight; arXiv:2406.11975.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in three dimensions for the motif-scaffolding problem. In *International Conference on Learning Representations (ICLR)*, 2023.

Robert Verkuil, Ori Kabeli, Yilun Du, Basile I. M. Wicky, Lukas F. Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, and Alexander Rives. Language models generalize beyond natural proteins. *bioRxiv*, 2022.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.18567.

Xinyou Wang, Liang Hong, Jiasheng Ye, Zaixiang Zheng, Yu Li, Shujian Huang, and Quanquan Gu. Towards a generative protein evolution machine with DPLM-Evo. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.00182.

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*, 620:1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

Angela Y. Wu, Quentin Trolliet, Abhinav Rajendran, Jakub Lála, and Stefano Angioletti-Uberti. Mind the gap: An embedding guide to safely travel in sequence space. *PLoS Computational Biology*, 2026. doi: 10.1371/journal.pcbi.1014433.

Zhijian Wu, Hongyan Yang, and Peter Colosi. Effect of genome size on AAV vector packaging. *Molecular Therapy*, 18(1):80–86, 2010.

Daiheng Zhang, Shiyang Zhang, Sizhuang He, Yangtian Zhang, Syed Asad Rizvi, and David van Dijk. STRIDE: Post-training LLMs to reason and refine bio-sequences via edit trajectories. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. arXiv:2603.03573.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1904.09675.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

Yang Zhang and Jeffrey Skolnick. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005. doi: 10.1093/nar/gki524.

Xinyi Zhou, Guanghui Shen, Yifan Chen, Gang Chen, and Pheng-Ann Heng. DivPro: diverse protein sequence design with direct structure recovery guidance. *Bioinformatics*, 2025. doi: 10.1093/bioinformatics/btaf258.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018. doi: 10.1145/3209978.3210080.

A THE SYSTEM IN FULL

The system receives a wild-type sequence, an operation, and a scalar budget. The operation embedding selects compression, expansion, or paraphrase; the budget embedding supplies the target ratio or identity ceiling; and the wild-type embedding supplies a functional context. The trunk produces deletion and insertion decisions in parallel. During training, the four-term preservation surrogate is included in the objective. During inference, eight candidates are ranked with the normalized training-oracle score, and only the selected candidate is sent to the frozen evaluation pipeline. The KEEP mask is a compression-time deletion constraint; alignment-level site retention is measured after decoding.

B BASELINE COVERAGE AND RUN AVAILABILITY

The baselines fall into three groups: learned length-changing generators (Baron et al., 2025; Devkota et al., 2026; Deutschmann et al., 2026; Wang et al., 2026; Zhang et al., 2026; Cheng et al., 2026); per-target search (Wu et al., 2026; Lála et al., 2025; 2026); and diversity-oriented design (Zhou et al., 2025; Silva et al., 2025; Dauparas et al., 2022; Hsu et al., 2022). Raygun and SCISOR are evaluated end to end under the same structural pipeline. For other methods we report capability coverage rather than a numerical comparison.

Table 4: Descriptive pooled evaluation-oracle TM for systems run end to end. Each available value averages the same $n = 40$ held-out proteins over the reported operation budgets; an em dash denotes an unevaluated pair. † marks a verbatim wild-type-preserving expansion. These summaries are retained as coverage checks rather than independent primary endpoints.

	n	compression	expansion
simple construction	40	0.710	$0.907^{\dagger}$
ISOFORM	40	0.664	0.801
SCISOR	40	0.630	—
Raygun	40	0.415	0.518
ESM2 editor	40	0.374	0.617

C THE FRONTIER IN FULL

Table 5: The preservation–divergence frontier, per operation and budget. Means over the same $n = 40$ held-out proteins in every row; paired comparisons are pooled per operation and given in the text. All listed budgets hit exactly (hit $= 1.00$ throughout). ident is the realised sequence identity to the wild type. TM is the TM-score from the AlphaFold2 structural branch for the design and $\mathrm{TM}^{\mathrm{sub}}$ that of the operation's simple construction (Section 4). The operation-level reference-homologue TM is constant within each group (0.804 for compression, 0.947 for expansion, and 0.950 for paraphrase) and is therefore reported here rather than as a repeated column. * marks a construction the feasibility test demotes to a structural reference; † marks one that retains the wild type verbatim and is therefore not a rewrite. The primary endpoint is the COMPRESS $0.60L$ row; the PARAPHRASE row is shown for completeness.

