RNAnneal-ss: Incorporating LinearFold-V and EternaFold Suboptimals

Motivation (FR3D/BGSU under400, non-rRNA; N=804)

On this benchmark, both **LinearFold-V** and **EternaFold** achieve higher **F1@100** than the current RNAnneal-ss pipeline. This suggests these methods generate *better candidate ensembles and/or better global scaffolds*, especially for longer RNAs, and motivates using their **suboptimal** structures inside RNAnneal-ss to improve **F1@100** and reduce **Fail@100**.

Key empirical signals

- Mean F1@100: RNAnneal-ss ≈ 0.784, LinearFold-V ≈ 0.794, EternaFold ≈ 0.820.
- "Oracle" per-target max across the three methods reaches mean F1@100 ≈ 0.846, indicating substantial headroom from ensembling.
- Fail@100 (best-of-100 F1 = 0): RNAnneal-ss $\approx 1.2\%$, LinearFold-V $\approx 0.6\%$, EternaFold $\approx 0.1\%$; all-three fail $\approx 0.1\%$.
- **Where the good structure appears:** LinearFold-V's best-of-100 is usually early (median rank 4), while EternaFold's best-of-100 is often late (median rank 29; $\sim 47\%$ of targets have best rank > 30). This argues for **including many EternaFold samples** or selecting an evidence-ranked subset rather than only top-1.

Strategy

1) External candidates (guaranteed coverage@100)

Treat LinearFold-V and EternaFold structures as **first-class candidates** in the final top-100 list (not only as internal scaffolds). To target F1@100, ordering is secondary; what matters is ensuring the pool contains diverse, plausible folds.

Recommended allocation (length-adaptive):

- $L \leq 150$: reserve 15–25 slots each for (LinearFold-V, EternaFold), keep remaining for RNAnneal-ss candidates.
- $151 \leq L \leq 300$: reserve 25–35 LinearFold-V and 35–50 EternaFold (late best ranks), reduce RNAnneal-ss share accordingly.
- $301 \leq L \leq 400$: reserve 35–45 each for (LinearFold-V, EternaFold), since both scale better and RNAnneal-ss currently lags in these buckets.

Within each reserved block, select a **diverse subset** (e.g., max-min Jaccard distance over base-pair sets) and keep at least one **top-1** structure from each method.

2) External scaffolds for MCMC sampling

Use a subset of external structures as additional **scaffolds** for RNAnneal-ss sampling:

- Build a **scaffold pool** from: CaCoFold SS_cons (and refined variants), MFE, LinearFold-V suboptimals, EternaFold MEA + samples.
- Run sampling in two regimes:
  - **Fixed-scaffold pass** (current default): fill in remaining positions around each scaffold.
  - **Unfixed-scaffold pass** (recommended for $L \geq 151$ and for external scaffolds): allow edits to scaffold pairs so a "good-but-wrong" scaffold can be repaired.
- Merge sampled candidates back into the global candidate pool and deduplicate by dot-bracket string.

3) Optional: external-pair evidence component

Convert the external ensembles into an **evidence map**:

- Compute per-pair frequencies across (LinearFold-V top-K, EternaFold samples).
- Add a new weight component (e.g., $\log(p + \epsilon)$ or p) blended into the RNAnneal-ss scoring weights with a length-adaptive coefficient.
- This can reduce Fail@100 by ensuring globally plausible long-range pairs become *available* to the sampler even when CaCoFold/thermo evidence is weak.

Implementation outline (code design)

- Add an "external candidate" step in the RNAnneal-ss pipeline driver:
  - LinearFold-V (`--zuker`) → up to 100 unique structures.
  - EternaFold MEA + `contrafold sample` → up to 100 unique structures (note: stochastic; no seed flag).
- Separate **sampling scaffolds** from **candidate list** to avoid treating injected external scaffolds as "refined" variants (and to keep existing refined-prefix heuristics meaningful).
- Add per-source quotas in the final top-100 selector (especially for $L \geq 151$) so external candidates are not pruned by proxy scoring.
- Add caching keyed by (*method*, *sequence*) to avoid re-running external tools across experiments.

Benchmark plan (focus: F1@100 and Fail@100)

- Primary ablations: (A) candidates-only quotas, (B) scaffolds for sampling, (C) external-pair evidence weights, (A+B), (A+B+C).
- Report overall and by-length buckets (30–80, 81–150, 151–300, 301–400).
- Track runtime overhead per target (especially for $L \geq 151$).