Computational Design of RNA Thermoswitches for High-Temperature Genetic Control in *Bacillus subtilis*

Data Science Project

Student Name: Md Abir Hossain

Student ID: 52322252

Programme: MSc in Data Science

School: School of Natural and Computing Sciences, University of Aberdeen

Project Supervisor: Dr Andrew Angel

Industry Supervisor: Dr Alexander Speakman (EVA Biosystems)

Submission date: 29 August 2025

Contents

Abstract

Bacteria such as *Bacillus subtilis* are increasingly engineered for industrial tasks, but many applications still need a simple, clean way to turn genes on and off without chemical inducers. RNA thermoswitches offer such control: compact RNA elements that keep the ribosome-binding site (RBS) sequestered at low temperature (OFF) and expose it at higher temperature (ON). For processes that include deliberate heat steps, it is desirable to deploy switches that activate well above typical growth temperatures. Here we target devices that turn on around 70 °C.

We present a fully computational approach to design high-temperature RNA thermoswitches. First, we crafted a prototype scaffold that occludes the RBS when cool and then explored a targeted neighbourhood of variants around this seed. Each candidate was evaluated with a thermodynamic RNA-folding model to estimate how RBS accessibility shifts with temperature, providing a physics-grounded criterion for ON/OFF behaviour. To accelerate screening, we then trained a lightweight machine-learning ranker to prioritise candidates predicted to switch near the target temperature with low basal leak and a steep transition, reducing reliance on repeated full folding calculations.

This pipeline produced a ranked, diverse shortlist of thermoswitch designs suitable for experimental testing in *Bacillus subtilis*. More broadly, the results suggest that combining thermodynamic scoring with learned ranking can streamline future design cycles, focusing effort on the most promising candidates and lowering the computational cost of exploring design space; while preserving the interpretability of a model at the core of the workflow.

1. Introduction

Living cells encounter abrupt temperature shifts in both natural and engineered settings, and translation can respond on short timescales before global transcriptional programmes fully settle. RNA thermoswitches (RNA thermometers; RNATs) exploit this by folding in a way that hides the ribosome-binding site (RBS) (Shine–Dalgarno; SD) at lower temperatures (OFF) and reveals it at higher temperatures (ON). In bacteria, this principle is conserved even though specific sequences and placements vary. Such switches are attractive for biomanufacturing workflows that include controlled heat pulses. [1], [2], [3].

RNA thermoswitches, also called RNA thermometers, regulate translation by coupling the folding of an mRNA leader to temperature. A canonical design places a hairpin across the ribosome-binding site (RBS). At low temperature the stem occludes the RBS (OFF), while warming destabilises the stem and exposes the RBS (ON) [1]. The RBS, often defined via the Shine–Dalgarno (SD) interaction between mRNA and the 16S rRNA anti-SD, is widely used for ribosome recruitment across bacteria though its usage and sequence vary substantially. [3], [4], [5].

In RNA, bases pair most stably as G–C and A–U, with G–U wobble pairs weaker. Consecutive pairs form a double-stranded stem capped by a single-stranded loop (a hairpin). At lower temperature the stem can sequester the RBS; as temperature rises, weaker pairs melt and the stem opens, exposing the RBS. [1].

Figure 1a schematically illustrates this mechanism: single-stranded RNA forms a base-paired hairpin that sequesters the RBS (OFF) at low temperature and opens to expose the RBS (ON) around 70 °C. We target around 70 °C because it is well above typical growth temperatures, aligns with deliberate heat steps used in many processes, and minimises spurious activation from small day-to-day fluctuations. High-temperature applications are common in bioprocessing, such as bio recycling of plastic waste, to better match the glass transition temperature of polymers. In particular, allowing cells to respond to short, high-temperature pulses via RNA-level regulation has applications in the production of composite biomaterials, enabling control of cells during manufacturing to optimise survival, for example, through control of sporulation (e.g., EVA Biosystems). [6], [7].

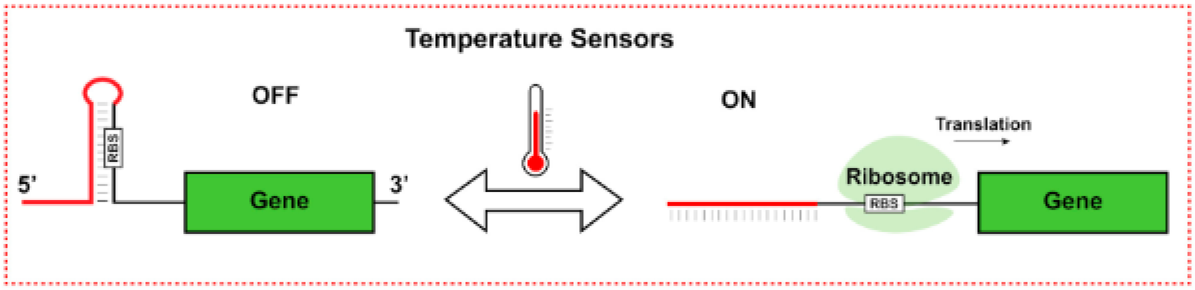

Figure 1a— Mechanism schematic: RBS occluded at low temperature (OFF) and exposed near the ~70 °C transition (ON). [8]

For applications, three properties matter: (i) low leak in the OFF state (RBS well-occluded), (ii) a steep activation near the target temperature (narrow transition), and (iii) a high ON plateau. These properties can be tuned by classic design levers: stem length and GC content (baseline stability), loop length (nucleation/closing), and purposeful "softening" of the anti-RBS using GU wobble pairs or a strategic mismatch to shift the melting window. Because GC pairs form three hydrogen bonds whereas AU pairs form two (and GU wobble pairs are weaker), base composition sets the stem's stability**:** more GC generally raises the melting temperature and strengthens the OFF state, while enriching AU or introducing GU/mismatches lowers stability, helping sharpen or shift the switch toward the target temperature [5].

We take a physics-based design approach, using nearest-neighbour energetics (a thermodynamic model that sums base-stacking contributions) to predict how the 5′-UTR (the untranslated leader region upstream of the start codon) folds across temperature. Aggregating these over the ribosome-binding site (RBS; Shine–Dalgarno/SD), the short motif that recruits the bacterial ribosome, yields an RBS-openness curve versus temperature, where RBS-openness is the fraction of SD nucleotides predicted to be unpaired at a given temperature. [9], [10], [11], [12], [13].

Aims and contributions: Our aim is to design RNA thermoswitches that stay OFF at ≤ 50–55 °C and switch ON around ~70 °C with low leak, a steep transition, and a solid ON plateau. By combining NUPACK software with a lightweight machine-learning helper, we can explore thousands of candidates in silico, yield the most promising ones, and send only a small, order-ready(i.e., each sequence that passes synthesis quality control, QC and meets the design parameters) shortlist to the lab, avoiding the expense and time of inserting every variant into bacteria and testing them experimentally. In short, ML helps us spot patterns we had miss and turns them into honest probabilities for each design, letting us focus experiments on the most promising sequences.

1) Software-first screening at scale: simulations plus ML ranking let us try out large data before any experimental work, cutting cost and cycle time while keeping a reliable output.

2) Clear design targets, measured consistently: each candidate is summarised by its RBS-openness curve and a few simple metrics; leak (OFF-state openness at low temperature), T50 (temperature where openness = 0.5), steepness (how sharp the transition is), and ON plateau, so choices are transparent and comparable.

3) Simple, reusable scaffold: a small set of sequence “knobs (tunable design parameters e.g., stem strength, loop size, placement) steers the switch towards the desired temperature.

4) Order-ready, diverse shortlist: we provide a compact dataset sized to typical synthesis batches, chosen to avoid near-duplicates so experiments start with a high-probability panel rather than trial-and-error.

Research questions:

RQ1: Can we tune an RNA thermoswitch to turn ON near 70 °C while maintaining low leak and a steep transition?

RQ2: Given limited labels and class imbalance, can a small, interpretable feature set support reliable Top-K ranking (PR-AUC/Precision@K)?

RQ3: Which QC and diversification steps best improve synthesis readiness and reduce redundancy without sacrificing predicted performance?

2. Data

2.1 Sequence Library

Table 2.1 — Dataset summary (N=700)

Metric	Value
Total sequences (N)	700
Unique sequences	700
Length (nt)	47.63 ± 1.12
GC fraction	0.29 ± 0.03
T50 (°C)	70.67 ± 2.13
Slope@70 °C	0.047 ± 0.010
% meeting spec	16.0%

We analysed 700 labelled designs in the final analysis snapshot. Our library consists of RNA designs created in silico by our thermoswitch pipeline (Methods 3); all sequences share a fixed Shine–Dalgarno (AGGAGG) RBS motif at annotated indices and differ only in scaffold knobs, i.e., the tunable parameters of the sequence scaffold: extra stem pairs (number of paired bases added to the stem), mid-stem GC bias (GC proportion in the central stem that sets baseline stability), and loop length (number of nucleotides in the hairpin loop). For each design we export the full sequence, RBS start–end indices, and knob settings, together with computed temperature-dependent curves. Sequences were produced across multiple runs by repeating local search from new random seeds; local search meaning iterative small edits around a seed sequence, scoring nearby variants and keeping improvements within the allowed knob ranges. The final snapshot merges these runs into a single table (library_combined.csv), alongside a few smaller demonstrator sets (e.g., library_100.csv). Lengths concentrate around 46–50 nt and GC fractions around 0.25–0.36 (see Results, Figs. 4.9–4.10). The RBS context follows SD-mediated initiation conventions but tolerates natural sequence diversity. (see Table 2.1) [4], [11].