op	budget	ident	n	TM	$\mathrm{TM}^{\mathrm{sub}}$
COMPRESS	$0.90L$	0.855	40	0.834	0.862
	$0.75L$	0.722	40	0.671	0.714*
	$0.60L$	**0.549**	**40**	**0.487**	**0.555***
EXPAND	$1.10L$	0.899	40	0.900	0.932†
	$1.25L$	0.872	40	0.862	0.915†
	$1.50L$	0.745	40	0.642	0.874†
PARAPHRASE	0.90	0.900	40	0.957	0.967
	0.80	0.800	40	0.918	0.952
	0.70	0.701	40	0.933	0.939

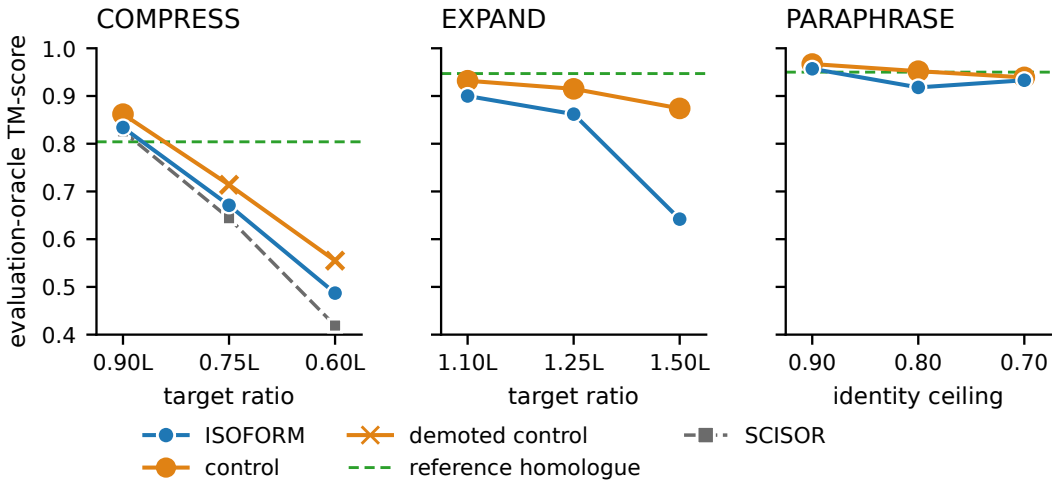


Figure 4: Preservation–divergence frontier. Evaluation-oracle TM across the commanded budgets ($n = 40$ per point). ISOFORM is blue, the operation-specific construction is orange, the reference homologue is green dashed, and SCISOR is gray. A cross marks a construction demoted to a structural reference by the feasibility check; all panels use the same $[0.4, 1.0]$ TM scale. Expansion is a specification stress test rather than evidence of successful rewriting.

D PER-PROTEIN SIGNIFICANCE, BY BUDGET

Because a wild type's budgets are not independent, every significance statement in this paper is a two-sided sign test over the 40 held-out proteins rather than over the 120 or 240 designs (Section 4). Table 6 gives the per-budget counts behind the pooled values quoted in Section 6; the pooled row is the protocol value we report.

