PrimerCast: Predictive Modeling of PCR Amplification with an AI-Ready Experimental Dataset

S. Chan Baek
Broad Institute of MIT and Harvard
Cambridge, MA 02142
baekseun@broadinstitute.org

Kenneth Bryan Hsu
Department of Systems Biology
Harvard University
Cambridge, MA 02138
khsu@broadinstitute.org

Yasha Ektefaie
Broad Institute of MIT and Harvard
Cambridge, MA 02142
yektefai@broadinstitute.org

Sameed Siddiqui
Lyra Labs
Cambridge, MA 02139
sameed@mit.edu

Krithik Ramesh
Lyra Labs
Cambridge, MA 02139
krithik@lyralabs.ai

Pardis Sabeti
Broad Institute of MIT and Harvard
Harvard University
Cambridge, MA 02138
pardis@broadinstitute.org

Abstract

Polymerase Chain Reaction (PCR) is the foundational technology for detecting pathogens from their genomic sequences. Rapidly selecting effective primers is critical in outbreak settings, yet predicting whether a candidate primer pair will amplify a target sequence depends on a complex mix of sequence–sequence interactions, nonlinear thermodynamics, and cross-genome variability. As a result, existing tools rely on hand-crafted heuristics and qualitative rules for design. From these candidates, high-performing primers are typically identified only through slow, trial-and-error laboratory testing. To overcome this bottleneck, we introduce the first AI-ready, large-scale, experimentally measured dataset of PCR amplification outcomes. The dataset comprises 50,760 unique primer–target reactions spanning 141 viral detection targets and 360 primer sets, with quantitative labels of amplification efficiency at unprecedented scale. Leveraging this resource, we train PRIMERCAST, a predictive model that forecasts amplification success and efficiency. PRIMERCAST consistently outperforms both widely used heuristic tools and prior baselines, enabling reliable, data-driven primer evaluation. By framing primer performance as a predictive modeling challenge and releasing an AI-ready dataset, we provide a foundation for faster, more reliable diagnostic test development and open the door for applying machine learning to broader nucleic acid technologies.

1 Introduction

The rapid development of highly effective diagnostics against an emerging pathogen is critical for timely public health responses. PCR diagnostics is the gold standard for viral surveillance because it is simple, fast, sensitive, widely accessible worldwide, and readily designed once a viral genome has been sequenced. The core of PCR assay development lies in designing two short oligonucleotide

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

primers, each approximately 20 DNA base pairs (bp) in length, that anneal to the target sequences 60-200 bp apart and catalyze amplification through appropriate thermal cycling.

Primer design has traditionally been relatively straightforward, relying on a handful of heuristic rules applied to a single target sequence to select perfectly matching subsequences. By contrast, for diagnostic assays that inherently aim to capture all circulating variants of a viral species at the time of development, viral sequence diversity poses a major challenge. Viruses accumulate mutations rapidly, and sequence divergence across the variants can reach up to 27% [1, 2]. Consequently, primers that perfectly match all variants are very rare. While conventional approaches use an artificial representative sequence, this strategy depends heavily on the degree of predefined sequence divergence, and minor variants are inevitably neglected. A critical unmet need is the ability to evaluate the PCR activity of each candidate primer against any possible target sequence.

Another important consideration is maximizing assay activity and specificity. The number of possible perfectly matching primer pairs for a single target often reaches several hundred thousand, underscoring the immense design space. However, existing tools lack the precision to identify the most effective candidates [3]. Conventional designs adjust sensitivity and specificity by manually inspecting mismatches across variants and screening against off-targets (e.g. human transcripts), followed by slow and labor-intensive laboratory testing. This process is brittle at scale: during the SARS-CoV-2 pandemic, primer flaws triggered recalls [4] and led to large performance variability [5] across approved assays, underscoring the urgent need for automated, data-driven primer design [3].

Machine learning offers a principled route to this goal by learning sequence-to-amplification mappings and optimizing over a combinatorial candidate space. Progress, however, has been constrained by data: prior studies use small experimental datasets ($<$ 4k reactions) or rely on surrogate labels from hybridization-energy calculators rather than measured PCR outcomes [6, 7, 8].