2.2 Physics-derived labels/targets

Table 2.2 — Spec thresholds used for labels

Metric	Spec/Threshold
closed_low	≥ 0.9
open_midhi	≥ 0.6

open_high	≥ 0.8
T50	69.0 – 72.5 °C
slope70	≥ 0.04 per °C

We derive labels from physics-based folding analysis. Using NUPACK v4, we compute equilibrium base-pairing probabilities with the Utilities function pairs and obtain per-base unpaired probabilities as the diagonal complement. Aggregating unpaired probabilities across the ribosome-binding site (RBS; Shine–Dalgarno/SD) yields an RBS-openness curve evaluated every 2 °C from 50 to 85 °C, where "openness" means the fraction of SD nucleotides predicted to be unpaired at that temperature. From this curve we extract metrics that summarise generated RNA thermoswitch sequence behaviour: closed_low (mean openness below ~68 °C), open_midhi (mean openness 72–78 °C), open_high (mean openness ≥ 80 °C), an interpolated T50 (temperature where openness = 0.5), and slope@70 °C (the local derivative of RBS-openness at 70 °C, i.e., per-degree change estimated by a finite difference). Nearest-neighbour thermodynamic parameters underlying these calculations incorporate AU/GC differences and GU wobble effects. Classification labels (pass/fail) are assigned by a specification that requires T50 within a target band (e.g., 69–72.5 °C), slope@70 °C ≥ 0.04 per °C, and acceptable openness in the OFF/ON regimes; these are the same guards later used for shortlist QC. [9], [10], [11], [12], [13].

2.3 Splits, leakage control, and features

We adopt a stratified 80/20 train–test split for all headline metrics and a stratified 5-fold cross-validation protocol for robustness reporting. To avoid leakage from near duplicates, we collapse exact sequence duplicates and monitor pairwise similarity using 3-mer Jaccard distance which means to split each sequence into overlapping 3-letter "words" (3-mers), compute similarity as shared 3-mers ÷ total unique 3-mers across both sequences, and define distance = 1 − similarity [5]; where appropriate, we ensure that extremely similar sequences do not straddle train/test folds. Model inputs combine global features (length, GC, scaffold knobs), k-mer frequencies (16 di- and 64 tri-nucleotides), and a positional one-hot window spanning the RBS (±10 upstream, +20 downstream). These features are compact, interpretable, and avoid using ground truth labels directly, which would induce target leakage. [14], [15].

3. Methods

3.1 Design & scoring

We start from a simple scaffold that places a hairpin over a fixed AGGAGG RBS. To obtain this design, rather than manual tuning or using NUPACK's sequence-design module, we used an automated local search (greedy hill-climb with multiple random restarts) over a small set of scaffold knobs; extra stem pairs, mid-stem GC bias, and loop length. At each step the anti-RBS is lightly softened (occasional GU wobble and, optionally, a single central mismatch), the candidate is scored from its RBS-openness curve, and only improvements within allowed knob ranges are kept. Simulation settings included using NUPACK v4 in RNA mode (*material='rna'*, nearest-neighbour energetics), we evaluate a single-strand complex on a 50–85 °C grid in 2 °C increments (50, 52, …, 84 °C) under [Na^+] = 1.0 M and [Mg^{2+}] = 0.0 M. From the Utilities function pairs we take the diagonal entries as per-base unpaired probabilities, average these across the RBS to obtain an openness-versus-temperature curve, and summarise five metrics: OFF leak (closed_low), ON levels (open_midhi, open_high), T50 (temperature where openness = 0.5), and the local slope at 70 °C (finite-difference estimate of steepness). [5], [3].

3.2 Library generation

We then generated multiple small libraries (e.g., library_100.csv, library_200.csv) with a reproducible script (make_library_csv) and later merged them into a single 700-sequence table (library_combined.csv). Each library is produced with the same search settings (LIB_PARAMS): ep_range=(2,3) (two or three extra stem pairs), pGC_range=(0.45,0.58) (mid-stem GC bias), loop_range=(7,9) (loop length), wobble_prob=0.30 (light GU wobble softening of the anti-RBS), center_mismatch=True (optionally inserting one central mismatch), clamp_ends=True (GC clamps at stem ends), and a local search of restarts=14 and steps=90. Local search here means a greedy hill-climb around a seed: at each step we propose small knob moves (e.g., extra_pairs ±1, pGC nudged within range, loop_len ±1), score the candidate from its RBS-openness curve, and keep it only if the score improves (within the allowed ranges). Reproducibility and diversity across files comes from the seed schedule (seed=seed0+i for record i), which triggers fresh random restarts per design.

For each design, we compute the RBS-openness curve on a 50–85 °C grid in 2 °C increments and write one row with: id, seq, len, rbs_start, rbs_end, knob values (extra_pairs, pGC, loop_len), and the derived metrics closed_low (OFF leak), open_midhi and open_high (ON levels), slope70 (local slope at 70 °C), T50 (linear interpolation where openness crosses 0.5), slope_max, and reentrant_pen (a penalty of 0.3 if the curve drops by >2% anywhere). We also store a composite score: score = 2.0*closed_low + 3.0*open_midhi + 1.5*open_high + 4.0*slope70 − reentrant_pen, and a compact text field curve containing the openness values across the temperature grid. [9], These weights were empirically tuned from pilot sweeps to match our design priorities, maximizing the transition steepness at ~70 °C (4.0), promoting RBS opening in the mid–high regime (3.0) and at high T (1.5), suppressing low-temperature leak (2.0), and penalizing any re-entrant closure so higher scores reflect sharper, cleaner ON behavior. [11], [13].

The example call, make_library_csv("results/library_100.csv",N=100),produces 100 such records per file; subsequent runs at the same settings yield additional files (library_200.csv, etc.), which we concatenate to form the final 700-sequence snapshot used for analysis.

3.3 Features, model and evaluation

For ranking we use compact, reproducible features: sequence length (nt) and GC fraction (G or C bases ÷ length); the three scaffold knobs which are extra stem pairs (integer count), mid-stem GC bias pGC (0–1), and loop_len (nt); 16 dinucleotide frequencies (normalised counts of all 2-mers over a sliding window); 64 trinucleotide frequencies (all 3-mers); and a positional one-hot window around the RBS spanning 10 nt upstream + the 6-nt SD motif + 20 nt downstream (36 positions × 4 nucleotides = 144 features, zero-padded if the sequence is shorter). [13], [14].

We train a gradient-boosting classifier (sklearn GradientBoostingClassifier) wrapped in a StandardScaler and isotonic calibration (CalibratedClassifierCV, 3-fold) with class-balanced sample weights; the calibrated scores are treated as approximate hit-rates. Data are split 80/20 (stratified) and we also run stratified 5-fold cross-validation. Evaluation covers discrimination (ROC-AUC, PR-AUC), calibration, and retrieval quality (Precision@K / Recall@K for K = 10, 20, 50). A train-time threshold for pass/fail can be chosen from the PR curve (e.g., F_β with β=2), but ranking itself uses the calibrated probabilities. [14], [16], [17], [18], [19], [20].

Shortlists apply guard bands, pre-filters on model predictions to enforce design targets, defined as predicted T50 ∈ [69.0, 72.5] °C and predicted slope@70 °C ≥ 0.040 per °C (candidates outside this window are excluded before ranking). We then apply a simple 3-mer diversity filter (max–min over 3-mer Jaccard distance) so the final Top-K avoids near-duplicates. (see Figure 4.4)

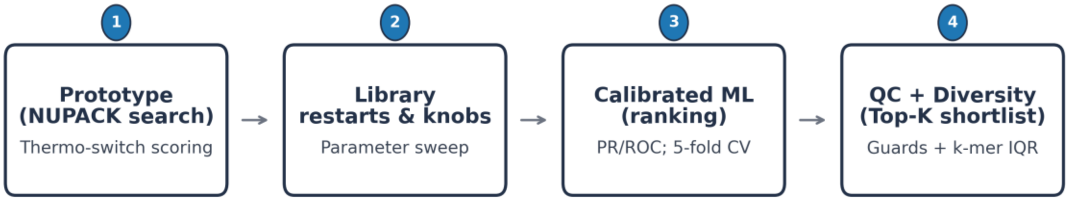

Figure 3a— Computational Pipeline (Overview)

Figure 3a summarises the computational workflow used in this study, from prototype design through library generation, calibrated ML ranking, and shortlist post-processing.

4. Results

4.1 Design outcome

Using the NUPACK-guided local search from Methods (3.1), we first obtained an optimised thermoswitch that occludes the RBS at low temperature and opens sharply around 70 °C. In brief, the guided search employs a greedy hill-climbing procedure with multiple random restarts over three scaffold knobs (extra stem pairs 2–3, mid-stem GC bias 0.45–0.58, loop length 7–9 nt): at each step we compute the RBS-openness curve with NUPACK v4 (material = “rna”) on a 50–85 °C grid in 2 °C increments under [Na^+] = 1.0 M, [Mg^{2+}] = 0.0 M and retain only moves that improve the composite score (details in Methods 3.1). The best candidate’s RBS-openness curve (Fig. 4.1) exhibits three desirable properties across temperature: (i) a low-leak OFF state from 50–65 °C where the RBS remains predominantly paired, (ii) a steep transition centred near 70 °C, and (iii) an ON plateau beyond ~75 °C where the RBS is mostly unpaired. The interpolated T50 is ≈ 70.5 °C, within the pre-specified 69–72.5 °C window, and the local slope around 70 °C is high, indicating switching within a narrow band rather than a gradual drift. [5], [9].

To answer, “if one can design a sequence by guided search, why perform shifts around it and add ML?”, a sequence found by a hill-climbing procedure is, by construction, a locally optimal solution under the chosen scoring function and initialisation; it does not ensure coverage of the constrained design space nor diversity for experimental testing. Accordingly, we expand around the prototype by repeating the same local search with different seeds to generate multiple libraries that are later merged. We then apply a calibrated gradient-boosting ranker to prioritise candidates, enforce guard bands on predictions (T50 ∈ [69.0, 72.5] °C; slope@70 °C ≥ 0.040 per °C), perform max–min 3-mer diversification to avoid near-duplicates, and run QC filters (GC range; homopolymers) before composing an order-ready shortlist. This yields a robust and diverse panel aligned to experimental throughput while keeping the physics-based criteria at the core. [3].

