Methods

We compare RNA secondary-structure similarity between two exons, A and B, of lengths $L = |A|$ and $M = |B|$. Below we describe each step of the metric's construction, explain every symbol and parameter, and outline how we tune them.

---

1. Top-percentile filtering

- **Input:** a TSV of all window comparisons, each row j with learned distance d_j.
- **Percentile** $p \in (0, 100]$: sort by d_j and retain the top-N windows:

$$N = \left\lceil \frac{p}{100} \times (\text{total rows}) \right\rceil, \quad j = 1, \ldots, N.$$

- **Rationale:** focuses on the most-similar structural matches and reduces noise.

---

2. Window notation and ranking

For each retained window j:

- (s_j^A, e_j^A): start/end positions on exon A (covers $a = s_j^A, \ldots, e_j^A$).
- (s_j^B, e_j^B): start/end positions on exon B.
- **Rank** $r_j \in \{1, \ldots, N\}$: sorted position by d_j (1 = best).

---

3. Per-cell matrix & numerator N_{ab}

Build an $L \times M$ matrix aggregating each window's contribution to base-pair (a, b).

3.1 Diagonal offset

$$\Delta_{ab}^{(j)} = \left| \, (a - s_j^A) \; - \; (b - s_j^B) \right| \quad (\geq 0).$$

3.2 Rank-decay weight

$$f_{\text{num}}(r_j) = r_j^{-\alpha_1}, \quad \alpha_1 > 0.$$

3.3 Off-diagonal decay

$$w_{\text{num}}^{(j)}(a, b) = \exp\!\left(-\beta_1 \, \Delta_{ab}^{(j)}\right), \quad \beta_1 \geq 0.$$

3.4 Numerator accumulation

$$N_{ab} = \sum_{j:\,(a,b)\in \mathrm{win}_j} \frac{1}{r_j^{\alpha_1}} \exp(-\beta_1 \Delta_{ab}^{(j)})$$

Interpretation: sums all "votes" from windows covering (a, b), weighted by rank and diagonal proximity.

---

4. Denominator D_{ab} (redundancy penalty)

To avoid over-counting overlapping windows:

4.1 Rank-decay penalty

$$f_{\mathrm{den}}(r_j) = r_j^{-\alpha_2}, \quad \alpha_2 > 0.$$

4.2 Positional-decay penalty

$$w_{\mathrm{den}}^{(j)}(a, b) = \exp(-\beta_2 \Delta_{ab}^{(j)}), \quad \beta_2 \geq 0.$$

4.3 Denominator accumulation

$$D_{ab} = \sum_{j:\,(a,b)\in \mathrm{win}_j} \frac{1}{r_j^{\alpha_2}} \exp(-\beta_2 \Delta_{ab}^{(j)})$$

Interpretation: measures the total "mass" of overlapping windows; larger D_{ab} more redundancy.

---

5. Combining signal & redundancy

Introduce $\gamma \in [0, 1]$ and define, for $D_{ab} > 0$,

$$M_{ab} = \frac{N_{ab}}{D_{ab}^{\gamma}}$$

- $\gamma = 0$: no penalty ($M_{ab} = N_{ab}$)
- $\gamma = 1$: full normalization ($M_{ab} = N_{ab}/D_{ab}$)

---

6. Global similarity score

Sum over all base-pairs:

$$G(A,B) = \sum_{a=1}^{L} \sum_{b=1}^{M} M_{ab}.$$

Rewards both **coverage** (many interacting cells) and **strength** (high weights).

---

7. Hyperparameter optimization with Optuna

Tune $\{p, \alpha_1, \alpha_2, \beta_1, \beta_2, \gamma\}$ to maximize separation of the known true-positive pair (ENSE00001655346.1, ENSE00004286647.1):

1. Compute G_{tp} and let $G_{(1)} \geq G_{(2)} \geq G_{(3)}$ be the next three best scores.

2. Define relative margin

$$\text{margin} = \frac{G_{\text{tp}} - \frac{1}{3}(G_{(1)} + G_{(2)} + G_{(3)})}{\frac{1}{3}(G_{(1)} + G_{(2)} + G_{(3)})}.$$

3. Enforce rank-1: penalize any trial that doesn't place the true positive atop.

4. Use Optuna's Bayesian sampler for 100–200 trials; inspect **optimization history**, **parameter importance**, and **slice plots**.

Table of parameters

Parameter	Symbol	Description
Percentile	p	Fraction of top-distance rows retained
Num-rank decay	α_1	Exponent in numerator rank weight $r^{-\alpha_1}$
Den-rank decay	α_2	Exponent in denominator rank weight $r^{-\alpha_2}$
Num-positional	β_1	Off-diagonal decay in numerator $e^{-\beta_1 \Delta}$
Den-positional	β_2	Off-diagonal decay in denominator $e^{-\beta_2 \Delta}$
Redundancy	γ	Exponent on D_{ab} penalty (0=no penalty;1=full)

This framework balances **signal amplification** with **redundancy control**, yielding a robust per-base and global similarity metric.