Table 6: External systems against the simple construction, per budget. *ahead* counts proteins on which the system has a strictly higher TM-score from the AlphaFold2 structural branch; *decisive* excludes ties, so the numerator/denominator is reported as ahead/decisive. The p column is the exact two-sided sign-test value over those decisive pairs. The pooled row is the protocol value reported in the body; it is not the mean of the rows above it, because a wild type's budgets are not independent.

system	budget	ahead / decisive	p
Raygun	COMPRESS $0.60L$	$0/40$	1.8×10^{-12}
Raygun	COMPRESS $0.75L$	$0/40$	1.8×10^{-12}
Raygun	COMPRESS $0.90L$	$1/40$	7.5×10^{-11}
Raygun	COMPRESS pooled	$0/40$	1.8×10^{-12}
Raygun	EXPAND $1.10L$	$3/40$	2.0×10^{-8}
Raygun	EXPAND $1.25L$	$1/40$	7.5×10^{-11}
Raygun	EXPAND $1.50L$	$1/40$	7.5×10^{-11}
Raygun	EXPAND pooled	$1/40$	7.5×10^{-11}
SCISOR	COMPRESS $0.60L$	$3/40$	2.0×10^{-8}
SCISOR	COMPRESS $0.75L$	$5/40$	1.4×10^{-6}
SCISOR	COMPRESS $0.90L$	$7/40$	4.2×10^{-5}
SCISOR	COMPRESS pooled	$2/40$	1.5×10^{-9}

E SITE GEOMETRY: PROCEDURE AND THE RESTRICTED SET

For every ordered pair of annotated KEEP sites retained in both the wild type and a candidate we compute $|d_{\text{cand}}(a, b) - d_{\text{WT}}(a, b)|$ between Cα atoms in the cached predicted structures, mapping wild-type sites onto the candidate with the same Needleman–Wunsch alignment the oracle uses. Cells with fewer than two jointly retained annotated sites are excluded from the pairwise geometry calculation. The number excluded is reported by operation and budget; an excluded cell is not treated as a zero displacement.

The design and the controls do not always retain the same sites, so a naive comparison can score them on different subsets. The pooled compression geometry analysis has a different denominator from the primary $0.6L$ retention table. We therefore report its retention and matched-pair counts separately and do not use them as the primary estimand. Table 7 restricts the released compression records to matched comparisons in which design and control retain the *same number* of site pairs, and aggregates before the sign test. We report this restricted analysis as a sensitivity analysis and do not interpret it as an unbiased estimate of all operation–budget cells.

The geometry figure reports the pooled matched-pair summary in the geometry views of Figure 2. The released geometry file contains 31 pre-filtered matched records but no protein identifiers or budget labels, so these values are secondary predicted-structure diagnostics rather than an independent protein-level estimate. The primary claim uses the $0.6L$ retention and TM table above. All quantities are computed on predicted structures and bear on geometry, not on measured activity.

Table 7: Restricted compression site-geometry summary. The released file contains 31 anonymised matched records and no operation or budget identifiers; design and control are restricted to the same number of site pairs. Lower displacement is better; bold marks the lowest value. p is the paired sign-test value.

operation	n	design	control	homologue	p
compression (pooled)	31	0.80	**0.27**	0.69	8.8×10^{-4}

F THE PARAPHRASE LADDER AND THE COUPLED-METRIC CORRECTION

d_{embed} is defined by ESM-2, which is both our training signal ($\lambda_{\text{e}} = 0.15$) and the generator of the masked-fill control, so a comparison that keeps it flatters whichever side owns ESM. Excluding it consistently from all three comparisons leaves a win against uniform substitution (-0.051), a tie against BLOSUM62 (-0.011—the same rule downgrades a win we had claimed) and a tie-to-loss