Table 4.1 — Headline performance

Metric	**Value**
ROC-AUC	0.951
PR-AUC (AP)	0.875
F_2-optimal Precision	0.75
F_2-optimal Recall	0.90
Precision@10	1.00
Recall@10	0.09
Precision@20	1.00
Recall@20	0.18
Precision@50	0.92
Recall@50	0.41
ECE (10-bin)	0.322

This prototype served a dual purpose. Practically, it provided an order-ready sequence, i.e., one that already passes our synthesis QC (GC fraction within 0.25–0.70 and no ≥7-nt homopolymer runs) and meets the design guard bands (predicted T50 ∈ [69.0, 72.5] °C and slope@70 °C ≥ 0.040 per °C), with annotated RBS start-end indices exported alongside the sequence, so it can be sent to a lab without further redesign. Methodologically, it defined feasible knob ranges for stem length/GC bias and loop size that we then used to generate the larger library via repeated hill-climbs with fresh random seeds. In other words, the library is not a blind random sample; it is concentrated around configurations that the prototype search identified as productive, which helps both modelling efficiency and downstream hit-rate. [14].

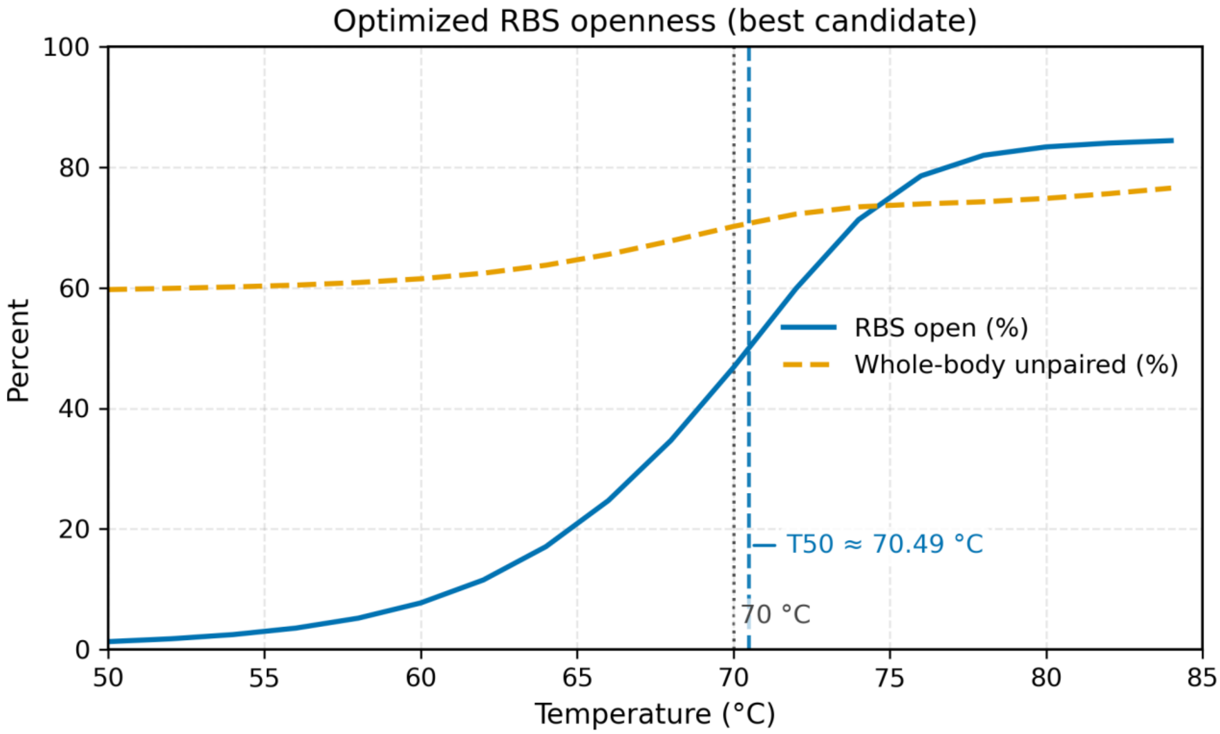

Figure 4.1 — Prototype RBS-openness vs temperature. Solid: RBS region; dashed: whole sequence unpaired fraction. Dotted line marks 70 °C; vertical line marks T50 ≈ 70.5 °C.

4.2 Screening performance

We next assessed how well the trained classifier separates sequences that satisfy the specification from those that do not. All results below are on the stratified 20% test split held out from training and calibration (the F_2 operating point is chosen on the training PR curve). On this held-out set the Precision–Recall curve achieves Average Precision, AP ≈ 0.867 (Fig. 4.2) and the ROC yields AUC ≈ 0.951 (Fig. 4.3), a combination that shows positives are ranked above negatives consistently while retaining precision as recall increases. The F_2-optimal point lies in the high-recall region, which matches our shortlist policy: we prefer to include borderline good candidates rather than miss them. [17], [21], [22] Since shortlists are selected by ranking rather than a fixed threshold, the reliability of the predicted probabilities matters. We obtain p(pass) by fitting a GradientBoostingClassifier and converting raw scores to probabilities with isotonic calibration (CalibratedClassifierCV, 3-fold) using only the training folds; we then evaluate predict_proba[:,1] on the test split. For Fig. 4.4, test-split predictions are binned into equal-width probability bins and we plot the observed positive fraction in each bin against the mean predicted probability (dashed diagonal = perfect calibration). The points are not exactly on the diagonal, lower-score bins are somewhat over-confident (below the line), and the highest bin is under-confident (above the line). Even so, the curve is monotonic with a healthy spread of scores, which preserves the ordering that drives Top-K selection and is sufficient for our rank-based shortlist. [18].

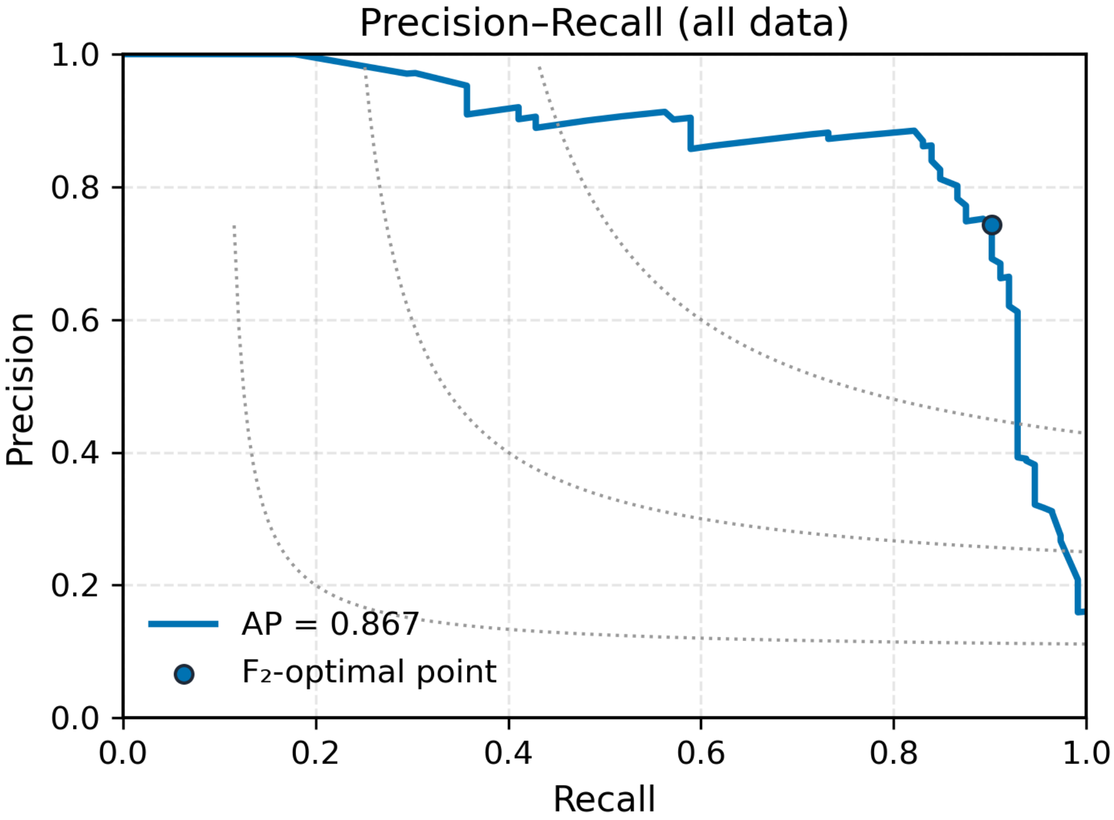

Figure 4.2 — Precision–Recall with iso-F1 curves. Dot marks the F_2-optimal operating point; AP ≈ 0.867.

