Miniaturizing, Modifying, and Magnifying Nature's Proteins with Raygun

Gökçe Uludoğan
PhD Candidate

LifeLU Reading Group | 17 April 2025

Kapil Devkota, Daichi Shonai, Joey Mao, Young Su Ko, Wei Wang, Scott Soderling, and Rohit Singh
bioRxiv preprint
doi.org/10.1101/2024.08.13.607858

Motivation

Context: Protein engineering landscape

- Rapid advances in de novo protein design (e.g., diffusion-based approaches, structure-based models) have expanded our ability to create novel proteins.

- However, engineering new proteins “from scratch” is not always ideal, especially when modifying an existing protein is more practical or necessary.

Motivation

The key gap: Template-based design with large-scale changes

- Typical template-based methods **excel at point substitutions** but **struggle with large insertions/deletions** (indels).

- Nature's proteins evolve through **combined substitutions and indels**, often leading to **significant structural or functional innovations**. Replicating that computationally has been a challenge.

Motivation

A need for a framework that:

- Uses an existing protein as a starting point (a “template”).

- Preserves overall fold and function while allowing significant deletions or insertions.

- Produces variants that remain structurally coherent and potentially functional, even if length changes by 10–50% or more.

Motivation

A need for a framework that:

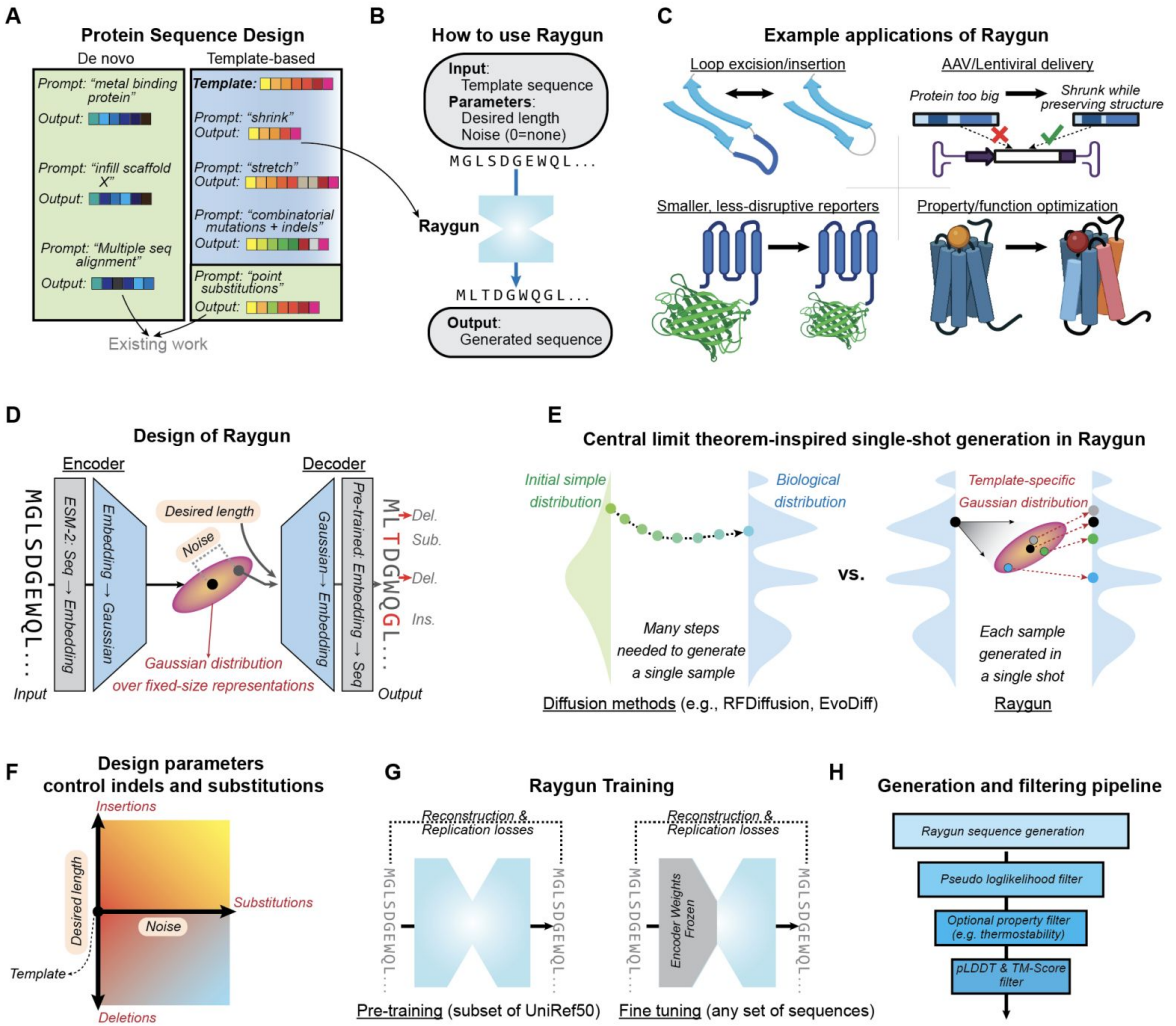
- Uses an existing protein as a starting point (a “template”).
- Preserves overall fold and function while allowing significant deletions or insertions.
- Produces variants that remain structurally coherent and potentially functional, even if length changes by 10–50% or more.

  > Raygun

  a generative framework that leverages protein language model (PLM) embedding but goes well beyond point substitutions