Here we introduce, to our knowledge, the first large-scale AI-ready dataset, consisting of more than 50,000 distinct primer–target pairs. These combinations vary widely in the count, position, and type of mismatches, enabling systematic evaluation of PCR activity for any primer–target pair. Leveraging a microfluidic PCR platform [9], we obtained high-quality experimental data in quadruplicate. Using this dataset along with an external dataset of primers, we train PRIMERCAST which outperforms traditional approaches based on mismatch counting or binding energy in classifying PCR reaction outcomes. To demonstrate the utility of PRIMERCAST, we applied PRIMERCAST to ten representative blood-borne viruses. In nine out of ten cases, the AI-designed primers achieved equal or superior predicted performance than experimentally validated primers with respect to variant coverage, mean efficiency across variants, and off-target reactivity against other viral species or human genes. To further validate these results, we performed a wetlab PCR experiment using PRIMERCAST to design primers for SARS-CoV-2 and compare to those released by the CDCs of several countries. We find PRIMERCAST designed primers achieves the highest sensitivity and specificity. By reporting this large experimental dataset, benchmark, and model, we establish a reproducible foundation for AI-guided diagnostics that can cut time-to-assay in future outbreaks.

2 Related Work

Primer3 and Primer-BLAST The currently most widely used primer design tool is Primer3 [3]. Primer3 evaluates primers by the weighted sum of thermodynamic and sequence parameters [7, 10]. While this approach is straightforward, the preset parameters is largely arbitrary. Consequently, selected primers often show a discrepancy from experimental outcomes. For example, the SARS-CoV-2 primers used by many high-performance tests are scored poorly under the default settings of Primer3, yet they are employed in many of the top-performing PCR tests [11]. Primer-BLAST is a representative tool for testing off-target amplification by searching for sequences primers could potentially bind to [12]. However, the number of mismatches between primers and their targets does not directly indicate whether PCR will occur, nor its amplification efficiency [8]. Neither Primer3, Primer-BLAST, or any currently existing tools can quantitatively predict PCR efficiency.

Machine learning for primer design Several machine learning approaches have been proposed for primer design. Pérez-Romero et al. (2023) developed an evolutionary algorithm-based pipeline for SARS-CoV-2 primer design [13]. This does not score primer candidates qualitatively and relies on Primer3 to generate and score primer candidates. Wang et al. (2025) introduced a VAE-based

model for automated primer design, but the scoring of generated primers was again conducted using bioinformatics tools only [14]. Finally, Kayama et al. (2021) constructed an experimental PCR dataset which was relatively small (n = 3,906) and of limited quality, as a large fraction of the true set consisted of primer–target pairs with more than 10 mismatches, a range irrelevant in actual PCR experiments [6]. Using this dataset, they reported a classifier that achieved a sensitivity of 70%. In summary, no models currently exist which can directly predict qPCR amplification from primer and target sequences. All currently reported models are reliant on conventional in-silico prediction of primers. To our knowledge, PRIMERCAST is the first quantitatively predictive model to be trained and validated on a high-quality experimental dataset.

3 Problem Definition

Let $\mathcal{P}_t$ denote the set of candidate primer pairs for target sequence t. Each candidate $P_i \in \mathcal{P}_t$ is an ordered pair $P_i = (p_f^i, p_r^i)$, where $p_f^{(i)}$ and $p_r^{(i)}$ are the forward and reverse primer sequences, respectively. We train two complementary machine learning models: a classifier and a regressor. Both models take as input the target sequence t and a primer pair p_f^i and p_r^i, then predicts whether amplification occurs. The classifier predicts a binary label:

$$f_\theta(p_f^i, p_r^i, t) \ \in \ \{0, 1\} \tag{1}$$

where a 1 indicates that the primer pair successfully amplifies the target as measured from PCR fluorescence above baseline. The regressor f_θ predicts the amplification efficiency as a continuous value:

$$f_\theta^{(r)}(p_f, p_r, t) \ \in \ \mathbb{R}_{\geq 0} \tag{2}$$

where 0 corresponds to the baseline signal, and 1 corresponds to the efficiency of the CDC-designed SARS-CoV-2 primers, although values greater than 1 are possible.

3.1 Primer Evaluation Metrics

We next define four evaluation metrics for each primer pair P^i, derived from the outputs of the classifier and regressor introduced above. These metrics describe variant coverage, amplification efficiency, and specificity against non-target sequences.

