RNAnneal-ss Ensembles on representative50 (under300) — v1

Objective

Improve **F1@100** on the 50-target representative50 subset (truth length ≤ 300) by leveraging the empirical observation from `docs/u300_50_scaffold_validation_v3` that **RNss(X)** typically improves the corresponding baseline **X**, especially at @100.

Methods

- **RNss(EF/LF/RS):** the three scaffold-seeded RNAnneal-ss variants from `u300_50_scaffold_validation_v3`.
- **Ens(RNss):** a post-hoc ensemble that pools the *top-100 candidate sets* from RNss(EF), RNss(LF), and RNss(RS), then selects 100 structures using:
  - length-adaptive per-source quotas that preserve most RNss(EF) candidates (70–90 of 100, increasing with length),
  - within-source diversity (Jaccard) to keep suboptimal coverage,
  - a global diversity fill step to resolve cross-source deduplication.
- **RNss(Ens3):** a single RNAnneal-ss run that replaces single-strand `Fold`/`AllSub` with an *aggregated* backend (RNAstructure + LinearFold-V + EternaFold). Included as an ablation (it did not outperform Ens(RNss) at @100).
- **Baselines:** RNAstructure, LinearFold-V, and EternaFold from `u300_50_scaffold_validation_v3`.

Results (@1 and @100)

Table 1: Overall performance on representative50 (N=50; metrics @1 and best-of-100).

Method	Mean F1@1	Med F1@1	Mean F1@100	Med F1@100	ΔF1	Fail@100
Ens(RNss)	0.542	0.530	0.723	0.738	0.181	0.0%
RNss(EF)	0.542	0.530	0.721	0.729	0.179	0.0%
RNss(LF)	0.490	0.413	0.695	0.710	0.205	0.0%
RNss(RS)	0.515	0.488	0.651	0.672	0.135	0.0%
RNss(Ens3)	0.510	0.491	0.717	0.719	0.207	0.0%
LF-V	0.519	0.498	0.684	0.701	0.165	2.0%
EFold	0.549	0.546	0.719	0.734	0.170	0.0%
RNAstr	0.516	0.460	0.563	0.506	0.047	4.0%

Table 2: Performance by truth-length bucket (50-nt buckets from the manifest).

Bucket	Method	N	Mean F1@1	Mean F1@100	Fail@100
30-79	Ens(RNss)	12	0.620	0.838	0.0%
30-79	RNss(EF)	12	0.620	0.834	0.0%
30-79	RNss(LF)	12	0.629	0.797	0.0%
30-79	RNss(RS)	12	0.634	0.809	0.0%
30-79	RNss(Ens3)	12	0.628	0.834	0.0%
30-79	LF-V	12	0.633	0.767	8.3%
30-79	EFold	12	0.621	0.823	0.0%
30-79	RNAstr	12	0.641	0.801	8.3%
80-129	Ens(RNss)	10	0.635	0.888	0.0%
80-129	RNss(EF)	10	0.635	0.874	0.0%
80-129	RNss(LF)	10	0.639	0.854	0.0%
80-129	RNss(RS)	10	0.653	0.864	0.0%
80-129	RNss(Ens3)	10	0.648	0.884	0.0%
80-129	LF-V	10	0.738	0.845	0.0%
80-129	EFold	10	0.668	0.872	0.0%
80-129	RNAstr	10	0.700	0.744	0.0%
130-179	Ens(RNss)	10	0.532	0.680	0.0%
130-179	RNss(EF)	10	0.532	0.680	0.0%
130-179	RNss(LF)	10	0.432	0.658	0.0%
130-179	RNss(RS)	10	0.511	0.616	0.0%
130-179	RNss(Ens3)	10	0.508	0.675	0.0%
130-179	LF-V	10	0.470	0.658	0.0%
130-179	EFold	10	0.529	0.677	0.0%
130-179	RNAstr	10	0.457	0.457	10.0%
180-229	Ens(RNss)	9	0.396	0.542	0.0%
180-229	RNss(EF)	9	0.396	0.548	0.0%
180-229	RNss(LF)	9	0.329	0.537	0.0%
180-229	RNss(RS)	9	0.343	0.428	0.0%
180-229	RNss(Ens3)	9	0.331	0.535	0.0%
180-229	LF-V	9	0.313	0.528	0.0%
180-229	EFold	9	0.396	0.548	0.0%
180-229	RNAstr	9	0.333	0.333	0.0%
230-279	Ens(RNss)	7	0.492	0.597	0.0%
230-279	RNss(EF)	7	0.492	0.602	0.0%
230-279	RNss(LF)	7	0.328	0.562	0.0%
230-279	RNss(RS)	7	0.370	0.462	0.0%
230-279	RNss(Ens3)	7	0.371	0.584	0.0%
230-279	LF-V	7	0.337	0.562	0.0%
230-279	EFold	7	0.497	0.611	0.0%
230-279	RNAstr	7	0.383	0.383	0.0%
280-300	Ens(RNss)	2	0.499	0.683	0.0%
280-300	RNss(EF)	2	0.499	0.683	0.0%
280-300	RNss(LF)	2	0.494	0.643	0.0%
280-300	RNss(RS)	2	0.420	0.470	0.0%
280-300	RNss(Ens3)	2	0.421	0.666	0.0%
280-300	LF-V	2	0.550	0.643	0.0%
280-300	EFold	2	0.499	0.683	0.0%
280-300	RNAstr	2	0.420	0.420	0.0%