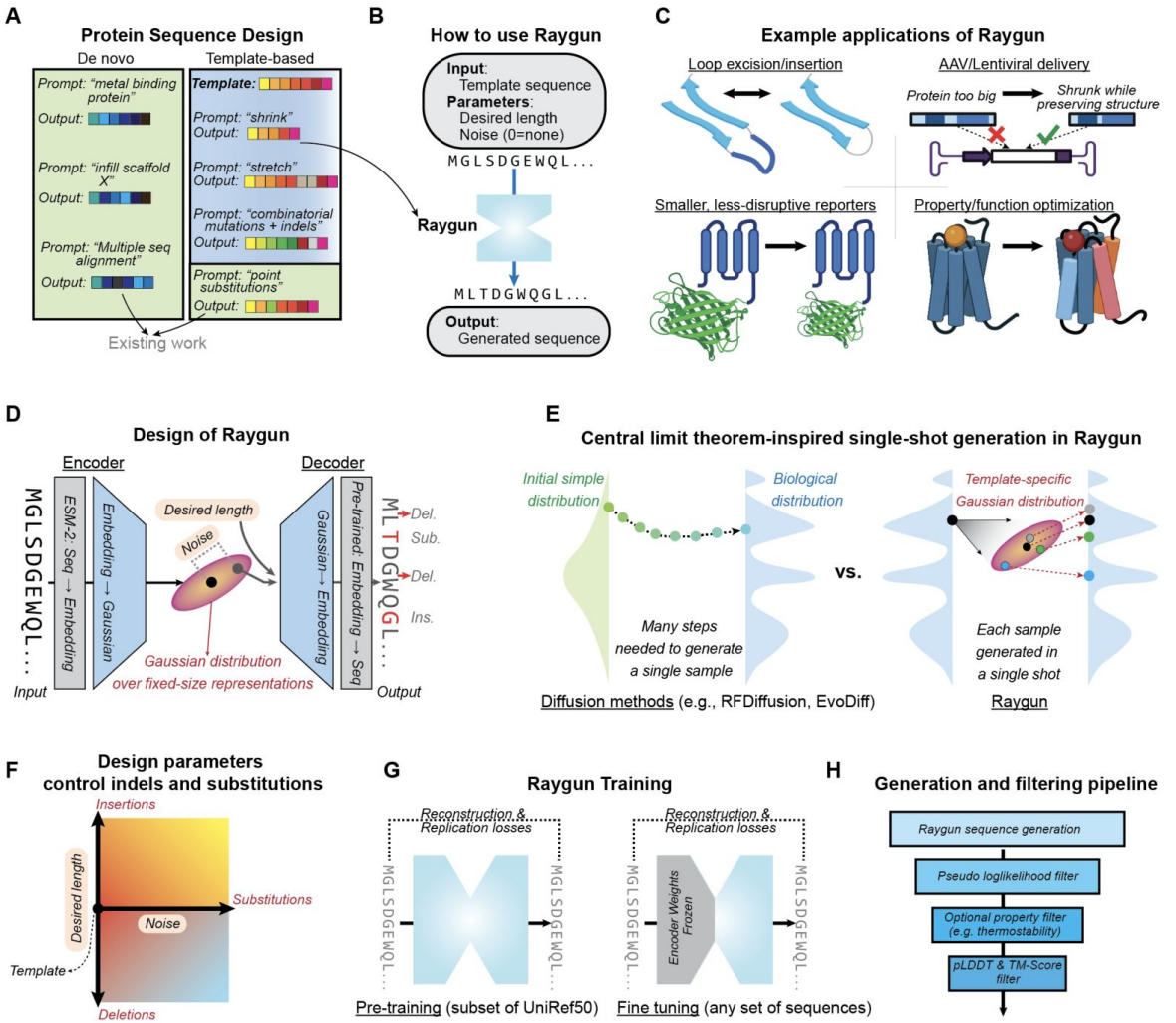
Raygun

- A **template-based** design model



Raygun

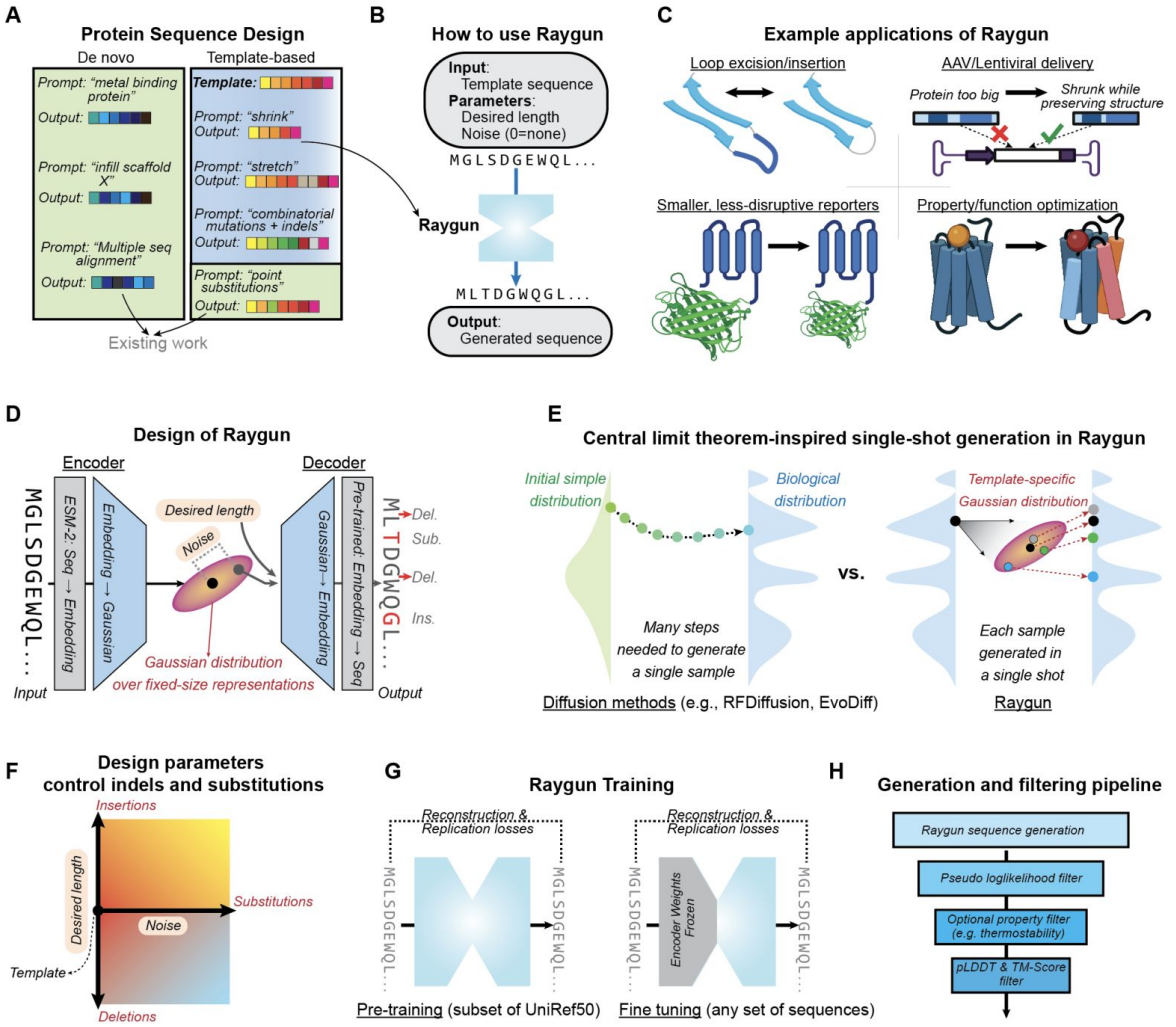
- A **template-based** design model
- It converts any protein sequence into a **probabilistic, fixed-size representation**, then generates new sequences by **sampling from that distribution.**



Raygun

Its novelty:

Mapping **variable-length PLM embeddings** into a **multivariate normal distribution** (MVN) of fixed dimensionality.



Raygun

Encoding

ESM-2 Residue-Wise Embeddings

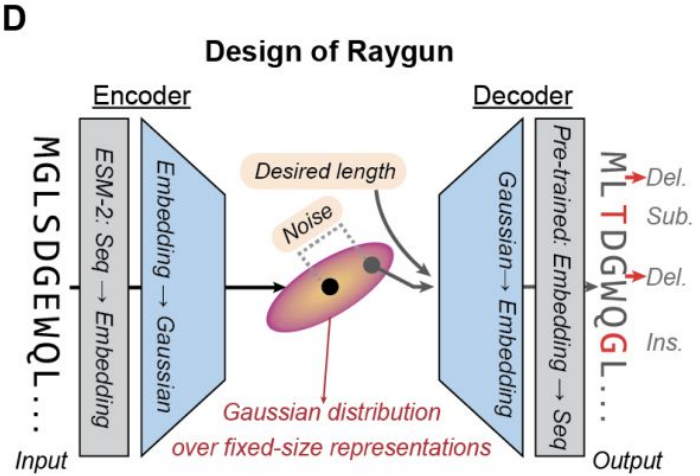
Leverages ESM-2's masked language modeling capabilities tc
capture rich local and global protein context for each residue

Partition into 50 Contiguous Blocks

Divides the variable-length embedding into K = 50 segments
(e.g., each segment is 10 residues for a 500-length template).

64,000-Dimensional MVN

Within each block, Raygun computes mean and standard
deviation, forming a fixed-size (64k-dim) Multivariate Normal
Distribution (MVN).



Raygun

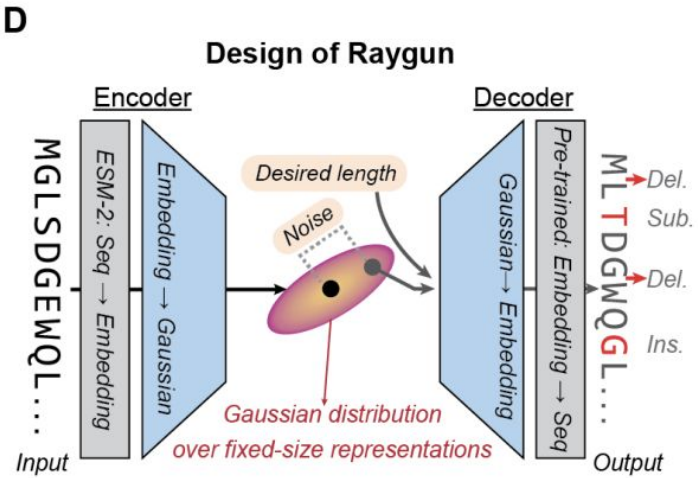
Encoding

Probabilistic Representation

The MVN's mean and covariance serve as a tractable distribution to enable single-shot sampling for new sequences.

Central Limit Theorem

Averaging over block-length embeddings approximates a normal distribution, aiding in compressing variable-length input while retaining sufficient structure to be decoded back to a protein sequence.



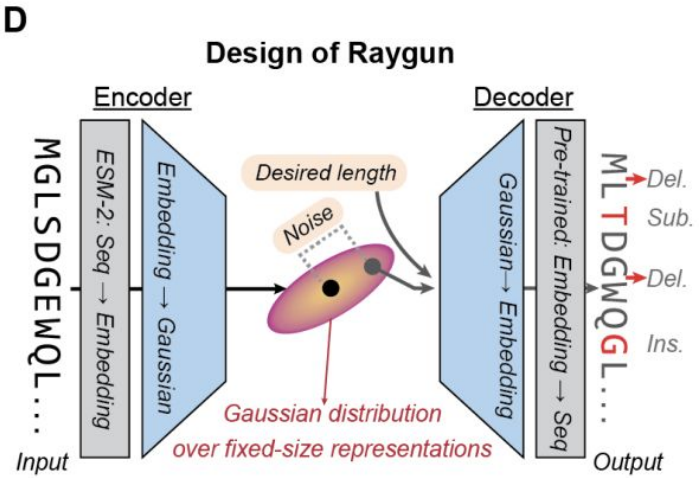
Raygun

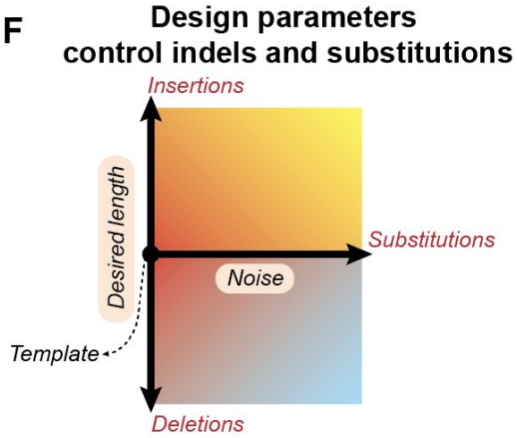
Sampling / Noise addition

Raygun samples from the MVN, scaled by a user-defined “noise” parameter.

Larger noise ⇒ more substitutions.

Adjusting desired output length ⇒ insertions or deletions



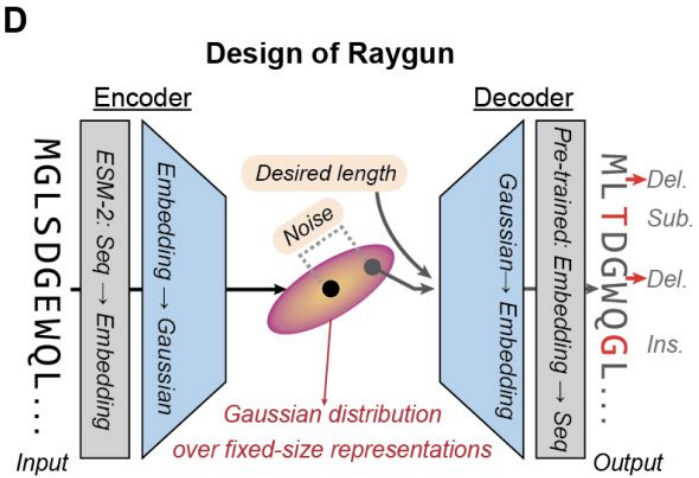


Raygun

Decoding

A second stage (decoder) reconstructs a new variable-length set of embeddings from the MVN sample.