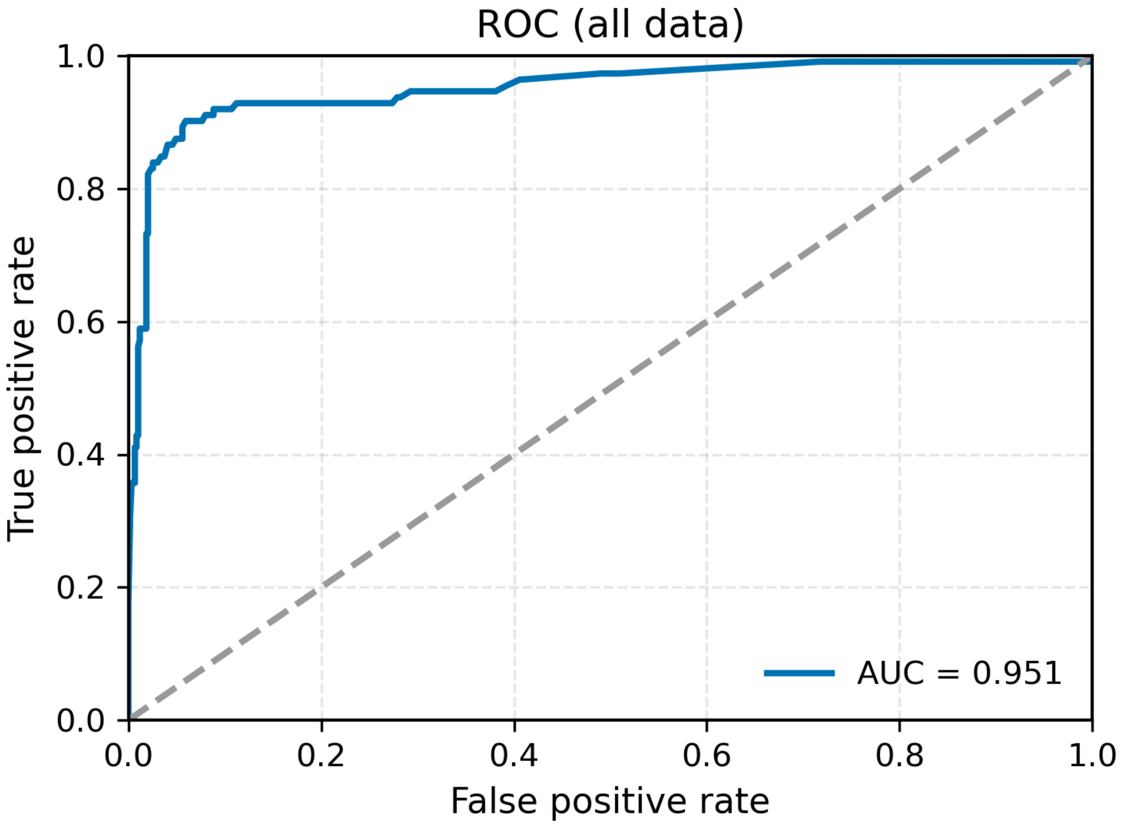

Figure 4.3 — ROC curve with AUC ≈ 0.951.

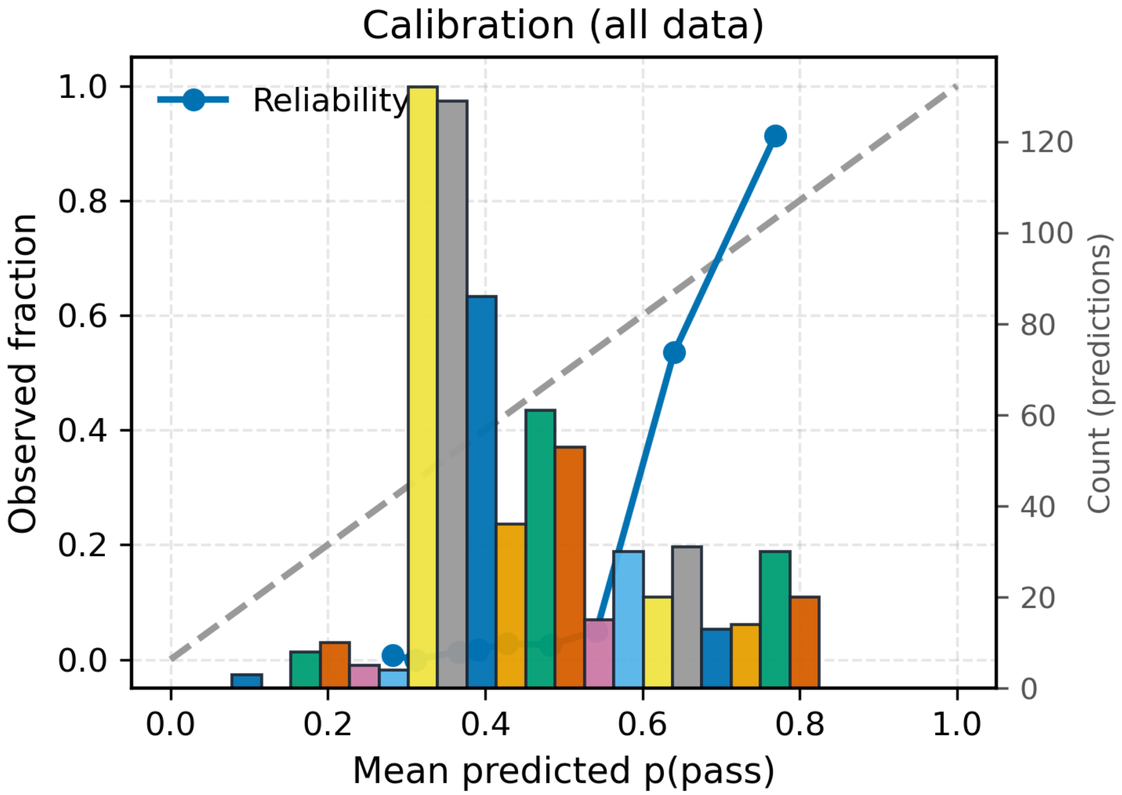

Figure 4.4 — Calibration (10 quantile bins) with marginal score histogram on a twin axis. Points close to the identity line indicate good probability calibration. [19], [20], [23].

4.3 Top-K retrieval

Ultimately, the practical question is how many good designs appear near the top of the ranked list. Using the descriptive all-data view, Precision@K is 1.00 for K=10 and K=20 and 0.92 for K=50 (Fig. 4.5). The companion Recall@K values are 0.09, 0.18, and 0.41, respectively (Fig. 4.6). Interpreted concretely, ordering the top 20 sequences is expected to yield ~20×1.00 = 20 predicted positives with nine of those being true hits given the base rate; expanding to 50 retains acceptable precision and recovers ~41% of all positives in the set. These values reflect the design difficulty: the pass label requires T50 in a tight band, a steep slope at 70 °C, and openness constraints, so even modest recall at stringent K can represent substantial enrichment over random selection. [18].

To check stability against sampling noise we repeated evaluation under stratified 5-fold cross-validation. The CV bars (Supplementary Fig. S2a–b) report mean±sd(standard deviation i.e., how much it deviates from the mean value): Precision@K = 0.32±0.15, 0.30±0.10, and 0.24±0.04 for K=10, 20, and 50; Recall@K = 0.14±0.07, 0.27±0.09, and 0.54±0.09. The absolute values are lower than the descriptive all-data view, as expected when training on fewer examples per fold, but the ordering is consistent: recall grows with K and precision decays gently, indicating the ranking is robust. The error bars are modest relative to the means, suggesting little fold-to-fold volatility. [24], [25].

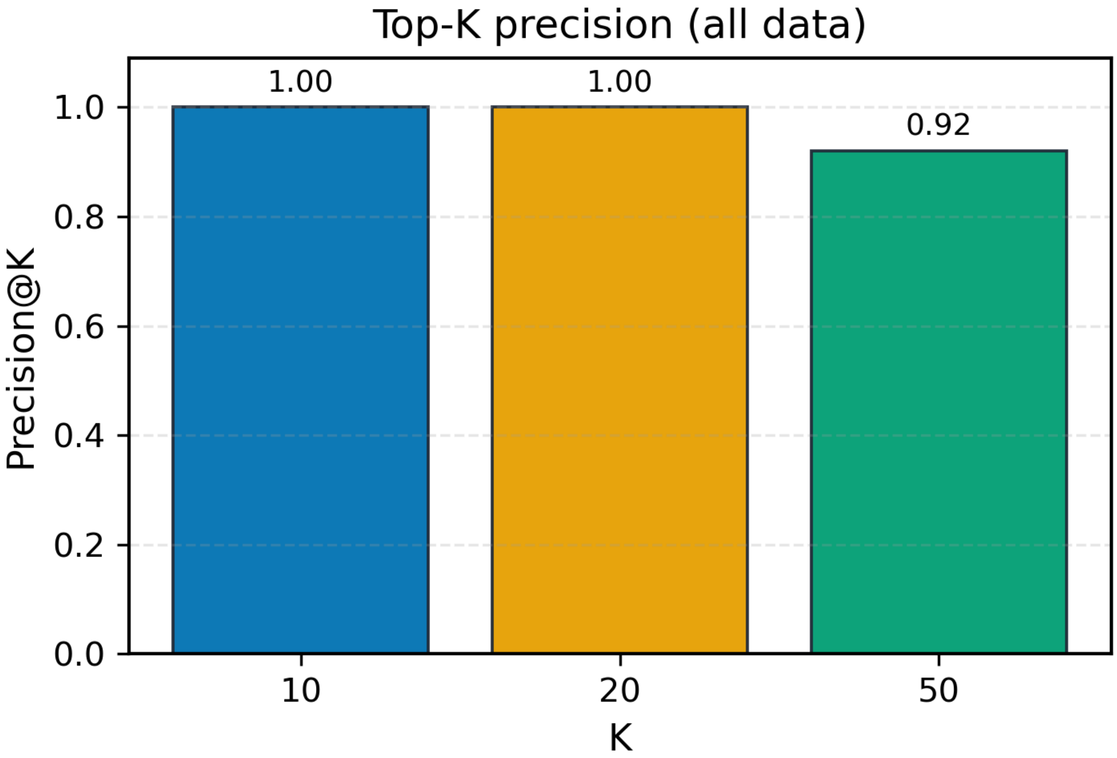

Figure 4.5 — Top-K Precision (descriptive view on all data): 1.00 at K=10 and K=20; 0.92 at K=50.

