RNAnneal-ss hyperparameter ablation (under300 representative27) — v1

Scope

This report studies sensitivity of **RNAnneal-ss** refinement/sampling hyperparameters on a **27-target** subset of FR3D/BGSU under400 with truth length ≤ 300, balanced across **short/medium/long** length groups (9 each).

Experimental setup

- **Common pipeline:** CaCoFold constraints + refinement + MCMC sampling (top-100 output).
- **Scaffold backend:** single-strand scaffolds via EternaFold (Fold + AllSub-like suboptimals). Duplex sampling uses RNAstructure `DuplexFold`.
- **Key fix included in all configs:** the refinement end-mask grid is enabled by using an end-mask step smaller than the max end-mask length (`step=10`, `max=30`).
- **Length-adaptive scaling:** several budgets (e.g., max scaffolds, refinement seconds) scale with length; the tables include per-target wall-clock runtime from `qc.json`.

Configs (one-factor-at-a-time around B4)

- **B4:** baseline (step=10, max-end-mask=30, min-unpaired=6, max-regions=4, max-seeds/solutions=50, kissing=50, max-scaffolds=15).
- **U4:** `--refine-min-unpaired 4`.
- **R10:** `--refine-max-regions 10`.
- **S150:** `--refine-max-seeds 150 --refine-max-solutions 150`.
- **K300:** `--refine-kissing-candidates 300`.
- **Sc30:** `--max-scaffolds 30`.
- **Hi:** combined (U4 + R10 + S150 + K300 + Sc30).
- **v3:** prior reference run (RNAnneal-ss with EternaFold scaffolds) for context.

Findings (this subset)

- Best mean F1@100: **R10** (0.6987); B4=0.6983; Δ=+0.0005.
- Mean F1@100 spread across configs is 0.0008 on this 27-target subset (Table 3 confirms the knobs did change internal budgets).
- Fail@100 is 0.0% for all configs on this subset.
- Runtime: R10 is -0.0s vs B4 (mean per target).

Results (@1 and @100)

Table 1: Overall ablation results (N=27; @1 and best-of-100). Runtime is wall-clock seconds per target.

Config	Mean F1@1	Mean F1@100	Fail@100	Mean time (s)	Med time (s)
v3	0.520	0.698	0.0%	30.8	18.9
B4	0.505	0.698	0.0%	32.1	18.7
U4	0.505	0.698	0.0%	32.1	19.0
R10	0.503	0.699	0.0%	32.1	18.1
S150	0.505	0.698	0.0%	32.1	19.0
K300	0.505	0.698	0.0%	32.0	18.5
Sc30	0.507	0.698	0.0%	33.9	18.4
Hi	0.504	0.699	0.0%	34.0	18.1

QC (did knobs change behavior?)

Table 2: Results by length group from the manifest (short/medium/long).

Group	Config	N	Mean F1@1	Mean F1@100	Fail@100	Mean time (s)
short	v3	9	0.537	0.798	0.0%	11.6
short	B4	9	0.532	0.798	0.0%	12.4
short	U4	9	0.532	0.798	0.0%	12.1
short	R10	9	0.523	0.798	0.0%	12.7
short	S150	9	0.532	0.798	0.0%	12.0
short	K300	9	0.532	0.798	0.0%	12.0
short	Sc30	9	0.537	0.798	0.0%	12.3
short	Hi	9	0.529	0.798	0.0%	12.3
medium	v3	9	0.528	0.675	0.0%	19.2
medium	B4	9	0.490	0.676	0.0%	19.7
medium	U4	9	0.490	0.676	0.0%	19.6
medium	R10	9	0.487	0.678	0.0%	19.2
medium	S150	9	0.490	0.676	0.0%	19.8
medium	K300	9	0.490	0.675	0.0%	19.9
medium	Sc30	9	0.490	0.676	0.0%	20.2
medium	Hi	9	0.484	0.678	0.0%	19.9
long	v3	9	0.494	0.620	0.0%	61.5
long	B4	9	0.493	0.620	0.0%	64.3
long	U4	9	0.493	0.620	0.0%	64.6
long	R10	9	0.497	0.620	0.0%	64.4
long	S150	9	0.494	0.620	0.0%	64.7
long	K300	9	0.493	0.620	0.0%	64.2
long	Sc30	9	0.493	0.620	0.0%	69.3
long	Hi	9	0.498	0.620	0.0%	69.7

Table 3: QC means from per-target `qc.json` (post length-adaptive scaling).