A small neural net then maps those embeddings back to a concrete protein sequence.

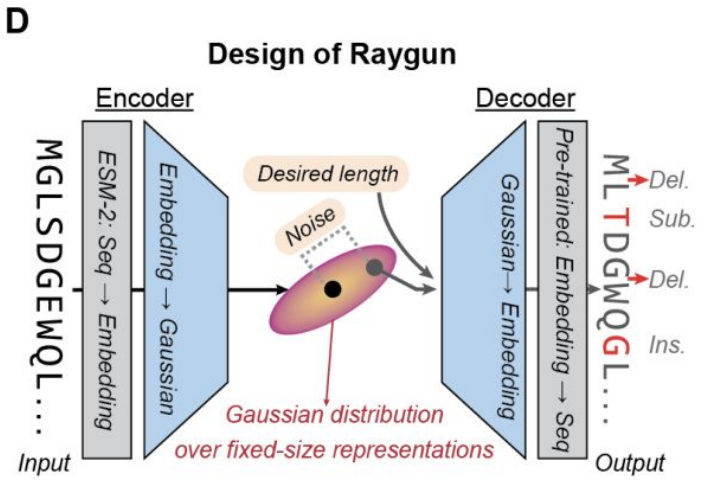


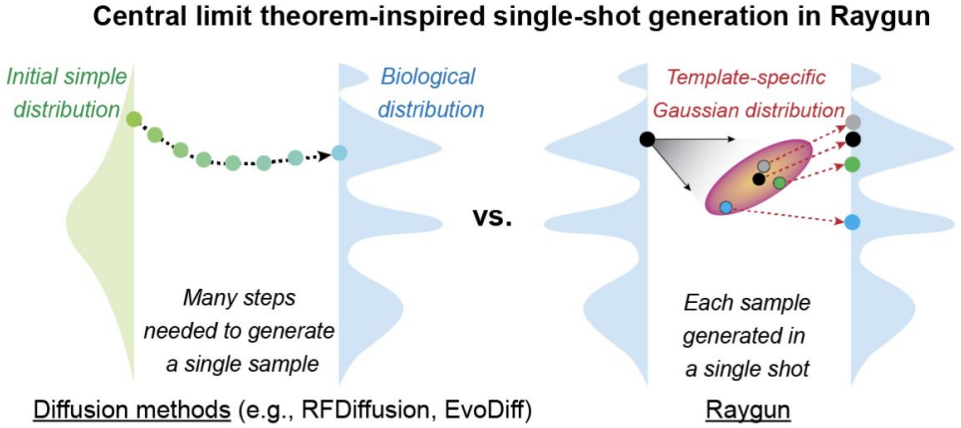
Raygun

Fixed-length representation rationale

Allows Raygun to handle any protein length by working in a single, uniform dimension internally.

Exploits the Central Limit Theorem concept that averaging contiguous embeddings approximates normality.





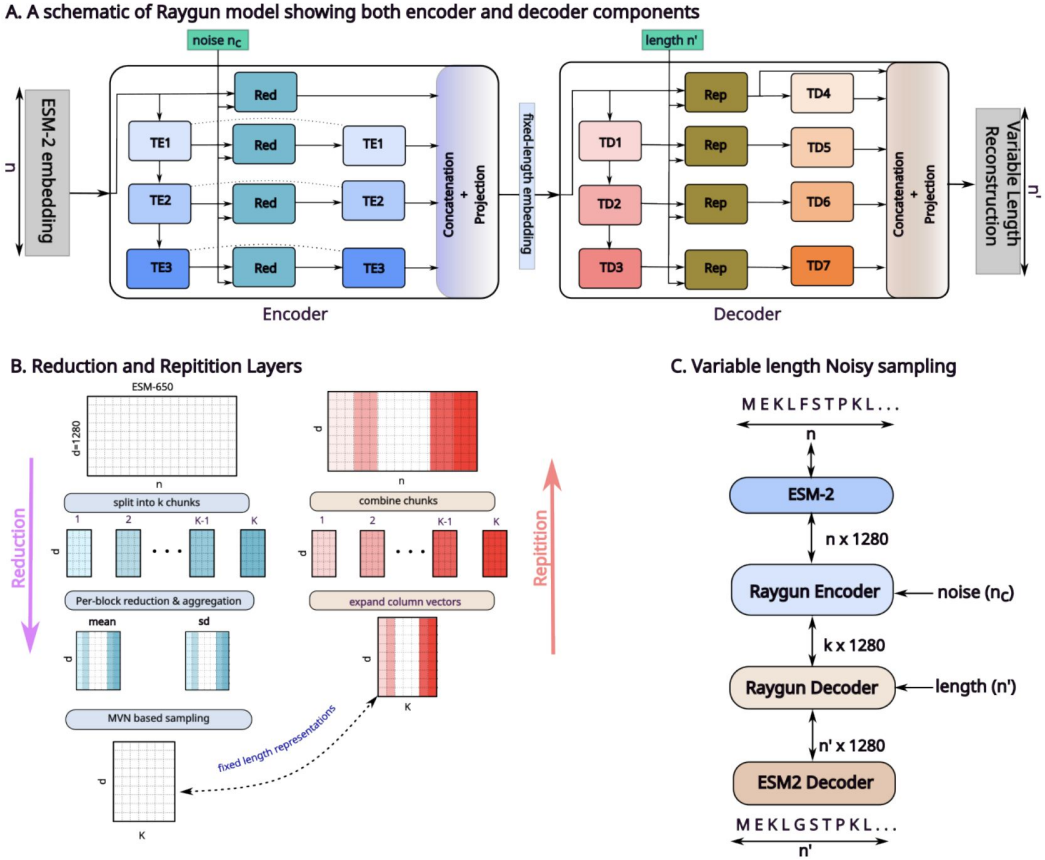
Raygun architecture

Encoder-decoder

- Encoder: T-Block & Reduction
- Decoder: T-Block & Repetition

T-Block Layers

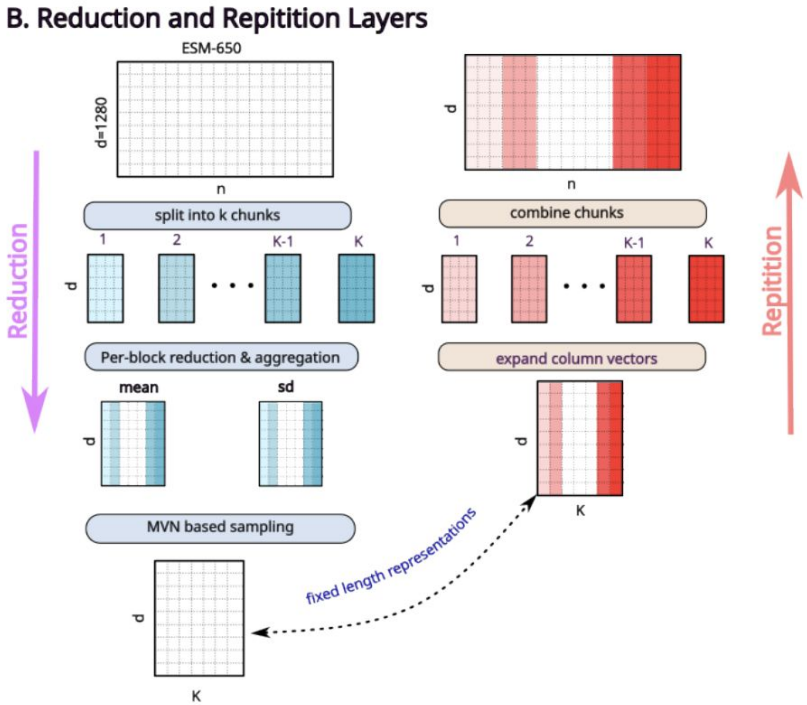
- Present in both encoder and decoder.
- Each T-Block is made up of:
  - ESM Transformer module for learning global context.
  - 1D Convolution block for enhancing local relationships.
- Surround and sit between the length-transforming layers to refine embeddings at each stage.
- Contain the majority of Raygun's ~701M parameters.



Raygun architecture

Length-Transforming Layers

- Reduction Layer (Encoder Stage)
  - Parameter-free.
  - Aggregates each of the 50 blocks into mean and standard deviation, capturing the Multivariate Normal Distribution (MVN).
  - During training, uses the block's mean.
  - During inference, samples from (mean, stdev) scaled by a user-defined noise parameter.
- Repetition Layer (Decoder Stage)
  - Parameter-free.
  - Expands the fixed-size MVN output back into a variable-length embedding according to the desired sequence length.