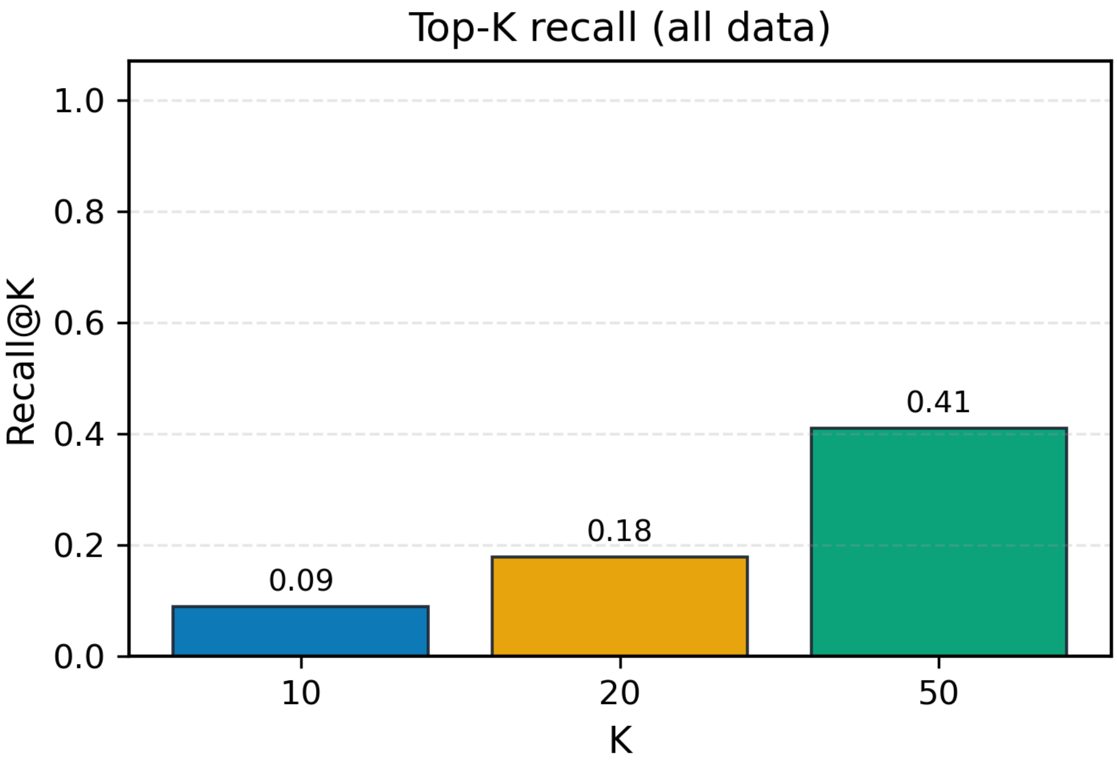

Figure 4.6 — Top-K Recall (descriptive view on all data): 0.09 at K=10; 0.18 at K=20; 0.41 at K=50.

4.4 Shortlist quality

Shortlists were filtered using the same guard rules as the specification (T50 ∈ [69, 72.5] °C; slope@70 °C ≥ 0.04). The distributions within the Top-20 (Figs. 4.7–4.8) confirm that these guards are comfortably met in practice: the T50 histogram is centred at μ ≈ 71.55 °C with most designs falling squarely inside the target band, and the slope distribution has μ ≈ 0.05 per °C with few, if any, values close to the minimum threshold. These properties make the shortlist resilient to minor modelling errors; small deviations are unlikely to push a design across a decision boundary.

Basic sequence QC supports synthesis readiness. The overall library shows tight control of length (μ ≈ 47.63 nt) and balanced base composition with GC around μ ≈ 0.29 (Figs. 4.9–4.10). We observed no long homopolymer runs in the shortlist (≤ 6 bases), and the GC band avoids extremes known to complicate DNA synthesis and cloning. Together, these checks reduce preventable experimental failure modes while keeping the RBS context fixed. [26], [27].

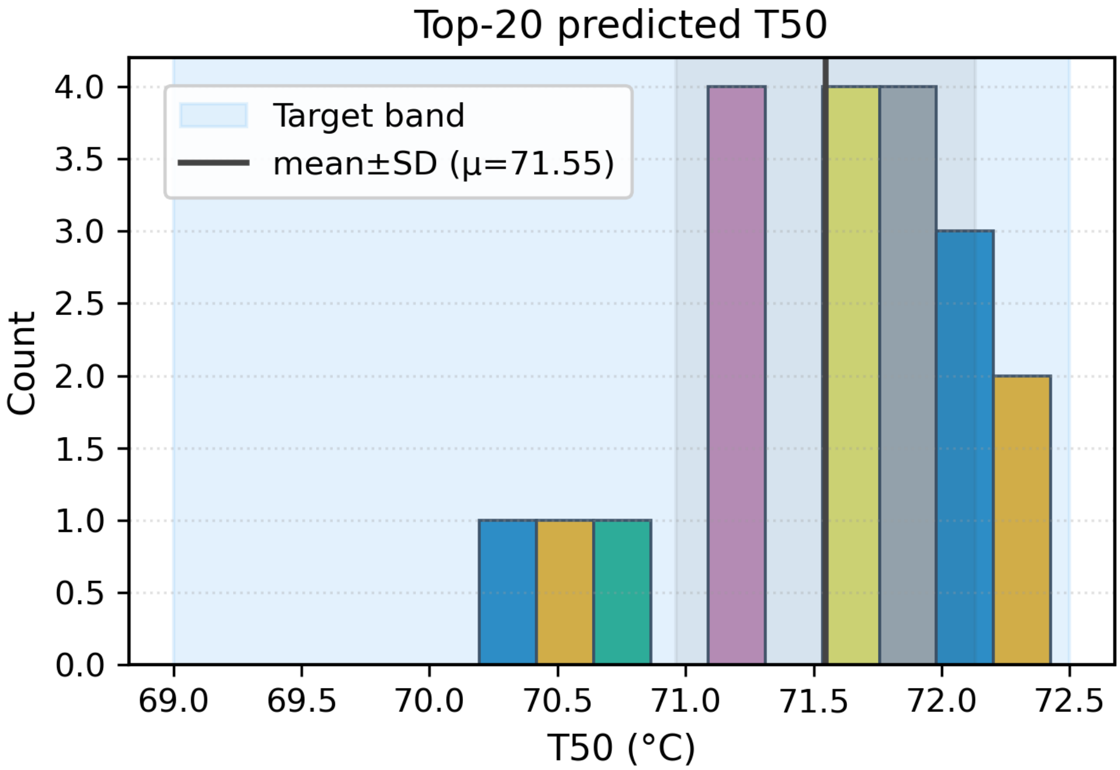

Figure 4.7 — Top-20 predicted T50 distribution with the 69–72.5 °C target band highlighted; mean ± SD overlay (μ ≈ 71.55 °C).

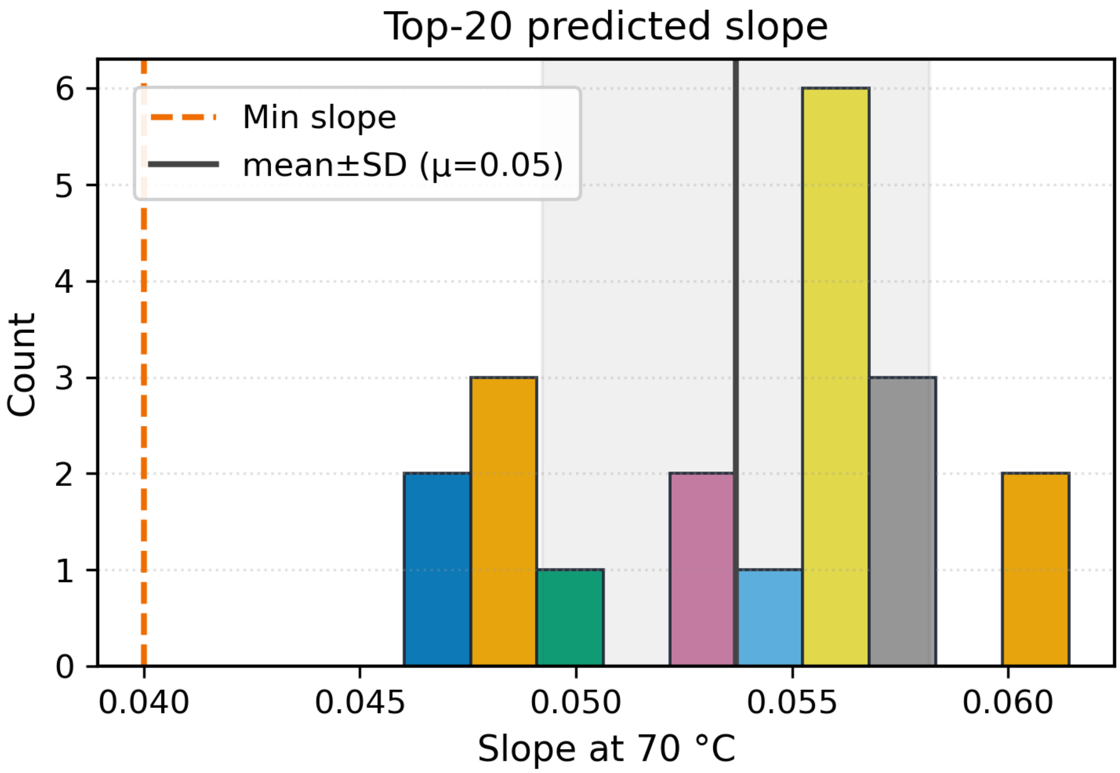

Figure 4.8 — Top-20 slope at 70 °C with the minimum threshold (0.04 per °C) shown as a dashed line; mean ± SD overlay (μ ≈ 0.05).