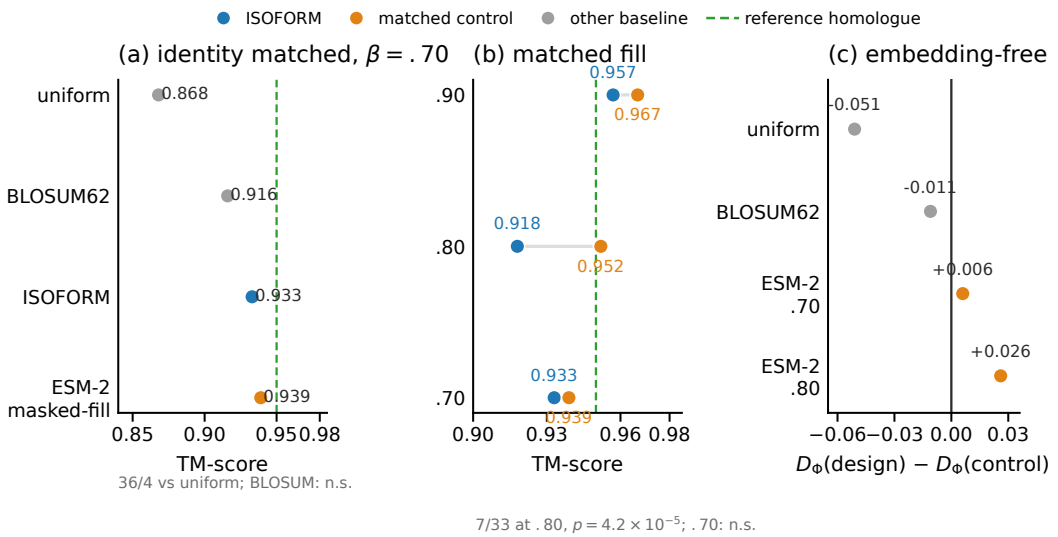


Figure 5: The identity-matched paraphrase ladder ($n = 40$ held-out proteins). Panel (a) shows TM-score for the design and controls at $\beta = 0.70$; the dashed line is the reference-homologue mean. Panel (b) shows paired means for ISOFORM and matched ESM-2 masked fill across identity ceilings. Panel (c) recomputes the aggregate difference after removing the ESM-defined embedding term from both sides; negative values favour ISOFORM. Sign-test results are reported in the text.

against ESM-2 masked-fill ($+0.006$). It does not rescue $\beta = 0.80$ ($+0.026$), where the design remains behind the control. Substitution conservatism tracks the budget: mean BLOSUM62 at substituted sites rises from -0.09 at ceiling 0.70 through -0.01 at 0.80 to $+0.31$ at 0.90, against homologues' $+0.24$.

G PARAPHRASE DECODING

For fixed-length paraphrase, deletion and insertion are implemented as in-place substitution: each selected position is remasked and refilled while the original residue is forbidden. The control uses the same identity-matching rule at each ceiling. This construction keeps the amount of sequence change comparable across the paraphrase ladder in Section 6.

H ABLATIONS AND REPRODUCIBILITY CHECKS

The structure-heavy rebalanced retrain is the only retraining ablation reported here. It uses 20k oracle-stage steps, matched to the base model. Across three training seeds its paired TM win rate is 0.158 ± 0.014, compared with 0.183 ± 0.014 for the base; for seed 42 the values are 0.150 and 0.175.

I IMPLEMENTATION DETAILS

Backbone and conditioning. The trunk is initialised from `esm2_t33_650M_UR50D` (650M parameters, context length 1024). A learned operation embedding, a scalar budget embedding, and the wild-type functional embedding are injected by FiLM modulation into the query and value projections of all 33 transformer blocks (hidden dimension 1280). Deletion and insertion heads are two-layer MLPs with hidden dimension 1280 and GELU activations; the token head projects from 1280 to the 33-token output vocabulary. For compression, the KEEP mask is derived from UniProt active-, binding-, and metal-site annotations and sets deletion logits to zero at those source positions. The mask is a generation-level deletion constraint. The reported site-retention statistic is computed after generation by aligning the candidate back to the wild type with Needleman–Wunsch and taking the

macro-average of retained-site fractions. Alignment gaps or shifts can therefore reduce the reported retention even when the source token was protected; we do not call this a violation of the mask.

Data and split. Rewrite pairs are mined from UniRef90 (2023_03) with MMseqs2 and aligned by Needleman–Wunsch with gap-open 11 and gap-extension 1. Identity thresholds are $\geq 30\%$ for COMPRESS and EXPAND and 30–80% for PARAPHRASE. The mined counts are 1.24M, 0.82M, and 2.51M pairs for COMPRESS, EXPAND, and PARAPHRASE, respectively. We generate 150k synthetic pairs: 80k minimal-domain/full-length pairs for COMPRESS and 70k structure-guided held-out insertions for EXPAND. The held-out set is disjoint from training at the UniRef90 cluster level, with maximum train–test identity below 25%.