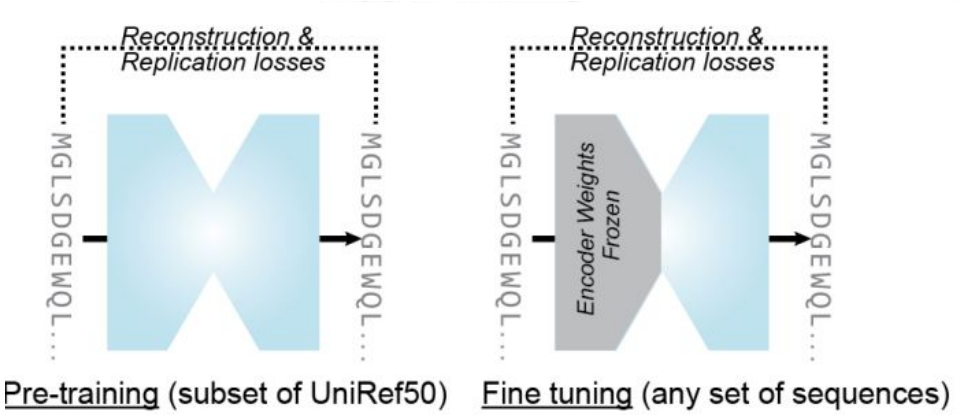


Raygun training

Self-supervised on ~80k proteins from UniRef-50, spanning lengths 100–1,000.

Three losses:

- **Sequence cross-entropy loss:** Compares decoded sequence to the original, using a pre-trained ESM-2-decoder.
- **Reconstruction loss:** Penalizes embedding deviations between input and output.
- **Replication loss**: Enforces consistency in the fixed-length representation when decoding to a shorter length and re-encoding.



Trained as **an autoencoder:**

Model compresses each input protein to a fixed-size MVN, then decompresses it back to the original length.

Raygun training

Self-supervised on ~80k proteins from UniRef-50, spanning lengths 100–1,000.

Three losses:

- **Sequence cross-entropy loss:** Compares decoded sequence to the original, using a pre-trained ESM-2-decoder.
- **Reconstruction loss:** Penalizes embedding deviations between input and output.
- **Replication loss**: Enforces consistency in the fixed-length representation when decoding to a shorter length and re-encoding.

Trained as **an autoencoder:**

Model compresses each input protein to a fixed-size MVN, then decompresses it back to the original length.

Let p be the input protein of length n, p' the Reconstructed Raygun sequence, ESM_p the ESM-650M embedding of p, $RGUN_p^{(50)}$ be the fixed length encoding of p and $RGUN_p^n$ be the reconstruction of p using Raygun to length n. Let p'' be another Raygun sequence obtained from p of length $n' < n$. Then the total and constituent losses become:

$$L_{total} = L_{ce} + L_{rr} + L_{rp} \tag{1}$$

$$L_{ce} = CrossEntropy(p, p') \tag{2}$$

$$L_{rr} = \|ESM_p - RGUN_p^n\|_2 \tag{3}$$

$$L_{rp} = \|RGUN_p^{(50)} - RGUN_{p''}^{(50)}\|_2 \tag{4}$$

Raygun validation

BLOSUM-weighted sequence identity score on validation proteins

- Ranges between -1 and 1

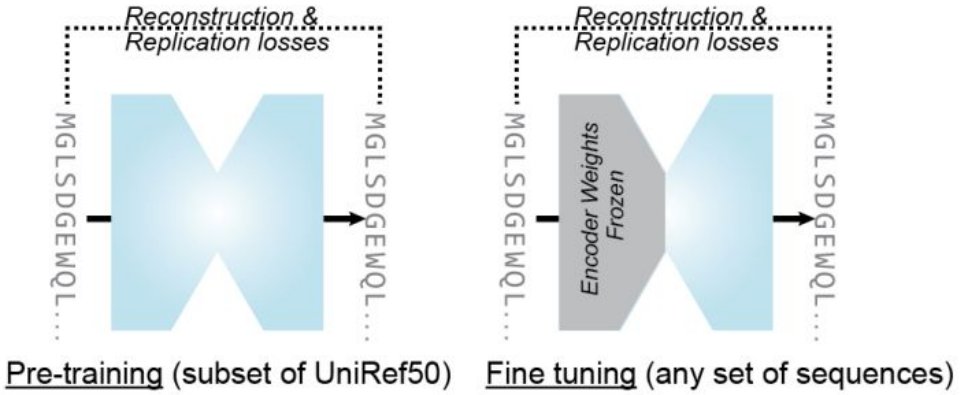
We used a BLOSUM-based sequence identity score, or simply "BLOSUM score" to evaluate the accuracy of Raygun on the validation dataset. Given the template protein S and the predicted Raygun candidate S', this score is computed as the ratio:

$$BLOSUM\ score(S, S') = \frac{\sum_i BLOSUM62(S(i), S'(i))}{\sum_i BLOSUM62(S(i), S(i))} \quad (5)$$

Raygun fine tuning

The model can be fine-tuned on **a specific protein family** or on **a set of related sequences** for higher fidelity decoding.

Only the decoder is unfrozen during fine-tuning, preserving the learned encoder.



Raygun generation and filtering pipeline

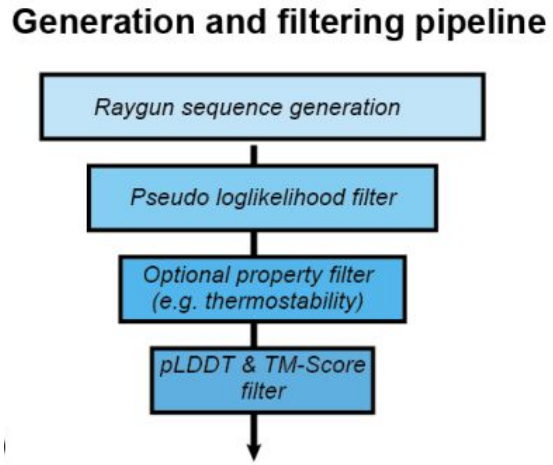
Raygun can generate thousands of candidates quickly (~0.3s each on a GPU).

Single-Shot Generation: Raygun can directly sample diverse candidates from the MVN.

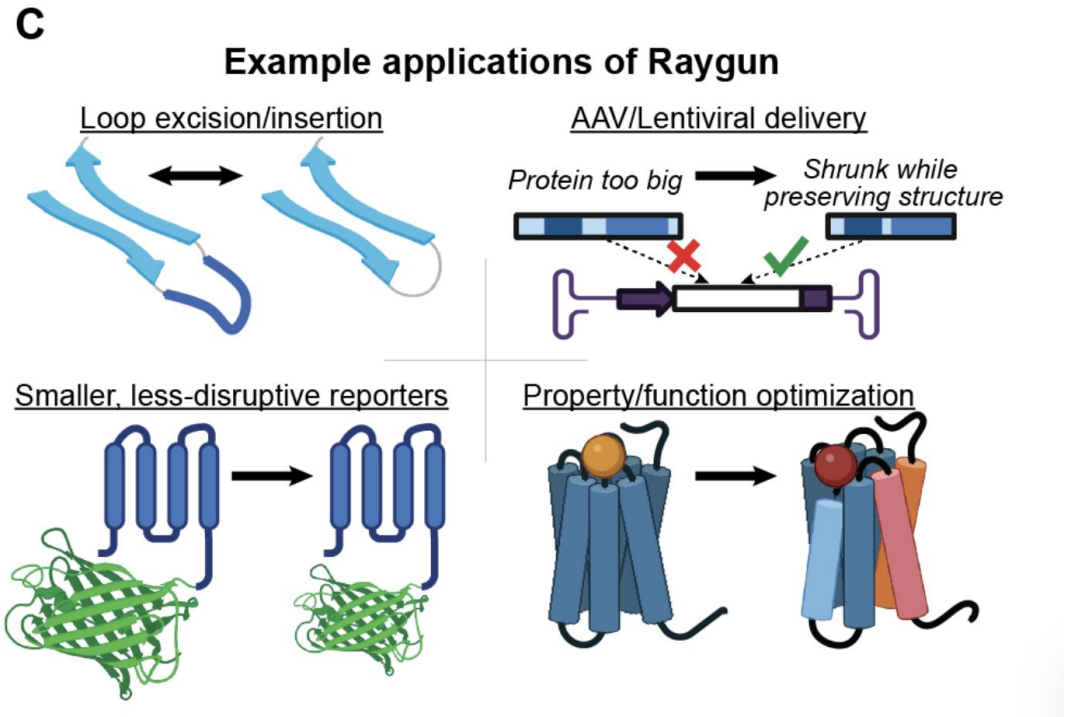
One-Step Recycling: Improves quality and diversity by reusing a newly generated candidate as the next input.

Candidates are filtered by:

- **Pseudo log-likelihood (pLL)** from ESM to check "evolutionary likelihood."
- **Optional property filter:** Functional site retention or specialized predictive models for brightness, binding, etc.
- **Structural filters** (e.g., AlphaFold pLDDT, TM-score).



Raygun applications



Protein editing using Raygun for proteins of different sequence lengths

Examples: Hemoglobin (147 aa), CCR1 (355 aa), lacZ (1,029 aa), mTOR (2,549 aa).