Variant coverage Let $\mathcal{T}$ denote the set of n variant sequences reported during a specific time period. For each $t \in \mathcal{T}$, the classifier $f_\theta^{(c)}(p_f, p_r, t)$ outputs 1 if amplification is predicted and 0 otherwise. The variant coverage u_i is defined as:

$$u_i = \frac{1}{n} \sum_{t \in \mathcal{T}} f_\theta^{(c)}(p_f, p_r, t) \tag{3}$$

Mean efficiency Let $\mathcal{S}_i \subseteq \mathcal{T}$ denote the subset of variants predicted positive by the classifier. For each $t \in \mathcal{S}_i$, the regressor $f_\theta^{(r)}(p_f, p_r, t)$ outputs a score corresponding to predicted amplification efficiency. The mean efficiency v_i is defined as:

$$v_i = \frac{1}{|\mathcal{S}_i|} \sum_{t \in \mathcal{S}_i} f_\theta^{(r)}(p_f, p_r, t) \tag{4}$$

Cross-species reactivity Let $\mathcal{T}_{\text{path}}$ denote the set of n_{path} sequences from unwanted pathogens. For each $t \in \mathcal{T}_{\text{path}}$, the classifier predicts amplification as above. The cross-species reactivity o_i is defined as:

$$o_i = \frac{1}{n_{\text{path}}} \sum_{t \in \mathcal{T}_{\text{path}}} [f_\theta^{(c)}(p_f, p_r, t) + 1] \tag{5}$$

Host-gene reactivity Let $\mathcal{T}_{\text{host}}$ denote the set of n_{host} human transcript sequences. For each $t \in \mathcal{T}_{\text{host}}$, the classifier predicts amplification as above. The host-gene reactivity h_i is defined as:

$$h_i = \frac{1}{n_{\text{host}}} \sum_{t \in \mathcal{T}_{\text{host}}} [f_\theta^{(c)}(p_f, p_r, t) + 1] \tag{6}$$

4 Methods

4.1 AI-Ready PCR Dataset Design

To enable the design of primers that are both sensitive and specific, the dataset must capture intended primer–target interactions (perfect or near-perfect matches with short amplicons) as well as unintended interactions (partial matches with variable amplicon lengths). To this end, we constructed a synthetic, biologically relevant PCR library. A wild-type target sequence was generated as a mosaic of human viral sequences: over 3 million sequences (41 families, 106 genera, 409 species) were collected from NCBI Virus [15], representative coding sequences from each species were fragmented, and 280 pentamers were sampled to evenly cover the frequency distribution. These pentamers were shuffled and concatenated iteratively to obtain a 1.4 kb wild-type target with diverse primer-binding site features, such as GC content, melting temperature, and secondary structure. From this sequence, 140 variants were created by random mutagenesis, introducing 0–10 mismatches in primer-binding sites and yielding 141 total targets. In parallel, 360 primer pairs were designed against the wild-type sequence, with expected amplicon lengths ranging from 50 to 1,400 bp. Combined, this design produced 50,760 unique primer–target combinations for experimental testing.

Microfluidic PCR experiment. We obtained 141 target sequences through chemical synthesis from Twist Biosciences and 360 primer pairs from Integrated DNA Technologies. Each PCR reaction consisted of a target sequence, a primer pair, and a master mix containing polymerase, nucleotides, and fluorescent EvaGreen dye that binds to amplified DNA. During each reaction cycle, fluorescence signal was measured to quantify the amount of amplified DNA present. For large-scale data collection, we used the Biomark X microfluidic PCR platform. Each microfluidic chip accommodates 96 targets and 96 primer pairs each. Once a run starts, the reagents are automatically combined into 96×96 microchambers. The instrument then performs up to 40 thermal cycles and records the fluorescence signals from each chamber using a high-resolution imaging system. We utilized this system to perform 221,184 individual PCR reactions, testing all 50,760 unique primer–target combinations in quadruplicate.

Data preprocessing. Raw fluorescence signals were normalized to account for well-to-well variations in reaction volume, background noise, and spatial biases across the microfluidic chip. Specifically, let E_i denote the raw EvaGreen fluorescence (reporting the amount of amplified DNA) at cycle i, and R_i denote the raw ROX fluorescence (serving as a passive reference dye that reflects reaction volume). B_E and B_R denote background signals for EvaGreen and ROX, respectively. The first-stage normalized signal at cycle i was computed as:

$$S_i^{(1)} = \frac{E_i/B_E}{R_i/B_R} \tag{7}$$

To correct for position-dependent variability, a second-stage normalization was applied. We set the first-cycle signal to 0, and scaled values linearly such that the 95th percentile of $S_i^{(1)}$ at the end-point across all reactions was mapped to 1. The resulting normalized signal $S_i^{(2)}$ therefore ranges between 0 and 1 and is comparable across reactions.

