u400 benchmark: Ensemble RNAnneal-ss vs baselines across K — v1

Scope

Full u400 non-rRNA benchmark (FR3D/BGSU representative set, truth length ≤ 400). We compare an **ensemble** (RNAnneal-ss + LinearFold-V + EternaFold candidate pooling) to **LinearFold-V** and **EternaFold** across $K \in \{1, 50, 100, 200, 500\}$.

Metrics

- **F1@K**: for each target, the *best* F1 among the first K ranked outputs (oracle within prefix).

- **MCC@K**: MCC of the same structure that achieved best F1 within the first K outputs.

Results

Table 1: Best-of-K F1 on u400 non-rRNA benchmark (N=804 targets; oracle within top-K prefix).

K	Ensemble mean	Ensemble med	LF-V mean	LF-V med	EFold mean	EFold med
@1	0.650	0.725	0.622	0.682	0.656	0.733
@50	0.755	0.837	0.788	0.843	0.803	0.863
@100	0.786	0.859	0.794	0.846	0.820	0.872
@200	0.830	0.889	0.796	0.847	0.833	0.889
@500	0.850	0.903	0.797	0.847	0.846	0.900

Table 2: Best-of-K MCC on u400 non-rRNA benchmark (N=804 targets; MCC for the structure achieving best F1 within top-K prefix).

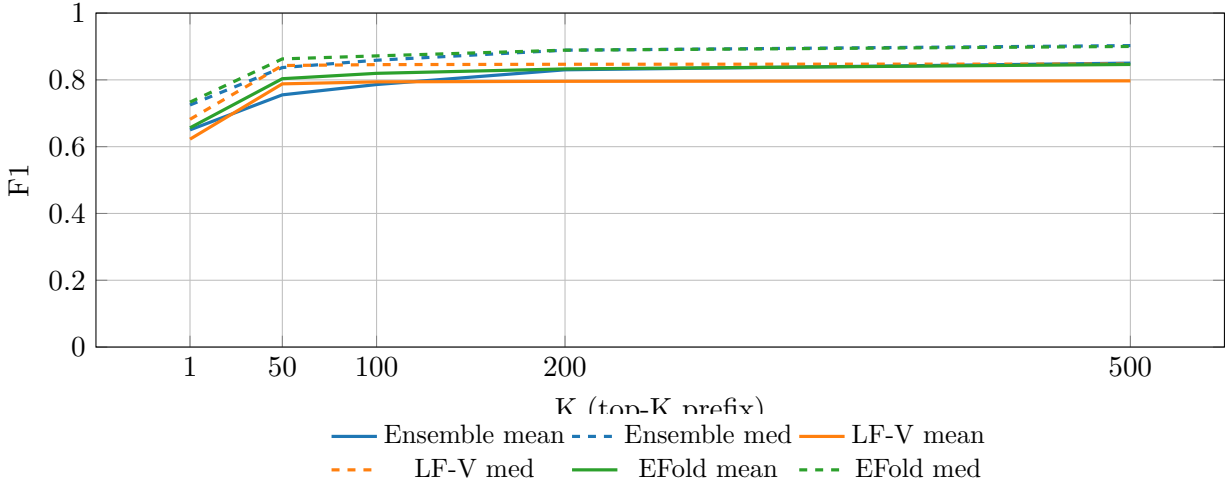| K	Ensemble mean	Ensemble med	LF-V mean	LF-V med	EFold mean	EFold med
@1	0.656	0.726	0.628	0.691	0.663	0.738
@50	0.761	0.842	0.796	0.847	0.810	0.865
@100	0.792	0.865	0.802	0.850	0.826	0.874
@200	0.837	0.889	0.803	0.850	0.839	0.889
@500	0.856	0.904	0.804	0.850	0.852	0.900

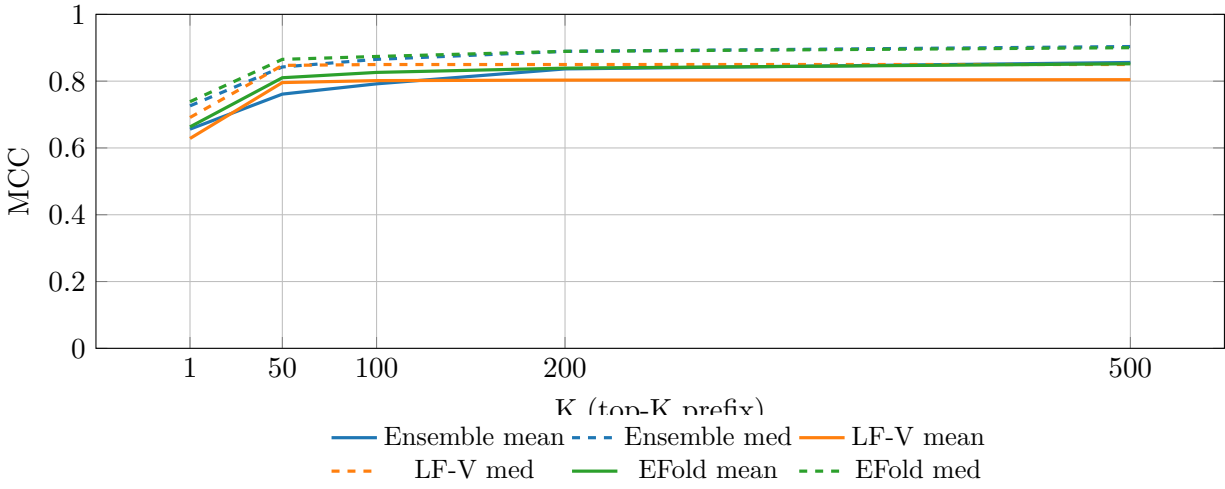Figure 1: F1 as a function of K (solid = mean, dashed = median; lines only).

Figure 2: MCC as a function of K (solid = mean, dashed = median; lines only).