- 2000 samples for each protein across lengths
- Set noise parameter to 0.5 to introduce moderate sequence variations
- Filtered the samples using length-adjusted ESM-2 pseudo-loglikelihood (pLL) scores.
- Kept top 5% of generated sequences distributed across the length spectrum

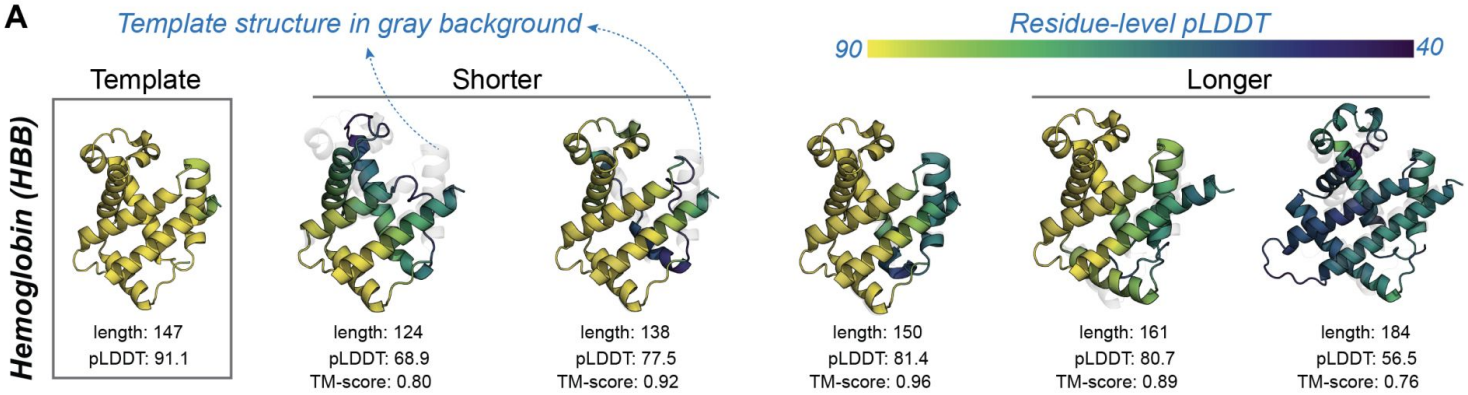
Key takeaway: Raygun effectively preserves protein structure across a wide range of output lengths, as evidenced by the high TM-scores

the degree of structure preservation varies depending on the properties of the template protein.

CCR1's 206 TM-score decreased to 0.68 when shortened by 17%, while mTOR could accommodate a 25% reduction to reach a similar TM-score of 0.69. As expected, greater deviations from the original length generally correspond to lower TM-score and pLDDT values

Protein editing using Raygun for proteins of different sequence lengths

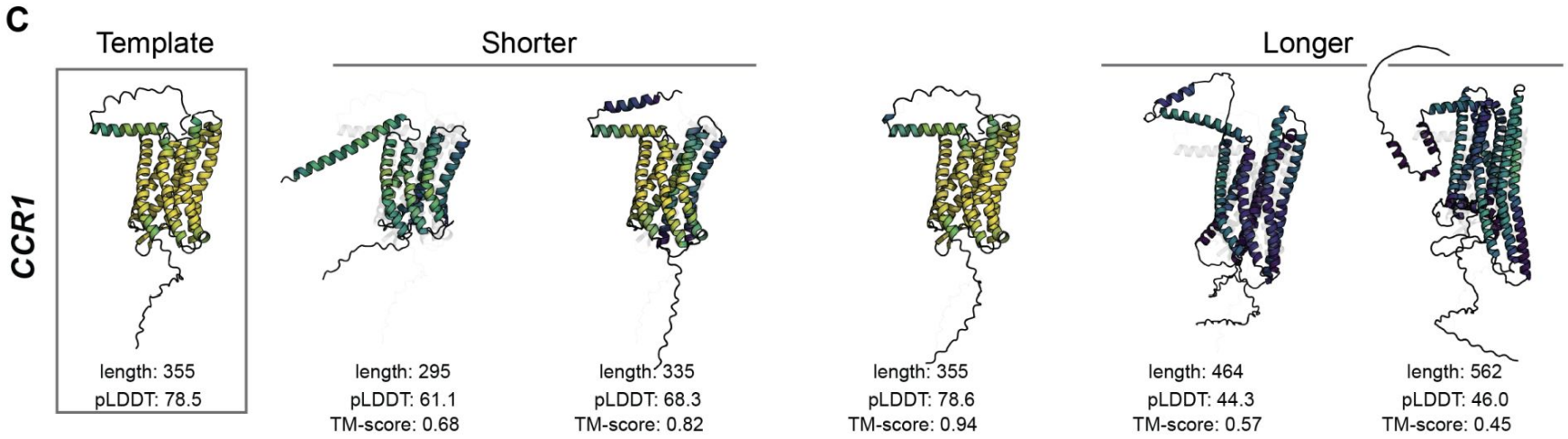
Examples: Hemoglobin (147 aa), CCR1 (355 aa), lacZ (1,029 aa), mTOR (2,549 aa).



Raygun effectively preserves protein structure across a wide range of output lengths as evidenced by the high TM-scores

Protein editing using Raygun for proteins of different sequence lengths

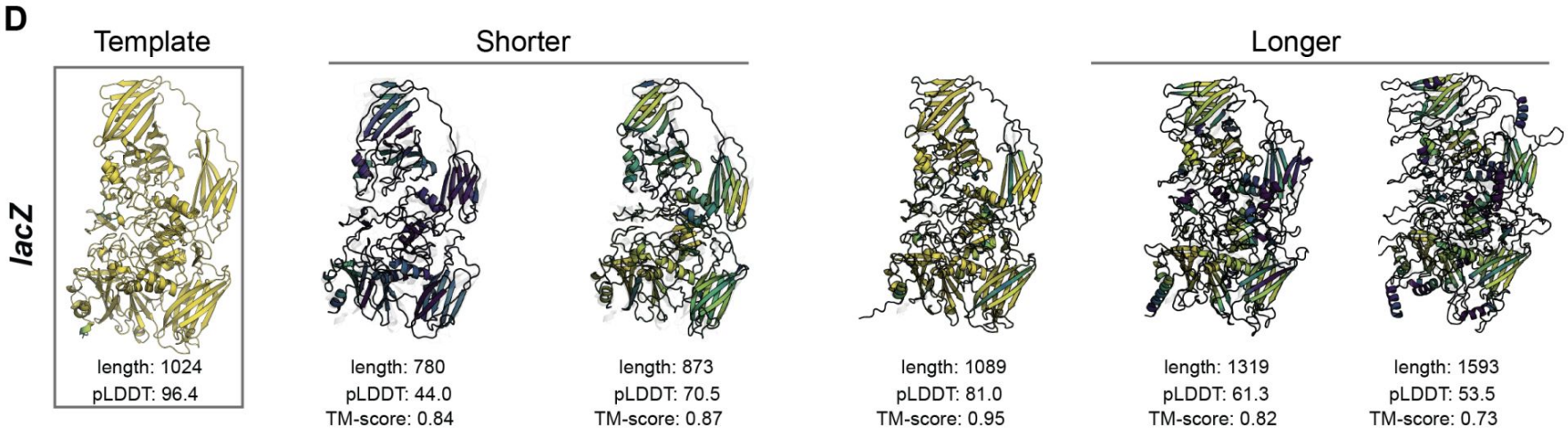
Examples: Hemoglobin (147 aa), **CCR1 (355 aa),** lacZ (1,029 aa), mTOR (2,549 aa).



Raygun effectively preserves protein structure across a wide range of output lengths as evidenced by the high TM-scores

Protein editing using Raygun for proteins of different sequence lengths

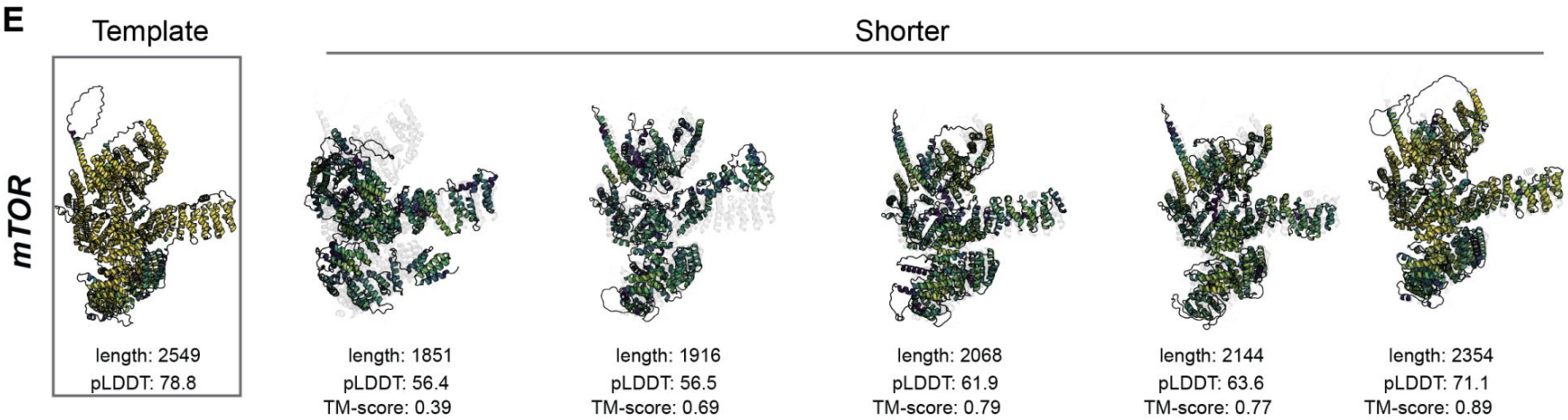
Examples: Hemoglobin (147 aa), CCR1 (355 aa), **lacZ (1,029 aa),** mTOR (2,549 aa).



Raygun effectively preserves protein structure across a wide range of output lengths as evidenced by the high TM-scores

Protein editing using Raygun for proteins of different sequence lengths

Examples: Hemoglobin (147 aa), CCR1 (355 aa), lacZ (1,029 aa), **mTOR (2,549 aa).**



Raygun effectively preserves protein structure across a wide range of output lengths as evidenced by the high TM-scores

Framework for Evaluating Template-Guided Design Methods

Structural Versatility

- Ability to generate diverse secondary structures without bias (e.g., avoiding excessive α-helical outputs).