1 - CDFs

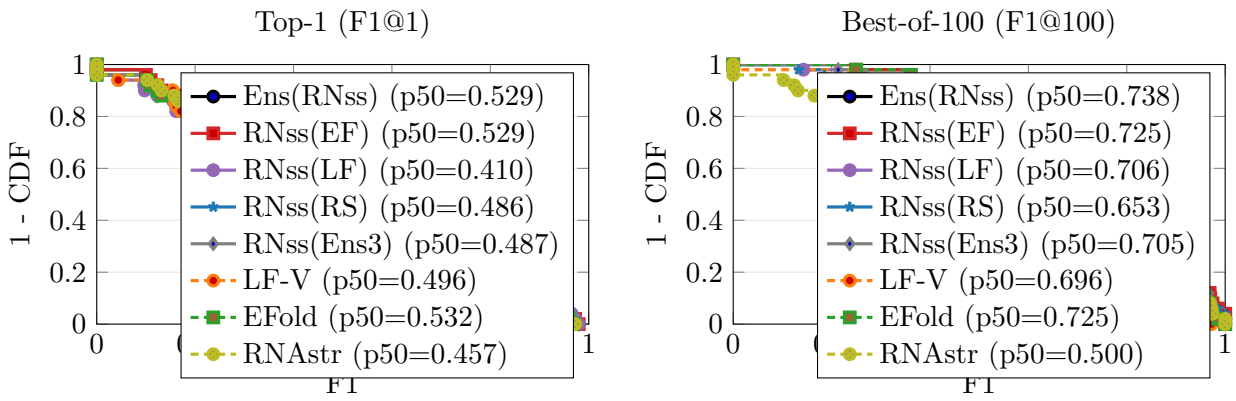

Figure 1: * F1@1

Figure 2: * F1@100

Figure 3: Empirical survival curves (1 - CDF). Legend includes the median (p50) F1.