Config	Eff. scaffolds	Eff. regions	Eff. seeds	Eff. kissing	Mean refined	Mean cand.
v3	19.0	5.0	63.4	63.4	207	945
B4	19.0	5.0	63.4	63.4	190	985
U4	19.0	5.0	63.4	63.4	194	980
R10	19.0	12.7	63.4	63.4	172	882
S150	19.0	5.0	190.2	63.4	220	1063
K300	19.0	5.0	63.4	380.4	191	980
Sc30	38.1	5.0	63.4	63.4	194	1138
Hi	38.1	12.7	190.2	380.4	191	887

1 - CDFs (F1)

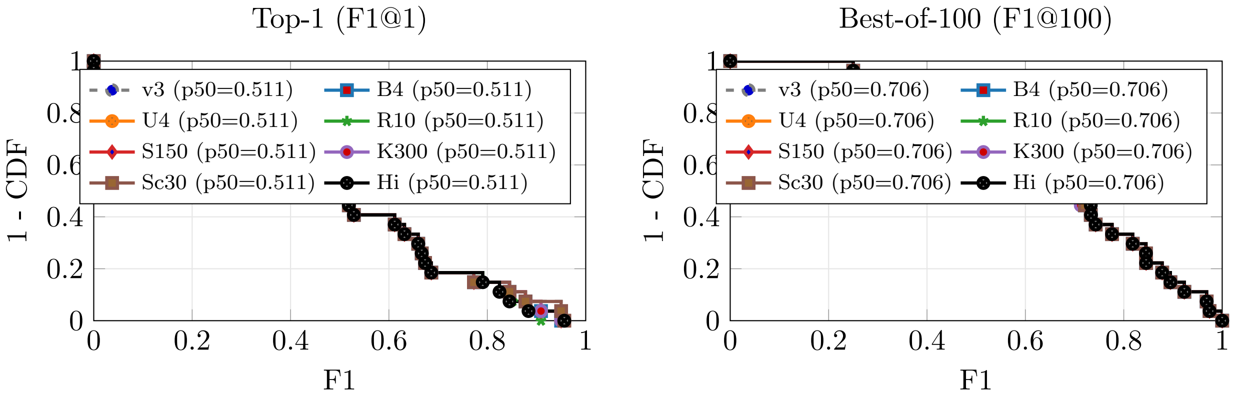

Figure 1: *
F1@1