Functional Site Preservation

- Retains and protects critical regions (active/binding sites) in the template.

Tunable Modification Range & Sequence Diversity

- Capable of both small substitutions and large indels, spanning minor to major transformations.

Scalability Across Protein Sizes

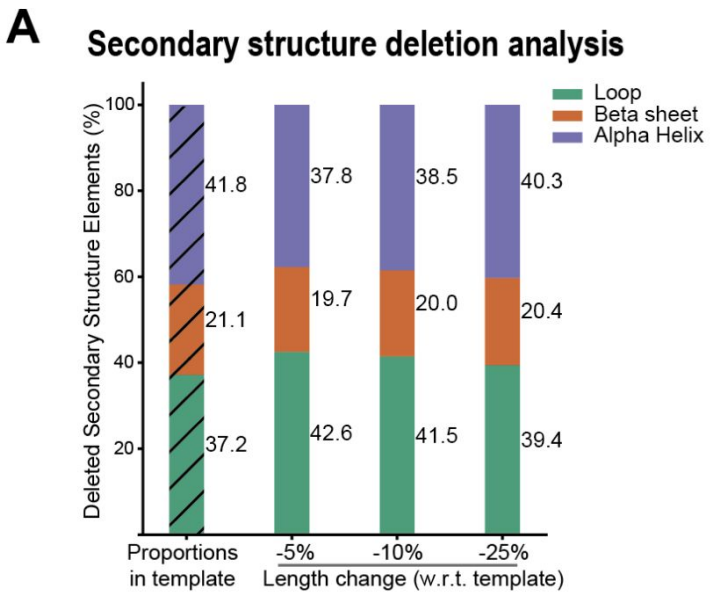
- Consistent performance from small to very large proteins, enabling wide applicability.

Information-Rich Representation

- A robust fixed-length embedding that captures enough structural information to guide accurate sampling.
- Demonstrated via clustering analysis (vs. baseline ESM-2)

Evaluating Raygun's structural versatility

- Shortened variants of 10,000 template proteins belonging to SCOP families representing α, β, α + β, and α/β structural classes
- Analysis of thousands of generated proteins showed:
  - Slight preference for removing loops or disordered regions, but balanced overall.
  - As the degree of miniaturization increases, more balanced in deletions



Evaluating Raygun's functional site preservation

Setup

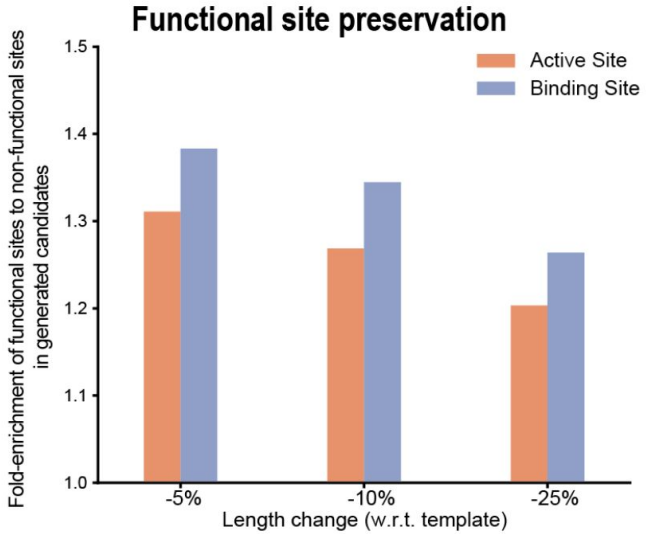
- 10,000 Proteins from UniProt (all with annotated active & binding sites).
- Generated Raygun variants at 5%, 10%, and 25% length reductions.

Evaluation Metric

- Compared the fraction of preserved functional sites to overall sequence identity.
- Ratio > 1 ⇒ Functional sites are conserved above chance.

Key Observations

- Binding sites are slightly better preserved than active sites.
- More Extensive Deletions ⇒ Lower preservation, reflecting the challenge of maintaining functional integrity with larger indels.



Controlling Raygun Outputs

Two Key Parameters

- **Noise:** Scales the covariance in the MVN, controlling substitution rates.
- **Output Length:** Dictates the extent of insertions/deletions (indels).

Noise Experiments (20 Proteins)

- **Noise Range:** 0.01 → 6, 100 candidates each, top 5 retained.
- **Interpretation:**
  - Low Noise (<0.5) ⇒ Minor, conservative edits.
  - High Noise (2+) ⇒ Greater sequence diversity but reduced structural similarity.

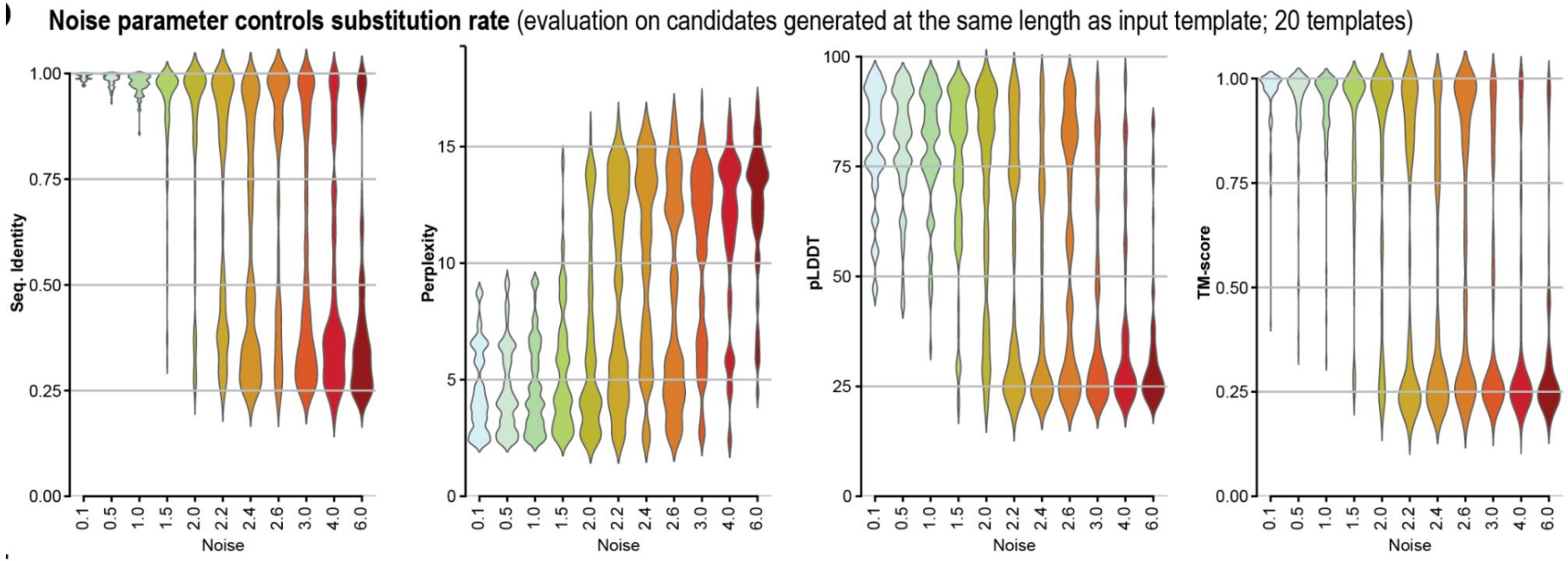
Length Variation

- Tested at noise levels 0.1, 0.5, 1.0.

The control allows for a spectrum of designs, from subtle tweaks to major structural modifications, with predictable trade-offs between sequence diversity and structural preservation.

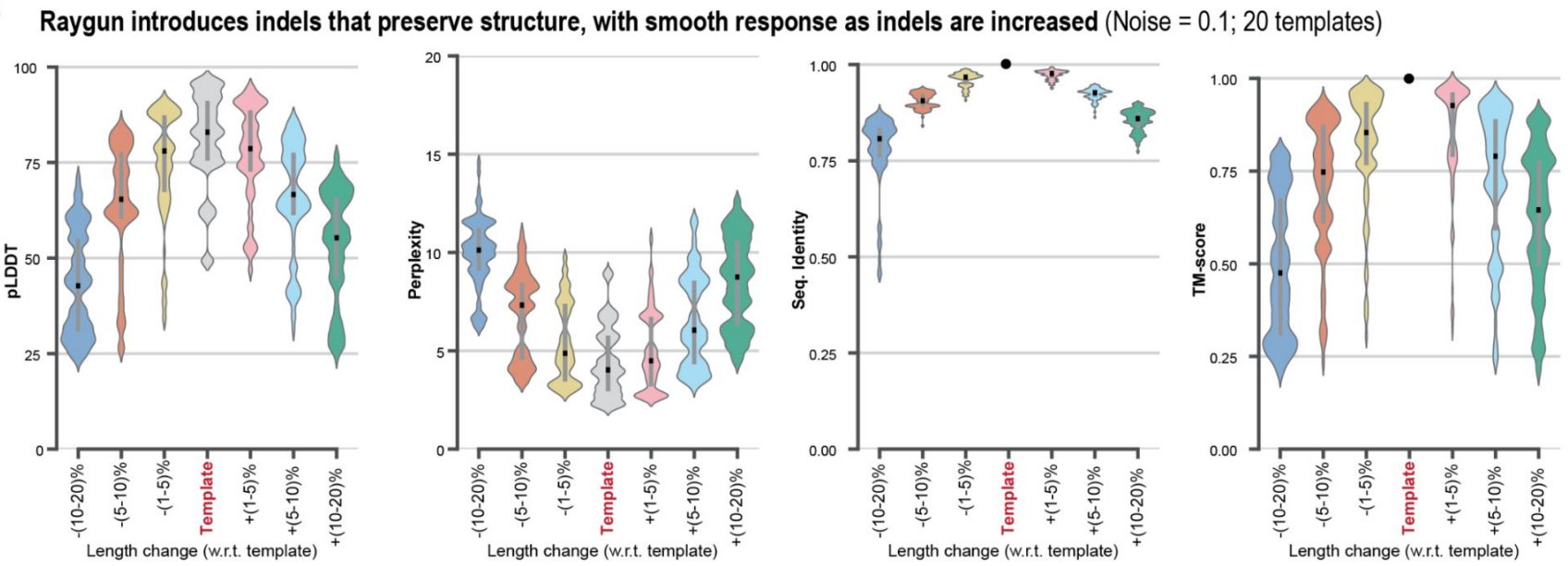
Raygun parameters provide fine-grained control over protein generation

Sequence identity decreases gradually up to ≈2.2, then drops sharply.
TM-score & pLDDT follow a similar trend.