Optimisation. We use AdamW with $(\beta_1, \beta_2, \epsilon) = (0.9, 0.98, 10^{-6})$, peak learning rate 10^{-4}, 10,000 linear-warmup steps, and cosine decay to 10^{-6}. The batch is 256 sequences per step accumulated over eight NVIDIA A100 80GB GPUs (effective batch size 2048). Base training runs for 200k steps followed by 20k oracle-stage steps; the rebalanced model uses the same 20k oracle-stage budget. Wall-clock time is approximately 14 days for base training and 3 days for oracle-stage training. The loss coefficients are $\alpha_1 = 0.5$, $\alpha_2 = 0.1$, $\alpha_3 = 2.0$, and $\alpha_4 = 1.0$. The reported oracle weights are $(0.40, 0.35, 0.15, 0.10)$; the rebalanced model uses $(0.68, 0.12, 0.12, 0.08)$.

Identity accounting. Commanded identity is measured on the global Needleman–Wunsch alignment and reported to three decimals. Because the identity is a ratio over the alignment length A, a one-residue change shifts the value by $1/A$. The paraphrase ceiling therefore uses a predefined ± 0.01 reporting tolerance; at $\beta = 0.70$ the mean is 0.701 and the maximum is 0.704. These one-step overshoots are reported as quantisation effects and are not interpreted as functional gains.

Table 8: Realised identity distribution for PARAPHRASE at ceiling $\beta = 0.70$ ($n = 40$). The bins reflect alignment quantisation; no sample is more than one step above the ceiling.

identity	count	fraction
0.696	3	7.5%
0.700	22	55.0%
0.704	15	37.5%

Oracle provenance. The training oracle uses ESMFold v1 for d_{struct}, alignment-restricted site correspondence for d_{motif}, ESM2-650M embeddings for d_{embed}, and DeepFRI, a ProtT5 enzyme-family classifier, and a geometric metal-site predictor for d_{prop}. Fluency is ESM-1v pseudo-log-likelihood. The evaluation structural oracle uses AlphaFold2 in single-sequence mode, with no MSA and three recycles, followed by TM-align. This structural branch is independent of the training structural oracle. ESM2 embeddings and the property predictors are shared between training and evaluation and are therefore reported as diagnostic components rather than independent evidence. Every TM-score in the paper comes from the AlphaFold2 structural branch.

Inference and candidate selection. For each prompt we generate $N = 8$ candidates at temperature $T = 1.0$. Candidates are ranked by the complete normalized training-oracle distance, including all four terms in Equation 1; the candidate identifier, score, and selected flag are recorded. The single highest-ranked candidate is passed once to the frozen evaluation pipeline. No evaluation-oracle score is used during selection. All 40 proteins in the primary table are shorter than the 1024-residue context limit, so no sliding-window decode is used for that analysis.

External controls and statistics. Raygun uses the largest released 8.8M model with top-pseudo-likelihood decoding and stochastic-decode noise fixed at 0.1. SCISOR uses the largest released 650M U90 checkpoint at exact target lengths and the same source-site annotations. Both are scored through the same AlphaFold2 structural branch. Per-protein bootstrap uses 10^4 resamples with a fixed seed. Exact two-sided sign tests and paired Cliff's delta are computed over the 40 proteins; no budget sweep is treated as an independent replicate. `check_i` and `check_f` fields in earlier tables duplicated the reported D_Φ values and are replaced by recomputed values and absolute check errors in the release.

J REPRODUCIBILITY AND RELEASE

The release contains the source code, model configuration, oracle configurations, per-protein evaluation table, and scripts used to regenerate the figures. The primary conclusions use the COMPRESS $0.6L$ table; biochemical validation remains outside the scope of this study. Components specialised to hazardous protein families are not released.