Precision/Recall Curves

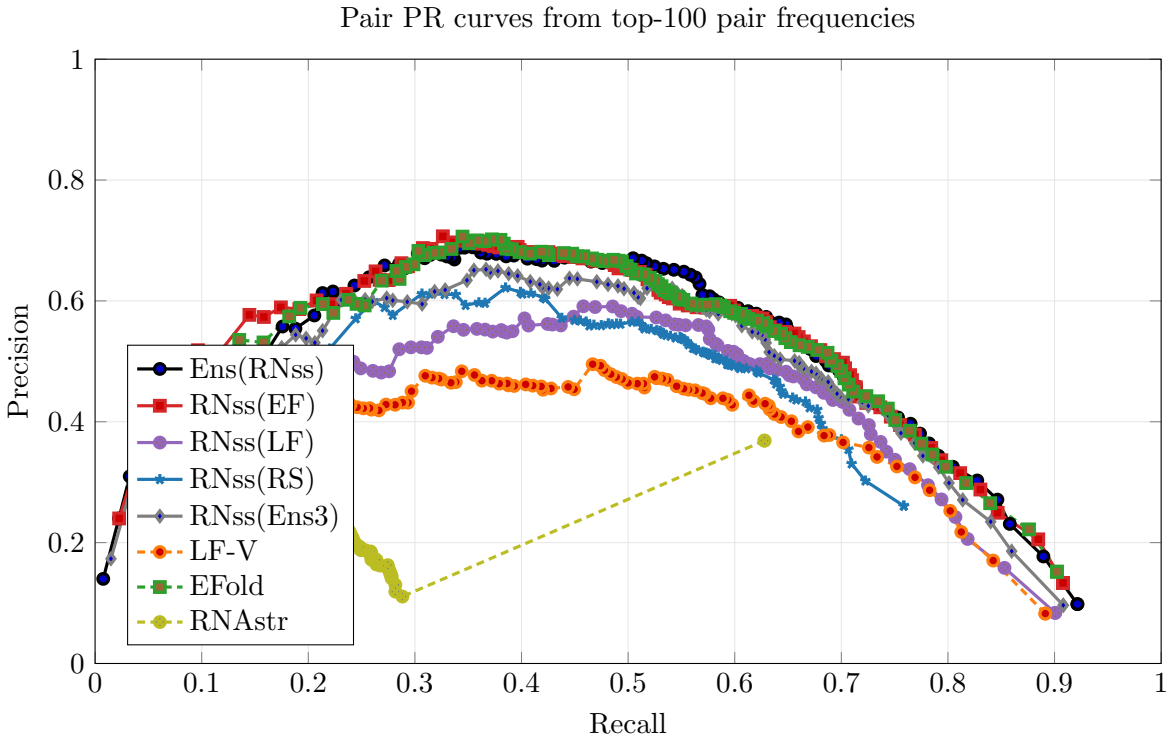

Figure 4: Pair-level PR curves computed by thresholding pair probabilities estimated as *frequency across the top-100 predicted structures.*

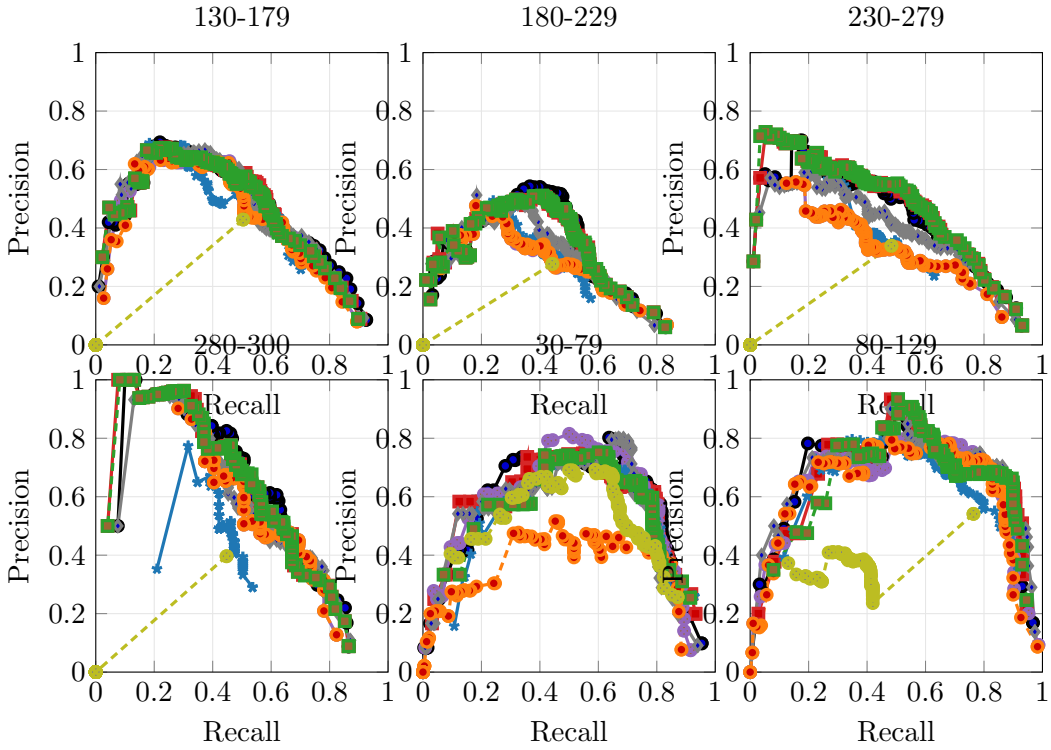

Figure 5: PR curves by 50-nt length bucket.