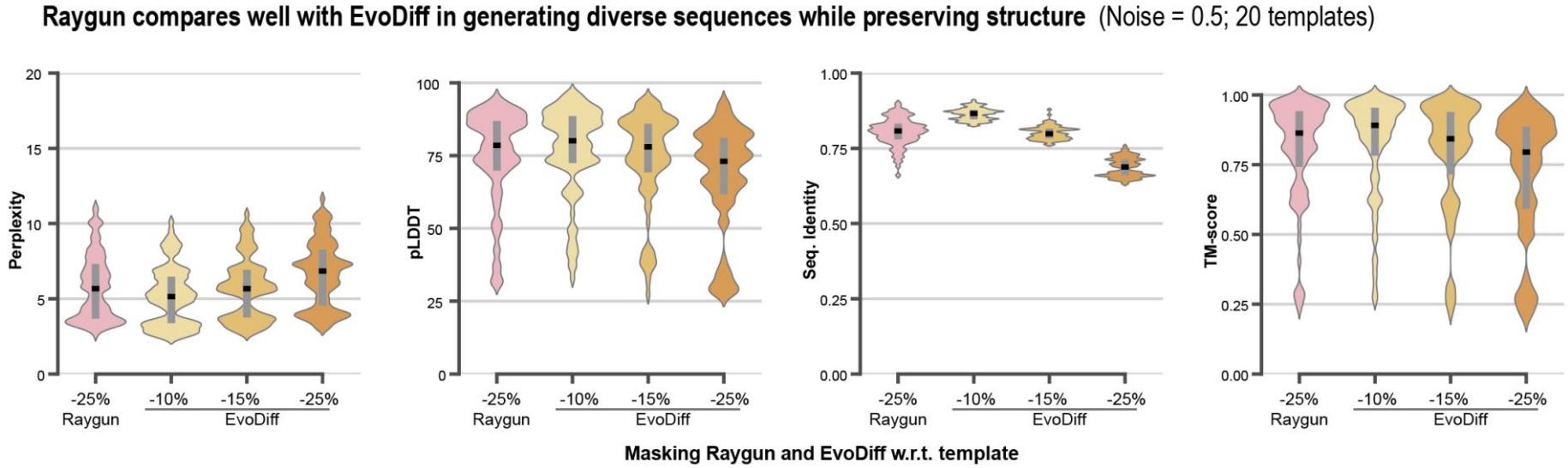
Raygun parameters provide fine-grained control over protein generation

Changes within ±10% often preserve structural integrity well.
Larger deviations (e.g., ±20% of template length) lower sequence identity & TM-score.



Raygun compares favorably with existing approaches in balancing sequence diversity and structural plausibility

Same-length comparisons with EvoDiff, where a part of the sequence has been masked and the methods are tasked with regenerating them.
At the same masking rate as EvoDiff, Raygun's candidates have better predicted structural properties.



Raygun's fixed-length representations better capture the organization of protein structural shapes

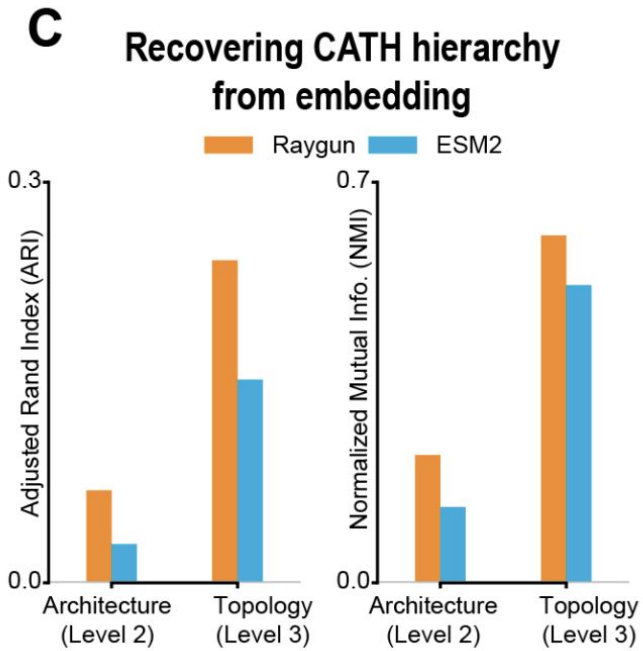
Evaluate whether Raygun's representations better capture structural organization than ESM-2's

Approach

- Used the CATH database, which classifies proteins by architecture (global folds) and topology (fine-grained structure).
- Applied agglomerative clustering to Raygun and ESM-2 representations.

Raygun outperformed ESM-2 both at Architecture Level (global structure) and Topology Level (fine-grained patterns)

Improvement most notable at higher-level (architecture) structure grouping.



Raygun's fixed-size embeddings better preserve structural hierarchy, enabling smoother transitions between related structures

Raygun demonstrates structural versatility while preserving functional domains

Evaluate Raygun's ability to maintain functional integrity across structural diversity and variable lengths by testing PFAM domain retention.

Selected PFAM domains spanning major SCOP structural classes.

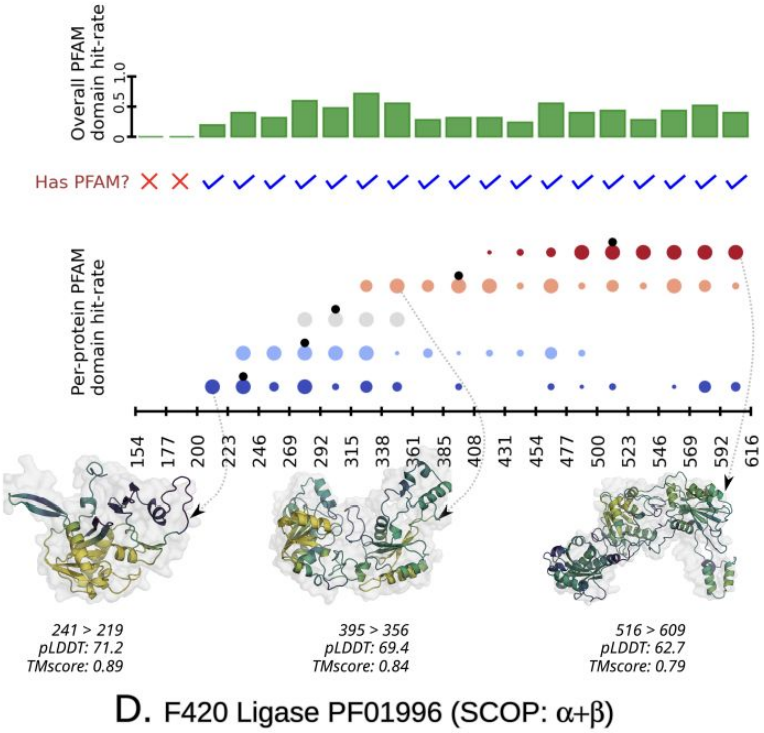
For each domain, 5 diverse template proteins, generated variants across 50–200% of median template length.

Assessed PFAM domain retention using HMMER.