Figure 2: *
F1@100

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

Runtime

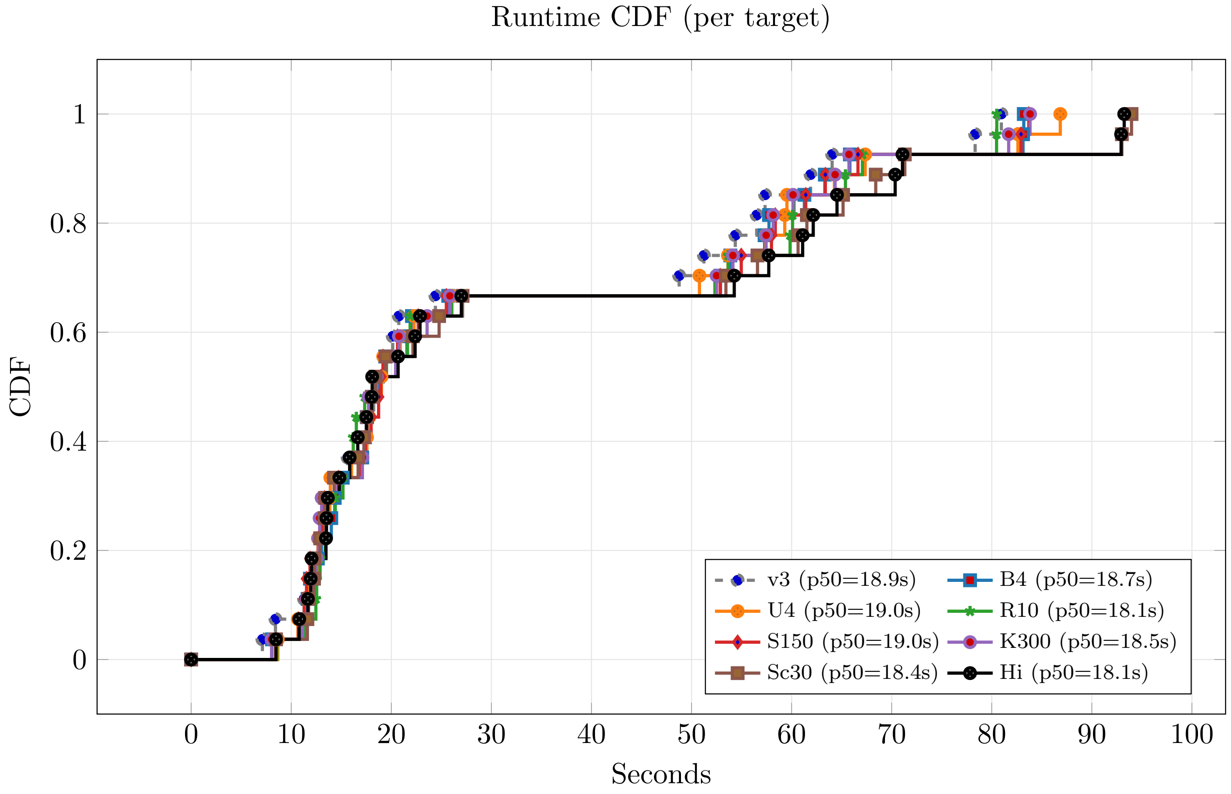

Figure 4: Empirical CDF of per-target wall-clock runtime. Legend includes p50 runtime.

Precision/Recall (v3 vs B4 vs Hi)

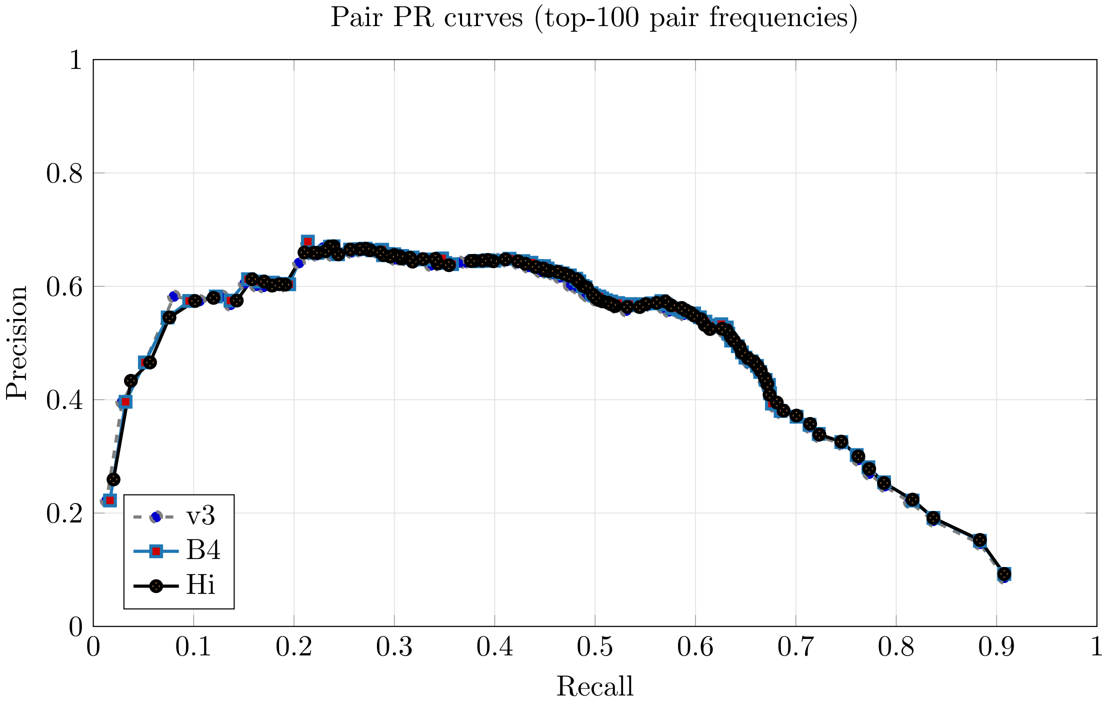

Figure 5: Pair-level PR curves computed by thresholding pair probabilities estimated as frequency across the top-100 structures.