Score Distributions

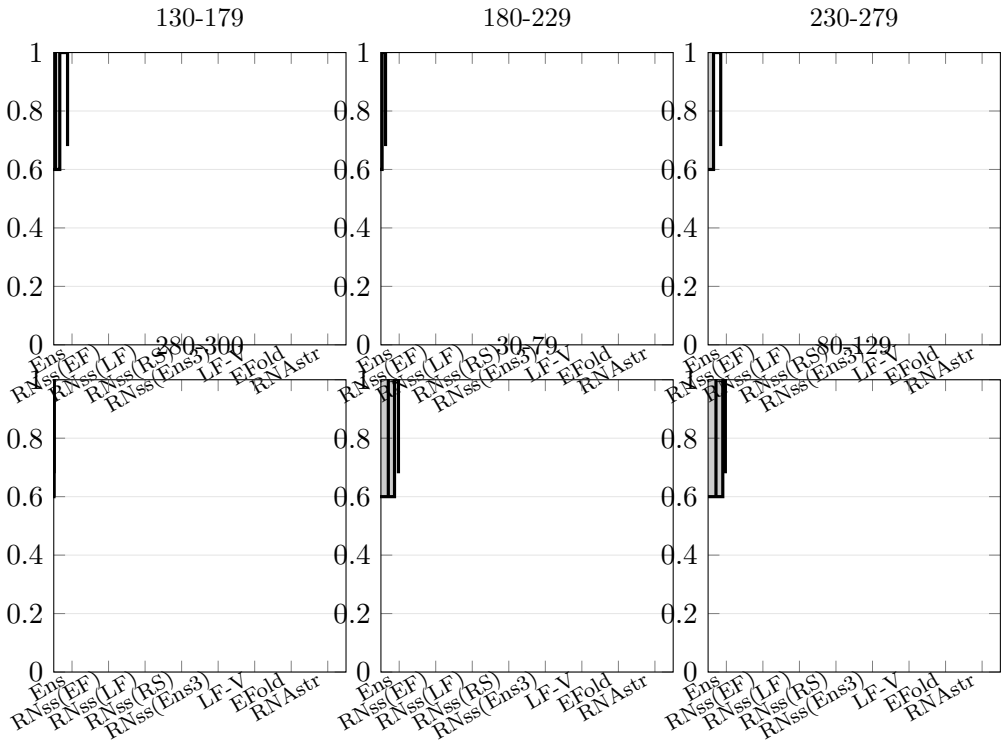

Figure 6: F1@1 distributions by length bucket (boxplots; outliers hidden).

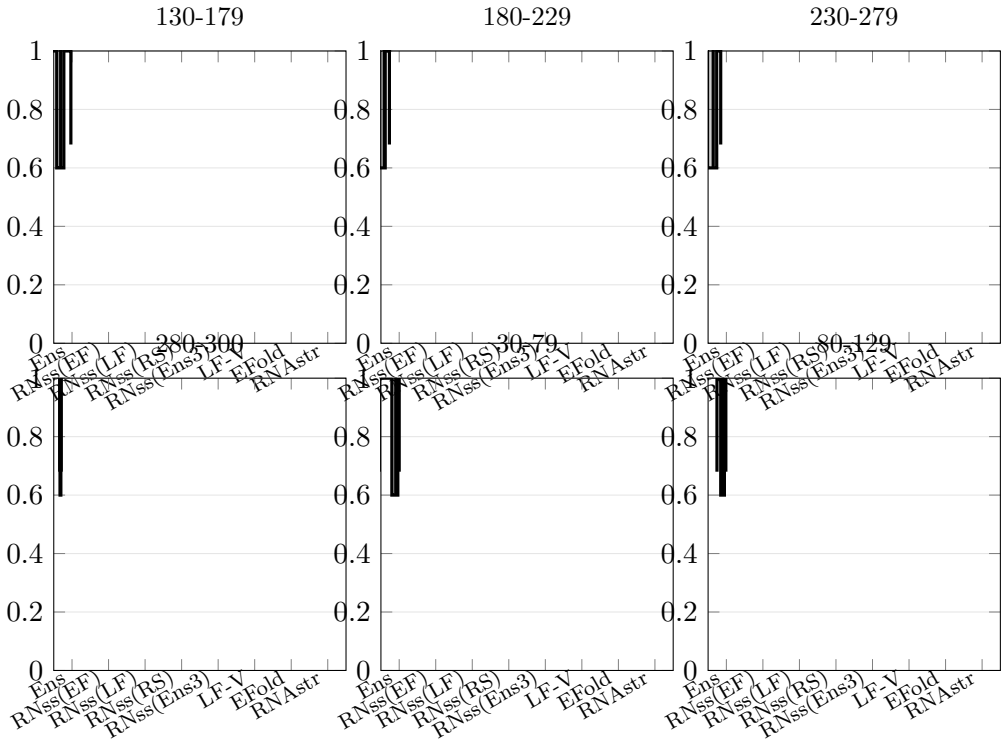

Figure 7: F1@100 distributions by length bucket (boxplots; outliers hidden).

Takeaways

On this subset, **Ens(RNss)** achieves the best mean F1@100 while preserving **Fail@100 = 0%**. The improvement over RNss(EF) is small but consistent with the premise that a small amount of additional diversity from other **RNss(X)** variants can recover complementary wins without sacrificing the high-yield EF candidate set.