Average domain retention: 48.25%

Highest: 57% (α/β class)

Lowest: 37% (α+β class)



Miniaturizing fluorescent proteins (FPs)

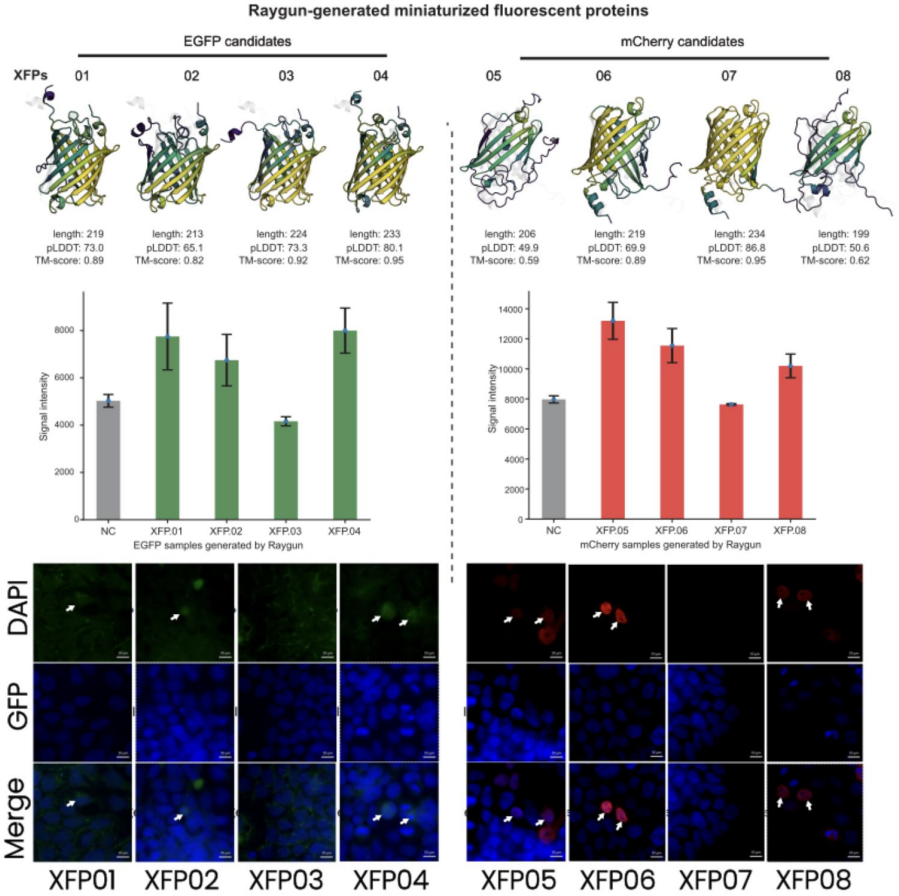
Motivation: FPs like eGFP and mCherry (~236–238 aa) can be too large for fusions with small proteins.

Raygun experiment:

- Generated candidates with ~5–15% length reductions (down to ~200 aa).
- Used sequence-based and brightness predictive filters.
- Chose eight final variants (XFP01–08) for experimental testing in HEK293 cells.

Results:

- Six of eight exhibited measurable fluorescence, including two at lengths of only 199 and 206 aa.
- One candidate changed the characteristic chromophore motif yet still fluoresced.
- These miniaturized designs are shorter than 96% of known FPs.



Miniaturizing TurboID

Background: TurboID (~320 aa) is a synthetic biotin ligase widely used for proximate proteomics.

Goal: Explore 1–20% length reductions, potentially up to 50%.

Method:

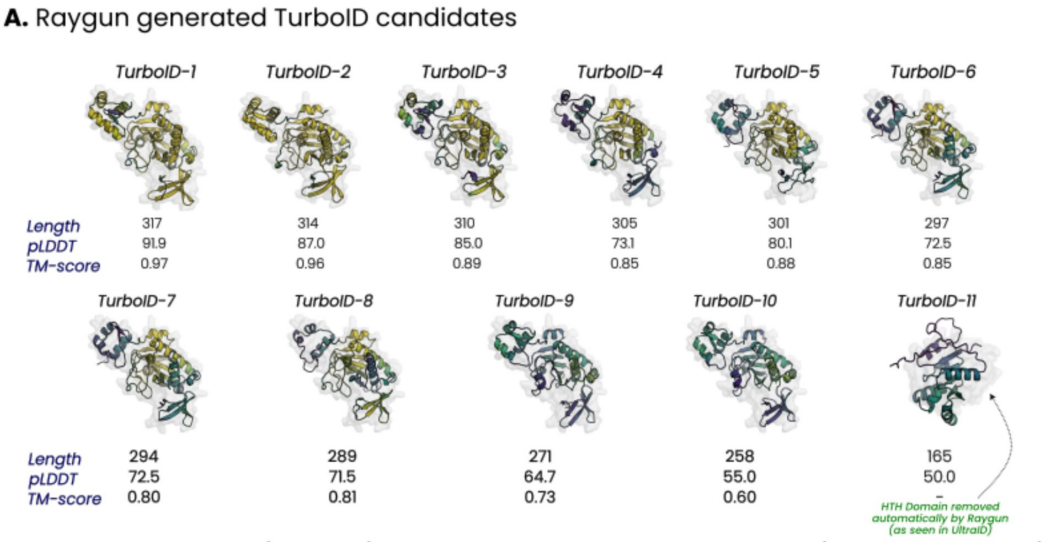
- Generated ~500k variants, filtered by pseudo log-likelihood, domain recognition, and thermostability metrics.
- Selected 11 variants for cell-based tests.

Experimental screening:

- 6 expressed in cells; among those, 2 showed moderate enzymatic activity.
- Even 50% length reduction was partially successful in expression but lost activity.

Implication: Confirmed that Raygun could remove entire domains (e.g., DNA-binding domain in BirA-based TurboID) in an unsupervised manner.

Miniaturizing TurboID



Background: TurboID (~320 aa) is a syn
proteomics.

Goal: Explore 1–20% length reductions

Method:

- Generated ~500k variants, filtered by pseudo log-likelihood, domain recognition, and thermostability metrics.
- Selected 11 variants for cell-based tests.

Experimental screening:

- 6 expressed in cells; among those, 2 showed moderate enzymatic activity.
- Even 50% length reduction was partially successful in expression but lost activity.

Implication: Confirmed that Raygun could remove entire domains (e.g., DNA-binding domain in BirA-based TurboID) in an unsupervised manner.

Magnifying EGF for higher binding affinity

Motivation: EGF–EGFR binding is central to many signaling pathways; EGF is only 53 aa.

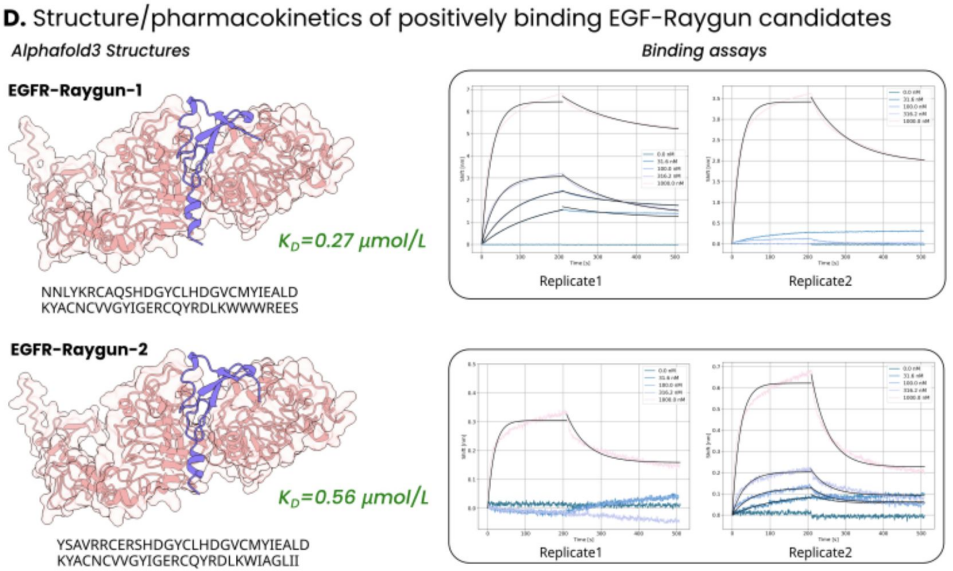
Approach:

- Used Raygun to expand EGF from 53 to 55–57 aa.
- Screened 10k variants with a tri-modal PLM (ProTrek) for EGFR-binding potential.
- Submitted top candidates to an external competition (AdaptyvBio).

Results:

- 4 variants were chosen for experimental testing.
- All four expressed; two had stronger EGFR binding than wildtype EGF (KD of 0.274 µM vs. 0.759 µM).

Demonstrated Raygun's ability to enhance binding even without a specialized binder module.



Discussion

Key conceptual advancements

- Probability distribution representation (multivariate normal) lets Raygun sample new protein lengths in a single shot, removing the typical complexity of chain growth or iterative “denoising” steps.
- Length-agnostic and domain-agnostic approach: proven on complex tasks like shrinking multi-domain enzymes or expanding short ligands for higher affinity.

Comparison with de novo methods

- Raygun outperforms or matches many “full generation” approaches in structural plausibility (pLDDT, TM-score) and can incorporate large indels without losing the original fold.
- Excellent for remodeling proteins that must remain largely consistent with a well-validated template.

Discussion

Challenges and limitations

- Extreme miniaturization can break function (e.g., 50% reduction in TurboID). These variants may require additional rounds of directed evolution or targeted site constraints.
- For specialized tasks (e.g., engineering enzymes for new catalytic specificity), one might need more advanced function-specific filters or further fine-tuning.

Implications for future protein engineering

- Potential to systematically explore large combinatorial spaces around known proteins (biosensors, enzymes, binders, structural components).
- Could accelerate therapeutic design by optimizing existing protein therapeutics for better expression, lower immunogenicity, or improved tissue penetration.

Conclusion

Raygun

- Introduces a robust, probabilistic fixed-length encoding for protein sequences.
- Achieves shrinkage, remodeling, and expansion of template proteins while preserving structure and, often, functionality.
- Demonstrates success in multiple domains: fluorescent proteins, biotin ligase, and EGF–EGFR binding.

Future directions

- Integrating Raygun with directed evolution pipelines to further refine function.
- Extending to more specialized tasks: e.g., domain swapping, designing multi-specific binders, or customizing immune epitopes.
- Community benchmarks for large-scale, template-guided protein engineering.

Take-home message

- Raygun exemplifies how protein language models, combined with clever encoding-decoding paradigms, can unlock entirely new engineering avenues.