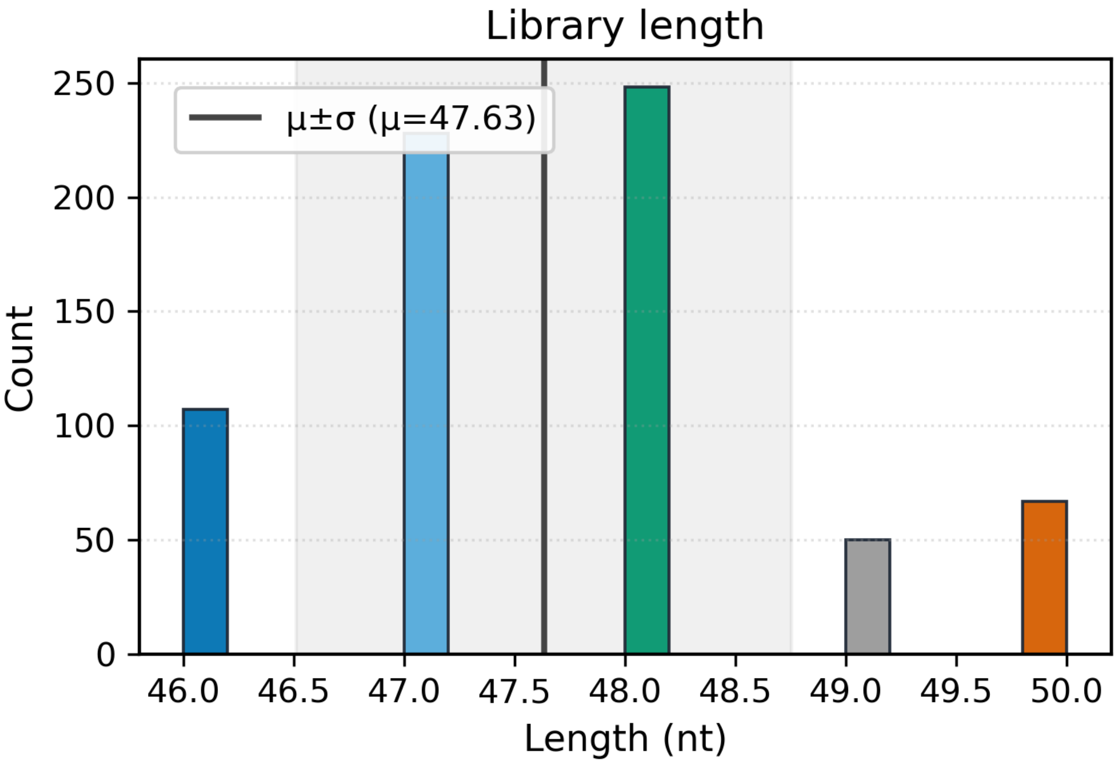

Figure 4.9 — Library length distribution (all sequences); mean ± SD overlay with μ ≈ 47.63 nt.

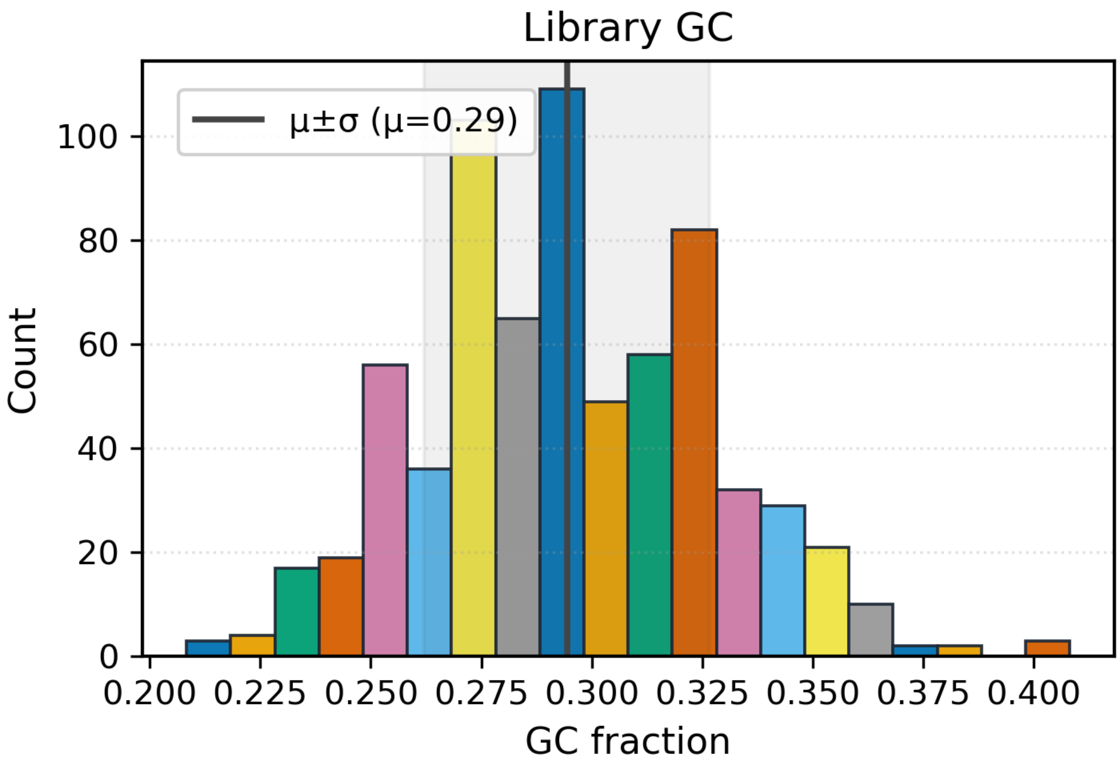

Figure 4.10 — Library GC fraction distribution (all sequences); mean ± SD overlay with μ ≈ 0.29.

4.5 Diversity of the shortlist

To mitigate correlated failures among near-identical sequences, we quantified sequence-level diversity using 1 − Jaccard on 3-mer sets and selected a subset via max–min greedy choice (Methods 3.3). The Top-20 distribution (Supplementary Fig. S3) has a median distance of 0.444 with an interquartile range of ~0.41–0.49, indicating a well-spread set. Diversity at this level is sufficient to sample multiple local sequence contexts around the same structural design, which improves the chance that at least some constructs will exhibit strong switching in vivo despite context-specific effects. [28]

4.6 Best predicted high-temperature thermoswitches

Table 4.2 — Top-20 shortlist

Rank	ID	Sequence	Length	RBS_start	RBS_end	extra_pairs	pGC	loop_len	p_pass	T50_pred	Slope70_pred

1	218	UUAAUUUAAAGC**AGGAGG**UGCUAAUAACUUUUUGCAUGUAUUAUAAAU	48	12	18	2	0.566	9	0.825	70.48	0.061
2	666	UUAUAUAUAAGAC**AGGAGG**CGCGAAGACUUUUUGUCAUGUUUAAUAUUU	49	13	19	3	0.541	8	0.757	71.12	0.047
3	19	UUAUUUUUUAGC**AGGAGG**GACAAGGCACUUUUUGCAUGAUAUUAAUAA	48	12	18	2	0.544	9	0.825	71.64	0.056
4	222	AAAAUAAAUAGC**AGGAGG**GUAUAGCGGCUUUUUGCAUGAAAAAAUAAU	48	12	18	2	0.522	9	0.775	71.90	0.049
5	457	UUUAUUAUAAGAC**AGGAGG**GGUAACGACUUUUUGUCAUGAUUUUAUAAA	49	13	19	3	0.45	8	0.761	72.43	0.052
6	199	AAUUAUAAUAGC**AGGAGG**UAACAUGCGCCUUUUGCAUGUUUAUAAUUU	48	12	18	2	0.553	9	0.811	71.24	0.048
7	629	UUUUAAUUAAGC**AGGAGG**UCAUCAGCGCUUUUUGCAUGAUAUUAAAAU	48	12	18	2	0.548	9	0.725	71.61	0.054
8	151	UAUAUAUAUAGAC**AGGAGG**UUUGGCAGGCUUUUUGUCAUGUAAAUUUUAA	50	13	19	3	0.58	9	0.746	72.00	0.056
9	30	AAAUUUUUUAGC**AGGAGG**CUCGAAGACUUUUUGCAUGAAAUUUAUAA	47	12	18	2	0.542	8	0.811	71.23	0.057
10	189	AUUUAAAAAAGC**AGGAGG**UAUGAGACGCUUUUUGCAUGUUUUUUAAAU	48	12	18	2	0.58	9	0.762	71.95	0.055
11	40	UAAUUAAAUAGC**AGGAGG**UUGGUGCAACUUUUUGCAUGAUUAUAUAAU	48	12	18	2	0.45	9	0.757	71.12	0.061

12	611	AAAAAUUUAAGC**AGGAGG**GGGCGAUACUUUUUGCAUGAAAAUUUUUU	47	12	18	2	0.483	8	0.825	71.79	0.056
13	213	UAUUAAAAAAGC**AGGAGG**GAGUGAACUUUUUGCAUGUUAUUAUUAA	46	12	18	2	0.45	7	0.725	71.89	0.048
14	626	UAAUAUUUUAGC**AGGAGG**UGUGCAGGCUUUUUGCAUGAAUAAUUUAU	47	12	18	2	0.58	8	0.775	70.69	0.056
15	446	AAUUUAUUUUGC**AGGAGG**GCAAGUACUUUUUGCAUGAAUUUUUAAU	46	12	18	2	0.45	7	0.762	72.09	0.048
16	560	UAUAUAUAUAGC**AGGAGG**GCUAAAAGACUUUUUGCAUGUUAUUAUUAU	48	12	18	2	0.555	9	0.825	71.55	0.057
17	676	UAUAUAAUAAGAC**AGGAGG**GAGCGGGCACUUUUUGUCAUGUAAUUUUUAU	50	13	19	3	0.45	9	0.825	72.30	0.054
18	35	AUUUAAUAUAGC**AGGAGG**CCGCGGAACUUUCUGCAUGUUUUAAAUAU	47	12	18	2	0.499	8	0.651	71.72	0.046
19	119	AUUAUUAUAAGC**AGGAGG**UACGAGGCACUUUUUGCAUGUUAUUUAUAA	48	12	18	2	0.551	9	0.669	72.01	0.056
20	436	AUAUUUUAUAGC**AGGAGG**CACUGCAGACCUUUUGCAUGUAUAUAUAUA	48	12	18	2	0.472	9	0.651	70.20	0.057

The final deliverable is an order-ready, diversified Top-20 list (ml_topK_diverse_clean.csv). Order-ready here means each sequence passes synthesis QC (GC fraction within 0.25–0.70 and no ≥7-nt homopolymer runs) and meets the design guard bands (predicted T50 ∈ [69.0, 72.5] °C and slope@70 °C ≥ 0.040 per °C); annotated RBS start–end indices are also provided for cloning/validation. Each entry includes the sequence, RBS start–end indices, scaffold knobs (extra_pairs, mid-stem pGC, loop_len), and predicted metrics (T50 and slope@70 °C, plus guard components). Based on the enrichment statistics above, we expect a meaningful fraction of true hits under our pass criterion. [3], [5], [11].

5. Discussion & Conclusion

This work set out to engineer RNA thermoswitches that turn on near ~70 °C with low leak and a steep transition, and to build a screening pipeline that can yield order-ready candidates without extensive experimental iteration. Combining physics-guided scoring, a constrained library generator, and a calibrated ML ranker produced a shortlist aligned with our aims; unlike fixed rules, the model learns which feature combinations matter and returns a probability of success, letting us sort big libraries faster with less compute and fewer lab tests. Across the analysis snapshot we observed high discrimination (AUC $\approx$ 0.95; AP $\approx$ 0.87), strong top-list enrichment (Precision@K = 1.00 at K = 10 and 20), stable cross-validation, and monotonic calibration sufficient for rank-based selection. Together, these results support the central claim: the pipeline can deliver credible Top-K candidates for a high-temperature use case in *Bacillus subtilis*. [1], [3], [12].

For the ~70 °C regime, our results support two points. First, our in-silico results predict low ribosome access at ≤~65 °C, a steep opening of the RBS around ~70 °C, and a sustained open state above that, i.e., little translation when cool, rapid turn-on near 70 °C, and strong translation when hot. This pattern is consistent with translation-level gating that could be triggered by short heat pulses. Second, the scaffold strategy-fixed SD RBS, GU-softened anti-RBS, and GC-clamped stems was sufficient to centre T50 within the target band while maintaining a high ON plateau and avoiding excessive leak, meaning a precise thermal window can be targeted by tuning a small number of interpretable knobs [5].

Several choices underpin this outcome. The physics-guided metrics correspond directly to the design goal in plain terms: how “off” the switch is at low temperature, how fully “on” it is at high temperature, where the turn-on midpoint sits (T50), and how abruptly it changes near ~70 °C (local slope). As these are computed from the thermodynamic ensemble rather than a single structure, they are robust and comparable across designs. For selection, we use a calibrated classifier to obtain well-behaved probabilities and we emphasise recall in evaluation and operating point (F_2 on the training PR curve; Precision@K/Recall@K for K = 10, 20, 50), while the shortlist itself is ranked by calibrated p(pass) (then proximity to 70 °C and predicted slope), not by a fixed threshold. Finally, QC + diversity mitigates practical risks: guard bands keep the predicted midpoint and steepness within bounds, and max–min 3-mer distance avoids near-duplicates, yielding a panel that is experimentally tractable and resilient to context-specific effects. [9], [11], [13], [19], [20], [26]–[28].

As this study is intentionally in silico, the principal constraints were computational and operational. Generating 100 optimised sequences required ~24 hours of end-to-end elapsed time on our workstation, reflecting (i) evaluation of RBS-openness over 50–85 °C at 2 °C increments using NUPACK ensemble probabilities and (ii) a multi-restart local search that iteratively scores nearby variants. With limited parallelisation and no caching of folding calls, throughput scales sublinearly with library size. We stabilised the pipeline by fixing a reproducible parameter set, logging seeds and CSV snapshots, and automating figure generation. Future work can reduce compute via batching, multi-process evaluation, memoisation across temperature points, and lightweight pruning; in parallel, experimental testing of a subset of candidates would close the loop. Since designs here were evaluated without a downstream "payload" (wider 5′-UTR/coding sequence), a next step is co-design or scoring in an extended mRNA context, acknowledging the added computational time. [1], [3], [4].

In conclusion, we deliver a decision-ready, reproducible pipeline for RNA thermoswitches that activate near ~70 °C, combining thermodynamic scoring (NUPACK-derived ensemble metrics), compact, interpretable features, calibrated probability ranking, and shortlist QC/diversification to yield candidates that are both plausible and order-ready. As a proof of principle for coupling thermodynamic modelling with machine learning, the approach is readily retargetable to other set-points or hosts by adjusting guard bands and modestly re-tuning anti-RBS softening. With a larger dataset, the ML component can increasingly act as a learned first-pass filter, narrowing the design space before full thermodynamic evaluation; feature analyses can highlight high-value regions to prioritise for compute and, ultimately, bench testing. This positions the workflow for iterative design–test–learn cycles that scale from in-silico screening to experimentally validated thermoswitches. [9], [14], [27], [28].

Acknowledgements

I acknowledge the use of OpenAI ChatGPT (GPT-5 Thinking) and OpenAI o3 as learning and productivity aids for Python and Nupack based code guidance and editorial clarity (June–August 2025). All outputs were checked, edited, and verified by the author.

Appendix — Glossary of Key Terms

AGGAGG (RBS motif): A common Shine–Dalgarno motif placed just before the start codon. It helps the ribosome bind the mRNA to start translation.

Anti-RBS: The RNA segment in the hairpin stem that is complementary to the RBS. When paired, it hides the RBS and blocks translation at low temperature.

Anti-Shine–Dalgarno (anti-SD): A sequence on the 16S rRNA in the small ribosomal subunit that base-pairs with the Shine–Dalgarno on mRNA to initiate translation.

AUC (area under curve): A single-number summary of a diagnostic curve. ROC-AUC measures separability across classes; PR-AUC focuses on performance on the positive class when positives are rare.

Average precision (AP): The area under the precision–recall curve, computed as the weighted average of precisions at different recall levels.

***Bacillus subtilis*:** A Gram-positive, spore-forming bacterium widely used in biotechnology; tolerant to stress and relevant to high-temperature workflows.

Base pair: Two RNA bases that pair by hydrogen bonds. Canonical pairs are G–C and A–U; G–U is a weaker wobble pair.

Calibration (probability): Making model scores match observed frequencies (e.g., designs scored 0.7 pass ~70% of the time). We use isotonic calibration.

Calibration curve / reliability plot: A plot comparing predicted probabilities to observed fractions in bins. A perfect model lies on the diagonal.

Class imbalance: When positive (pass) examples are much fewer than negatives. Metrics like PR-AUC, Precision@K, and Recall@K are preferred.

Closed_low (OFF closedness): The fraction of the RBS predicted to be paired at low temperature; computed as 1 − openness (higher = tighter OFF).

Cross-validation (5-fold stratified): Split data into five equal parts with the same class balance; train on four, test on one, and report mean±sd across folds.

Dataset snapshot (release): A frozen copy of the CSV files, parameters, and seeds used for the reported analysis so results can be reproduced.

Diversity filter (3-mer Jaccard): Ensures selected sequences are not near-duplicates by comparing sets of all 3-base substrings and favouring higher distances.

Ensemble (RNA folding): The collection of all possible RNA structures weighted by thermodynamic probabilities, not just a single “best” fold.

Expected hits: Sum of calibrated probabilities across a shortlist; an estimate of how many true passes we should see when we test them.

Feature (ML): A numeric description of a sequence (e.g., length, GC fraction, k-mer frequencies, positional one-hot around the RBS) used by the model.

F_2 score: A combined score that weights recall higher than precision ($\beta = 2$): $F\beta = (1+\beta^2)\cdot(P\cdot R)/(\beta^2\cdot P + R)$. Useful when missing positives is costly.

GC clamp: Using G–C pairs at the ends of a stem to stabilise the hairpin so it forms reliably at low temperature.

GC fraction: Fraction of G and C nucleotides in a sequence; higher GC generally stabilises RNA stems.

Gradient Boosting (classifier): An ML method that adds small decision trees to improve predictions step-by-step; used here with probability calibration.

GU wobble: A weaker, non-canonical G•U pair used to "soften" the hairpin stem so it melts near the target temperature.

Guard filters / guard bands: Pass/fail checks for shortlisted designs (e.g., T50 within 69–72.5 °C; slope@70 °C ≥ 0.04).

Hairpin (stem–loop): A common RNA structure: two complementary stretches form the stem; an unpaired region forms the loop.

Heat pulse / recovery: Intended use pattern—short exposure to ~70 °C followed by rapid cool-down so translation resumes while the mRNA remains open.

In silico: Work done by computer rather than in the laboratory (e.g., simulations, model training).

Isotonic calibration: A non-parametric method mapping raw model scores to calibrated probabilities while preserving rank order.

Jaccard distance: A measure of dissimilarity between two sets: $1 - (\text{intersection} \div \text{union})$. Used here on sets of k-mers.

K-mer (di- / tri-): All contiguous substrings of length k (e.g., 16 dinucleotides and 64 trinucleotides). Their frequencies summarise sequence composition.

Leak: Undesired translation in the supposed OFF state; assessed via low-temperature RBS openness (lower is better).

Library: A collection of candidate RNA designs generated by the computational pipeline.

Local search / hill-climb: An optimisation strategy: propose small edits to a candidate and keep changes that improve the score; repeat from several random restarts.

Machine-learning ranker: A model that orders sequences so the top of the list is enriched for likely passes; here a calibrated gradient-boosting classifier.

Melting / denaturation (RNA): Temperature-driven loss of base pairing in the hairpin stem that exposes the RBS (switch turns ON).

NUPACK: RNA folding software that computes base-pairing probabilities using nearest-neighbour thermodynamic rules.

One-hot encoding (positional): Representing each base at each position as a 4-vector (A,U,G,C), used to capture local patterns around the RBS.

Open_high / open_midhi: Predicted RBS openness when warm; open_midhi summarises just above the transition (~72–78 °C), open_high summarises the high-temperature plateau (≥80 °C).

Precision: Of the designs predicted positive, the fraction that are truly positive: TP/(TP+FP).

Precision@K: Among the top K ranked designs, the fraction that are true positives—indicates shortlist purity.

Probability of pass, p(pass): The calibrated model score for a design; can be read as an approximate hit-rate.

PR-AUC (precision–recall AUC): Area under the precision–recall curve; preferred when positives are rare.

RBS (ribosome-binding site): The mRNA region that recruits the ribosome for translation initiation, often with an AGGAGG-like Shine–Dalgarno motif.

RBS openness: Predicted fraction of bases across the RBS that are unpaired at a given temperature (averaged over the SD window).

Recall (sensitivity / true-positive rate): Of all true positives, the fraction retrieved: TP/(TP+FN).

Recall@K: In the top K designs, the fraction of all true positives recovered: TP_in_topK ÷ total_positives.

Release (analysis) snapshot: The set of files, parameters, and seeds frozen for the reported results so others can reproduce the figures.

Reliability (calibration): Agreement between predicted probabilities and observed frequencies.

Ribosome: The cell’s protein-synthesising complex; translation begins when it binds near the RBS.

ROC curve: Receiver-Operating-Characteristic curve; shows the trade-off between sensitivity and false-positive rate across thresholds.

ROC-AUC: Area under the ROC curve; higher indicates better separability.

SD (Shine–Dalgarno) sequence: Purine-rich sequence upstream of the start codon that base-pairs to the anti-SD on 16S rRNA to recruit the ribosome.

Seed (random): Initial value for a pseudorandom number generator so runs are repeatable.

Shortlist: The final set of designs proposed for synthesis/testing after ranking, guard bands, QC, and diversity filters.

Slope@70 °C: The local rate of change of RBS openness with temperature around 70 °C (per °C); higher = sharper ON transition.

Spec-pass definition: The logical rule that marks a design as a "positive" (e.g., thresholds on closed_low, open_midhi, open_high, T50, slope@70 °C).

Stem: The double-stranded part of a hairpin formed by base pairs between complementary segments.

T50: Temperature at which predicted RBS openness reaches 0.5 (50%); an intuitive switching point.

Thermodynamic nearest-neighbour model: Standard model in RNA folding where each base pair's contribution depends on its neighbour; parameters derive from melting experiments.

Thermostable reporter/enzyme: A protein retaining activity at elevated temperatures; useful for readouts following short heat pulses.

Top-K list: The K highest-ranked designs; we report Precision@K and Recall@K for K ∈ {10, 20, 50}.

Translation: The process by which the ribosome reads mRNA to build a protein.

Unpaired probability: For each nucleotide position, the probability of being unpaired in the ensemble; equals the diagonal of NUPACK's pairs-matrix.

Window around the RBS: The fixed region (e.g., ±10 upstream, +20 downstream) used to compute positional one-hot features for ML.

References

[1] B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell*, 6th ed. New York, NY, USA: Garland Science, 2014. ISBN: 978-0815344322. doi:10.1201/9781315735368. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9781315735368

[2] F. Narberhaus, T. Waldminghaus, and S. Chowdhury, "RNA thermometers," *FEMS Microbiology Reviews*, vol. 30, no. 1, pp. 3–16, 2006. doi:10.1111/j.1574-6976.2005.004.x. [Online]. Available: https://academic.oup.com/femsre/article/30/1/3/2367535

[3] J. Kortmann and F. Narberhaus, "Bacterial RNA thermometers: Molecular zippers and switches," *Nature Reviews Microbiology*, vol. 10, no. 4, pp. 255–265, 2012. doi:10.1038/nrmicro2730. [Online]. Available: https://www.nature.com/articles/nrmicro2730

[4] J.-D. Wen, S.-T. Kuo, and H.-H. D. Chou, "The diversity of Shine–Dalgarno sequences sheds light on the evolution of translation initiation," *RNA Biology*, vol. 18, no. 11, pp. 1489–1500, 2021. doi:10.1080/15476286.2020.1861406. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/15476286.2020.1861406

[5] J. Kortmann, S. Sczodrok, J. Rinnenthal, H. Schwalbe, and F. Narberhaus, "Translation on demand by a simple RNA-based thermosensor," *Nucleic Acids Research*, vol. 39, no. 7, pp. 2855–2868, 2011. doi:10.1093/nar/gkq1252. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC3077793/

[6] Y. Su, C. Liu, H. Fang, and D. Zhang, "**Bacillus subtilis**: a universal cell factory for industry, agriculture, biomaterials and medicine," *Microbial Cell Factories*, vol. 19, p. 173, 2020. doi:10.1186/s12934-020-01436-8. [Online]. Available: https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-020-01436-8

[7] D. E. Cameron, C. J. Bashor, and J. J. Collins, "A brief history of synthetic biology," *Nature Reviews Microbiology*, vol. 12, pp. 381–390, 2014. doi:10.1038/nrmicro3239. [Online]. Available: https://www.nature.com/articles/nrmicro3239

[8] M. Y. Chan, P. Zhang, and M. R. Sorokina, "RNA Thermometer: A Review of Regulatory Mechanisms and Applications," *Applied Sciences*, vol. 11, no. 10, p. 4532, 2021. doi:10.3390/app11104532. [Online]. Available: https://doi.org/10.3390/app11104532

[9] J. N. Zadeh, C. D. Steenberg, J. S. Bois, B. R. Wolfe, M. B. Pierce, A. R. Khan, R. M. Dirks, and N. A. Pierce, "NUPACK: Analysis and design of nucleic acid systems," *Journal of Computational Chemistry*, vol. 32, no. 1, pp. 170–173, 2011. doi:10.1002/jcc.21596. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/jcc.21596

[10] R. M. Dirks and N. A. Pierce, “A partition function algorithm for nucleic acid secondary structure including pseudoknots,” *Journal of Computational Chemistry*, vol. 24, no. 13, pp. 1664–1677, 2003. doi:10.1002/jcc.10396. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/jcc.10396

[11] R. M. Dirks, J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce, “Thermodynamic analysis of interacting nucleic acid strands,” *SIAM Review*, vol. 49, no. 1, pp. 65–88, 2007. doi:10.1137/060651100. [Online]. Available: https://epubs.siam.org/doi/10.1137/060651100

[12] J. S. Reuter and D. H. Mathews, “RNAstructure: Software for RNA secondary structure prediction and analysis,” *BMC Bioinformatics*, vol. 11, p. 129, 2010. doi:10.1186/1471-2105-11-129. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-

[13] *NUPACK 4 User Guide*. California Institute of Technology. [Online]. Available: https://docs.nupack.org/

[14] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://jmlr.org/papers/v12/pedregosa11a.html

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009. doi:10.1007/978-0-387-84858-7. [Online]. Available: https://link.springer.com/book/10.1007/978-0-387-84858-7

[16] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001. doi:10.1214/aos/1013203451. [Online]. Available: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation--a-gradient-boosting-machine/10.1214/aos/1013203451.full

[17] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLOS ONE*, vol. 10, no. 3, e0118432, 2015. doi:10.1371/journal.pone.0118432. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432

[18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. ISBN: 978-0521865715. doi:10.1017/CBO9780511809071. [Online]. Available: https://www.cambridge.org/highereducation/books/introduction-to-information-retrieval/669D108D20F556C5C30957D63B5AB65C

[19] A. Niculescu-Mizil and R. Caruana, “Predicting Good Probabilities With Supervised Learning,” in *Proc. ICML*, 2005, pp. 625–632. doi:10.1145/1102351.1102430. [Online]. Available: https://dl.acm.org/doi/10.1145/1102351.1102430

[20] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in *Proc. KDD*, 2002, pp. 694–699. doi:10.1145/775047.775151. [Online]. Available: https://dl.acm.org/doi/10.1145/775047.775151

[21] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. doi:10.1016/j.patrec.2005.10.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

[22] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proc. ICML*, 2006, pp. 233–240. doi:10.1145/1143844.1143874. [Online]. Available: https://dl.acm.org/doi/10.1145/1143844.1143874

[23] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proc. ICML*, 2017, pp. 1321–1330. [Online]. Available: https://proceedings.mlr.press/v70/guo17a.html

[24] R. Kohavi, “A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection,” in *Proc. IJCAI*, 1995, pp. 1137–1143. [Online]. Available: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf

[25] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1993. doi:10.1201/9780429246593. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.1201/9780429246593/introduction-bootstrap-bradley-efron-tibshirani

[26] Twist Bioscience, “What are the sequence acceptance criteria for Express Genes?,” FAQ. [Online]. Available: https://www.twistbioscience.com/faq/gene-synthesis

[27] Integrated DNA Technologies (IDT), “What types of sequence motifs should be avoided when ordering gBlocks Gene Fragments?,” FAQ. [Online]. Available: https://www.idtdna.com/pages/support/faqs/what-types-of-sequence-motifs-should-be-avoided-when-ordering-gblocks-gene-fragments

[28] R. W. Hamming, “Error detecting and error correcting codes,” *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950. doi:10.1002/j.1538-7305.1950.tb00463.x. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/j.1538-7305.1950.tb00463.