Activity score definition. In qPCR, amplification is typically quantified by the cycle threshold (Ct), the cycle at which the fluorescence signal crosses a predefined threshold (0.1 in our case). Because Ct is sensitive to target input amount and low-level contamination, we convert Ct values into a normalized activity score using two reference measurements per target: (i) a no-target control (NTC) for each primer pair to capture primer-specific contamination/background, and (ii) amplification of the same target by a universal primer pair to account for target amount variability. For each target t and primer pair p, let

$$\begin{aligned} &\mathrm{Ct}_{t,p} \text{ be the Ct of } p \text{ on } t, \\ &\mathrm{Ct}_{p,0} \text{ be the Ct of } p \text{ with no target, and} \\ &\mathrm{Ct}_{t,u} \text{ be the Ct of a universal primer on } t. \end{aligned}$$

If amplification is not detected within 40 cycles, we set $\mathrm{Ct} = 40$. Because lower Ct indicates stronger amplification, we define the activity score by linearly mapping $\mathrm{Ct}_{0,p} \mapsto 0$ and $\mathrm{Ct}_{t,u} \mapsto 1$:

$$\mathrm{Act}(t,p) \;=\; \frac{\mathrm{Ct}_{0,p} - \mathrm{Ct}_{t,p}}{\mathrm{Ct}_{0,p} - \mathrm{Ct}_{t,u}}. \tag{8}$$

By construction, $\mathrm{Act}(t,p) = 0$ matches the NTC baseline of the primer pair, and $\mathrm{Act}(t,p) = 1$ matches the universal-primer amplification on the same target. We clip $\mathrm{Act}(t,p)$ to a lower bound of 0 for downstream analyses. For robustness, all Ct values are aggregated across four technical replicates using the average. 235 reactions (0.5%) were excluded because the Ct in the no-target control was below 36 or if the standard deviation of Cts across replicates exceeded 0.6. Remaining 50,525 data points were used for subsequent analyses.

4.2 Model training

Data splitting. To prevent information leakage across splits, we partitioned the dataset at the level of primer pairs rather than individual primer–target combinations. Because target sequences were generated through mutagenesis, primers that appear in multiple splits would inevitably share highly similar target sites. Thus, all 360 primer pairs were first divided into training, validation, and test sets in a 70:14:16 ratio, and the corresponding primer–target reactions were assigned accordingly. The test set shares no forward or reverse primers with the other sets, while only two reverse primers are shared between the training and validation sets.

Sequence representations. Because forward and reverse primers anneal at variable distances, the input must accommodate the variable inter-primer region. We compared three representations: (i) *Features-only*, a vector of classical primer-design features (Table 3) [7]; (ii) *Full-Seq*, which encodes the entire 1.4 kb target sequence together with the primer pair; and (iii) *Core-Seq*, which encodes only the primer-binding site sequences, augmented with a small set of scalar features. The Full-Seq and Core-Seq settings use the same tokenization: nucleotides are one-hot encoded over $\{A, C, G, T, \text{mask}\}$, yielding a $5 \times n$ tensor per sequence, where the *mask* channel marks padding and unbound positions. We form the final sequence input by concatenating the primer tensor and the target tensor along the channel dimension to obtain a $10 \times n$ matrix. In Core-Seq, we additionally append 12 handcrafted features—including amplicon length and melting temperature (Table 3)—motivated by the facts that (i) initiation at the binding sites is often rate-limiting, and (ii) elongation time is largely determined by amplicon length rather than detailed internal sequence content. Empirically, Core-Seq matched or exceeded Full-Seq predictive performance on both classification and regression tasks while reducing training and inference time by more than an order of magnitude (Tables 4, 5); we therefore use Core-Seq as the default representation in subsequent experiments.

Hurdle formulation. Activity scores are strongly bimodal, reflecting the all-or-nothing nature of PCR. We therefore adopt a two-stage (hurdle) setup: (i) a classifier $f_\theta^{(c)}$ predicts the probability of amplification, and (ii) a regressor $f_\theta^{(r)}$ predicts normalized efficiency for amplified pairs. The classifier is trained on labels $\mathbf{1}\{\mathrm{Act}(t,p) > 0\}$; the regressor is trained on pairs with $\mathrm{Act}(t,p) > 0$ using a MSE loss.

Model families and tuning protocol. Because this is, to our knowledge, the first large AI-ready PCR dataset, we benchmarked inductive biases rather than committing to a single architecture. We evaluate three families: *(a) Tabular baselines* on handcrafted features (Table 3); logistic/linear models [16], tree ensembles [17, 18], MLPs [19], SVMs [20], and k-NNs [16]) [21]; *(b) Sequence encoders* on Full-Seq and Core-Seq inputs (CNN [22], LSTM [23], Transformer [24], and state-space model Lyra [25]); and *(c) Late-fusion hybrids* that combine a sequence encoder with an MLP on handcrafted features. All models use the same sequence tokenization and Core/Full-Seq representations; hyperparameters are selected by exhaustive search on a held-out validation set with early stopping, and we report test metrics at the best validation checkpoint. Full grids and settings are provided in Appendix 8. We define PRIMERCAST as the best performing models.

4.3 In-silico validation of PRIMERCAST for primer design for blood-borne viruses

We applied PRIMERCAST to ten representative blood-borne viruses: CCHFV, CHIKV, EBOV, HCV, LASV, mpox-Ia/Ib, mpox-IIb, WNV, YFV, and ZIKV. For each virus, 98–2000 full-length genome sequences published since 2022 were downloaded from NCBI Virus (Table 6) [15]. When more than 2,000 variant sequences were available, a random subset of 2,000 was retained.

From each sequence, we extracted all possible 20-nt candidate primers using a 5-nt sliding window for both forward and reverse directions. Candidate primers with GC content $\geq 60\%$ or self-dimerization

stability $\Delta G < -6$ kcal/mol were excluded. Remaining primers were mapped to their target genomes using Bowtie2 [26] to obtain approximate binding sites, and all forward–reverse pairs separated by 75–200 bp were enumerated as candidate primer pairs.

For each candidate primer pair, we first computed the variant coverage u and mean predicted efficiency v as described in earlier. The top 100 primer pairs for each virus were selected according to the geometric mean of u and v. These primers were further evaluated for specificity by mapping against other blood-borne virus genomes and the human transcriptome, from which the off-target rates o and h were computed. Finally, the optimal primer pairs were identified by maximizing the geometric mean of u, v, o, and h.

4.4 Experimental validation of PRIMERCAST for SARS-CoV-2 primer generation

We conducted wetlab validation experiments to assess the sensitivity and specificity of PRIMERCAST candidate primers against standard primers used by the CDCs (Centers for Disease Control) from the US, China, Hong Kong, Japan, Korea, and Thailand [27, 28, 29, 30, 31, 32]. We conducted a standard RT-qPCR (reverse transcriptase) experiment evaluating the amplification efficiency of these primers to viral RNA targets (Supplementary Methods A.2) [33]. We measure for off-target activity by including human transcript only reactions. From these amplification curves, we derived quantitative comparisons of on-target and off-target sensitivities of manually designed and optimized primer sets as well as those of PRIMERCAST.

5 Experiments

5.1 Validation of PCR prediction models

Datasets We evaluated our models on two datasets. (1) We used the held-out test set from our synthetic dataset comprising 50,525 primer–target reactions spanning a broad space of primer features and primer–target interaction modes (Sections 4.1 and 4.2). The test set contained 5,854 true pairs and 2,324 false pairs (PRIMERCAST). (2) For external validation, we constructed a benchmark dataset from Origene PCR primers and targets, supplemented with synthetic variants. This dataset contained 33,860 true pairs and 38,168 false pairs (ORIGENE) [34](Supplementary Materials A.1).

Baselines We compared our models against two conventional baselines used in primer design. The first baseline uses the number of mismatches between primer and target sequences as a predictor of PCR outcome [35], where increasing mismatches leads to poorer outcomes. The second baseline relies on the predicted binding free energy (ΔG) of the primer–target duplex, a thermodynamic criterion used by many primer design tools [7]. Both baselines were evaluated in classification and regression settings by directly using mismatch counts or binding energies as continuous predictors.

Evaluation Metrics For classification tasks, we report AUROC and AUPRC. [36] For regression tasks, we report Spearman correlation coefficient ($|\rho|$) [37] and normalized mean square error (NMSE) [16].

Results We found that the Lyra+MLP model had best performance both in classification and regression. These models make up PRIMERCAST. PRIMERCAST consistently outperformed conventional baselines in both classifying amplifying reactions and predicting PCR efficiency (Table 1). On the synthetic dataset (PRIMERCAST), our classifier achieved an AUROC of 0.930, compared with 0.807 for mismatch count and 0.855 for free energy-based predictions. AUPRC also showed a similar trend (0.965 vs. 0.538–0.579). On the external dataset (ORIGENE), our classifier likewise surpassed both baselines (AUROC 0.950 vs. 0.743–0.809; AUPRC 0.910 vs. 0.610–0.680), demonstrating its generalizability. Regression could only be evaluated on PRIMERCAST since external datasets do not provide quantitative activity scores. Our regressor achieved a markedly higher correlation ($|\rho|$ = 0.842) and lower error (NMSE 0.396) than baselines ($|\rho|$ = 0.533–0.565; NMSE 5.64–24.6).

5.2 Benchmarking PRIMERCAST against existing primers for blood-borne viruses

Datasets We evaluated PRIMERCAST on ten representative blood-borne viruses: CCHFV, CHIKV, EBOV, HCV, LASV, mpox-Ia/Ib, mpox-IIb, WNV, YFV, and ZIKV. For each virus, we collected

Table 1: Performance of baselines and PRIMERCAST on the PRIMERCAST and ORIGENE datasets

	Classifier				Regressor	
	PRIMERCAST		ORIGENE		PRIMERCAST	
	AUROC	AUPRC	AUROC	AUPRC	$\lvert\rho\rvert$	NMSE
Mismatch count	0.807	0.579	0.743	0.610	0.533	5.64
Free energy	0.855	0.538	0.809	0.680	0.565	24.6
PRIMERCAST	**0.930**	**0.965**	**0.950**	**0.910**	**0.842**	**0.396**

full-length sequences published since 2022 from NCBI Virus at a maximum of 2000 variants (Table 6) [15].

Baselines We benchmark PRIMERCAST against primers produced by convention: (i) derive consensus sequences from circulating genomes; (ii) generate and score candidates with Primer3; (iii) screen off-targets with Primer-BLAST; and (iv) select a small panel via laboratory screening. The comparison set comprises primers currently in use that were designed with this conventional workflow.

Evaluation Metrics Performance was assessed using the four metrics defined in Section 3: variant coverage (u), mean predicted amplification efficiency across variants (v), cross-species reactivity (o), and human-gene reactivity (h). Coverage was the fraction of variants classified positive. Mean efficiency was the average regressor score across reactive variants. Cross-species and human-gene reactivity were the fractions of non-target pathogen sequences and human transcripts, respectively, predicted to be reactive.

Results In nine out of ten cases, the AI-designed primers achieved equal or superior performance across all evaluation metrics, including variant coverage, mean amplification efficiency, and off-target reactivity against non-target viruses and human transcripts (Table 2). For mpox-Ia/Ib and mpox-IIb, the top-ranked AI designs exactly matched the primers selected manually, indicating concordance between computational and experimental approaches. For highly diverse viruses such as LASV, manual designs performed poorly, likely because consensus sequences, while representing an average of variants, differ substantially from any actual sequence.

Table 2: Performance of AI-designed (PRIMERCAST) and manually designed primers

	Coverage (%)		Mean efficiency		Cross reactivity (%)		Host reactivity (%)	
	PRIMERCAST	Manual	PRIMERCAST	Manual	PRIMERCAST	Manual	PRIMERCAST	Manual
CCHFV	100	100	**1.07**	0.63	0.00	0.00	0.00	0.00
CHIK	**98.6**	96.0	**1.09**	1.03	**0.00**	0.14	**0.00**	0.06
EBOV	97.0	97.0	**1.11**	1.10	0.00	0.00	**0.00**	0.01
HCV	**98.9**	98.4	1.00	**1.04**	0.00	0.00	**0.00**	0.05
LASV	**91.9**	90.0	**0.98**	0.94	**0.00**	0.02	0.00	0.00
mpox-Ia/Ib	94.2	94.2	1.04	1.04	0.00	0.00	0.01	0.01
mpox-IIb	97.2	97.2	1.07	1.07	0.02	0.02	0.06	0.06
WNV	**99.0**	98.6	**1.08**	1.05	**0.00**	0.02	**0.00**	0.01
YFV	**96.8**	93.3	**1.09**	0.42	**0.00**	4.57	**0.00**	0.45
ZIKA	**93.9**	92.4	**1.09**	1.04	0.00	0.00	**0.00**	0.03

5.3 Experimental validation of PRIMERCAST performance against current SARS-CoV-2 primer sets

qPCR setup To assess sensitivity and specificity, PRIMERCAST primer and published primers by various CDCs were subject to RT-qPCR experiment in the presence of SARS-CoV-2 and human lung total RNAs, respectively.

Evaluation Metrics We evaluated the activity of a primer via cycle threshold (Ct), the cycle at which the fluorescence signal crosses a predefined threshold. It is inversely proportional to the sensitivity. If amplification is not detected, we set Ct as 41.

Results We observed that PRIMERCAST primers (designated P1-P7) were broadly comparable or exceeded the performance of existing CDC primers. In terms of off-target detection against human transcripts, all but one primer had Ct higher than 40, indicating no detectable off-target amplification. In terms of on-target detection against SARS-CoV-2 RNA, three primer sets had Ct values matching or lower than that of the most sensitive currently used CDC primer set, validating a strong agreement between in-silico predictions and experimental performance.

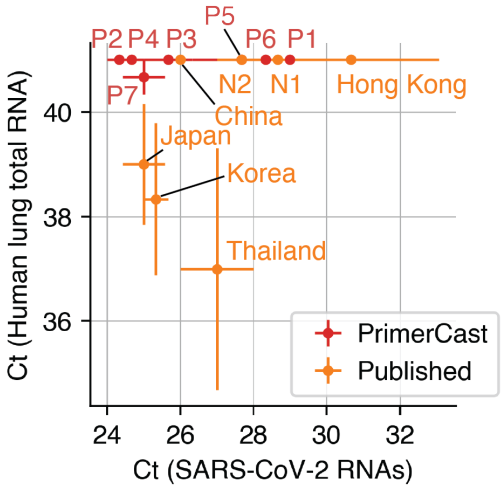

Figure 1: Wetlab validation of PRIMERCAST primer sets and comparison with the published primers. The x-axis and y-axis represent the on-target and off-target activities, respectively. Scatter points indicates the average over triplicates, and the error bars represent standard deviation.

6 Conclusion

PCR is a cornerstone technology for detecting and harnessing nucleic acids, with central roles in infectious disease testing. In outbreak response, the speed of setting up PCR tests and their performances directly translate into saved lives and reduced economic loss. However, the lack of rationally designed, large-scale PCR datasets has hindered the development of effective machine learning approaches. Here, we generated a high-quality dataset of 50,525 primer–target reactions through large-scale microfluidic PCR. Leveraging this dataset, we trained both classification and regression models to predict PCR outcomes, which consistently outperformed conventional heuristic methods. Building on these models, we developed PRIMERCAST, an AI-based model for primer design. In contrast to traditional approaches that rely on indirect proxies, PRIMERCAST directly and quantitatively evaluates PCR activity, exhaustively screens candidate primers across variants and potential off-targets, and identifies globally optimal designs. Applied to blood-borne viruses, PRIMERCAST produced primers with sensitivity and specificity equal to or exceeding those of primers that are experimentally validated and currently used in clinical diagnostics. In addition, we validate the performance predicted in-silico via wetlab experiments, showing strong prediction generalization to real-life performance. PRIMERCAST reduces the time scale of primer design from weeks to hours. The dataset reported here will provide a broadly useful resource for developing machine learning applications in diverse nucleic acid technologies, such as quantification, multiplexed testing, and viral genomics.

Competing Interests

The authors declare no competing interests.

Code Availability Statement

Our code is available at this anonymized Github link.

References

[1] GISAID Data Science Initiative.

[2] Michael D. Bowen, Pierre E. Rollin, Thomas G. Ksiazek, Heather L. Hustad, Daniel G. Bausch, Austin H. Demby, Mary D. Bajani, Clarence J. Peters, and Stuart T. Nichol. Genetic diversity among lassa virus strains. *Journal of Virology*, 74(15):6992–7004, Aug 2000.

[3] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C. Faircloth, Maido Remm, and Steven G. Rozen. Primer3—new capabilities and interfaces. *Nucleic Acids Research*, 40(15), Jun 2012.

[4] Justin S. Lee, Jason M. Goldstein, Jonathan L. Moon, Owen Herzegh, