

# RNAnneal-ss top-K output size test (under300 representative50) — v1

## Goal

Measure whether emitting more structures improves oracle best-of- $K$  accuracy, and where diminishing returns start, by evaluating prefixes of a single ranked output list.

## Setup

- **Targets:** representative50 subset (truth length  $\leq 300$ ) from FR3D/BGSU under400.
- **Predictor:** RNAnneal-ss (v3-high settings), run once with `--top-k 500`.
- **Scaffolds:** EternaFold (Fold + AllSub-like). Duplex sampling uses RNAstructure `DuplexFold`.
- **Metric:** for each  $K \in \{1, 50, 100, 200, 500\}$  compute the *oracle* best F1 among the first  $K$  outputs.

## Summary

Table 1: Top-K oracle metrics vs output size (N=50; higher is better).

K	Mean F1	Med F1	Fail	Mean P	Mean R	Mean eff.K
@1	0.536	0.530	6.0%	0.513	0.589	1.000
@50	0.629	0.617	2.0%	0.605	0.683	50.000
@100	0.649	0.691	2.0%	0.625	0.703	99.980
@200	0.673	0.697	0.0%	0.649	0.727	194.620
@500	0.733	0.754	0.0%	0.710	0.785	426.840

## Do longer RNAs benefit more?

- Mean  $\Delta F1$  @50: short=0.109, long=0.080 (long-short=-0.029).
- Mean  $\Delta F1$  @100: short=0.146, long=0.085 (long-short=-0.061).
- Mean  $\Delta F1$  @200: short=0.176, long=0.105 (long-short=-0.071).
- Mean  $\Delta F1$  @500: short=0.233, long=0.167 (long-short=-0.067).
- Spearman  $\rho$  between truth length and  $\Delta F1$  @500: -0.119.

Table 2: Mean/median  $\Delta F1$  (best-of- $K$  minus best-of-1), stratified by RNA length.

Group	K	N	Mean $\Delta F1$	Med $\Delta F1$
short( $i=150$ )	@50	23	0.109	0.068
short( $i=150$ )	@100	23	0.146	0.074
short( $i=150$ )	@200	23	0.176	0.091
short( $i=150$ )	@500	23	0.233	0.099
long(151-300)	@50	27	0.080	0.066
long(151-300)	@100	27	0.085	0.074
long(151-300)	@200	27	0.105	0.086
long(151-300)	@500	27	0.167	0.164

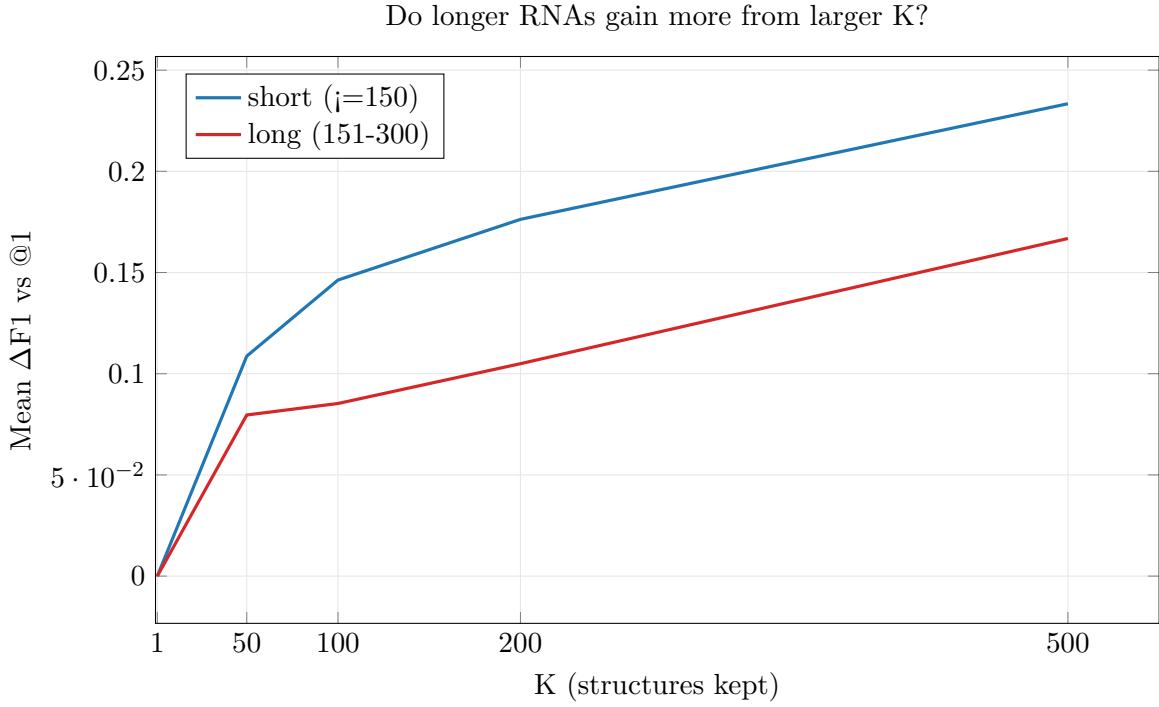


Figure 1: Mean improvement (best-of- $K$  minus best-of-1) for short vs long targets.

## Diminishing returns

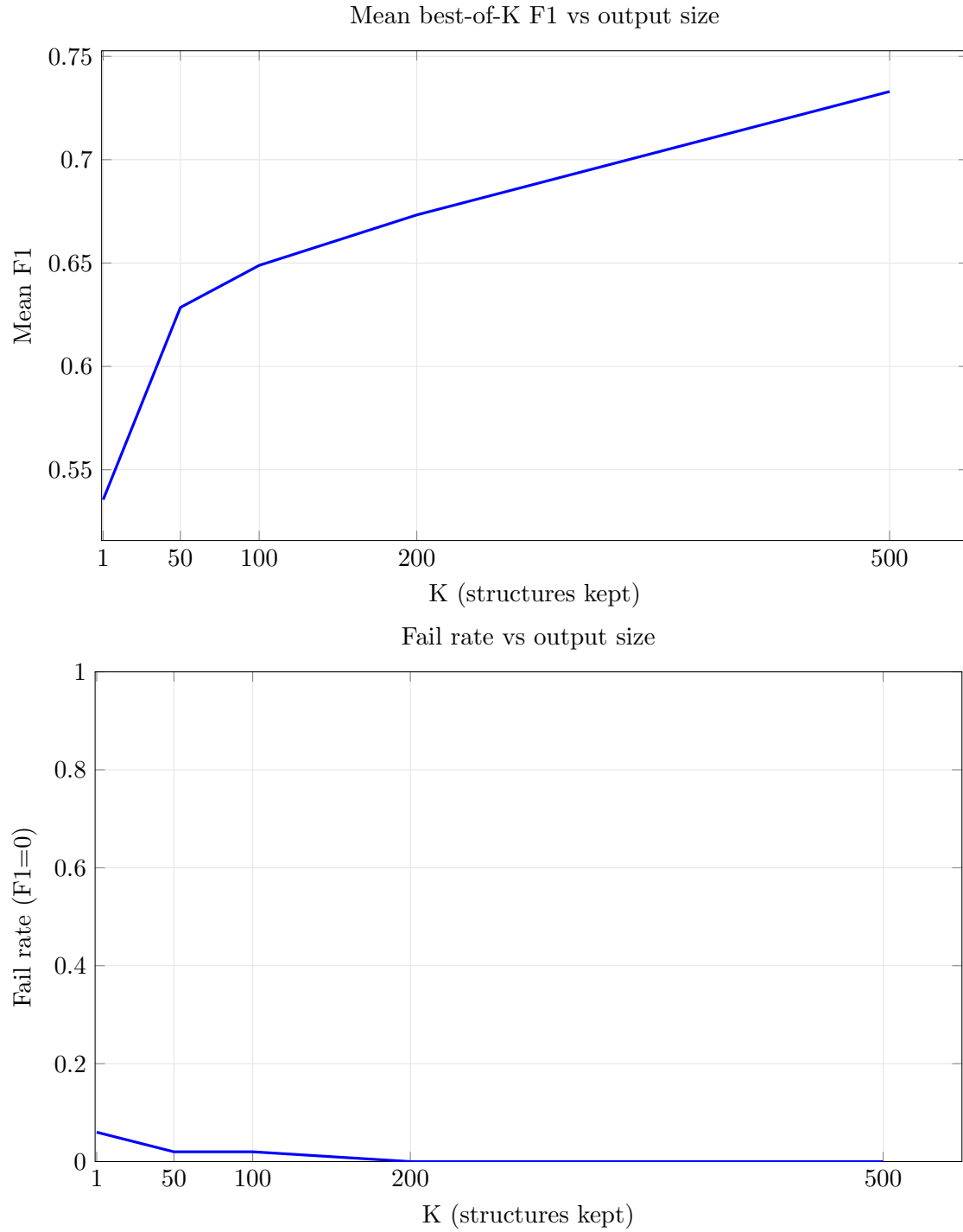


Figure 2: Mean best-of- $K$  F1 and fail rate as a function of output size.

## CDF

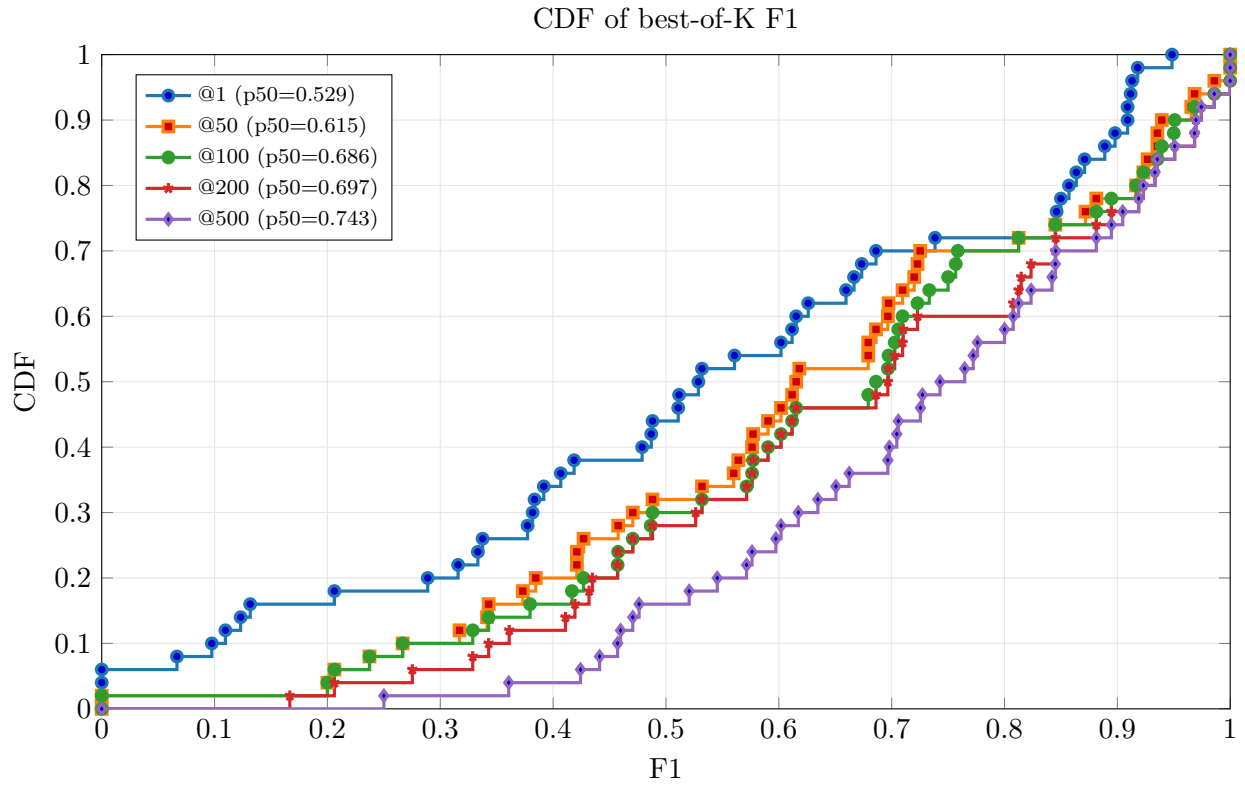


Figure 3: Empirical CDF for best-of- $K$  F1. Legend includes the median (p50) F1.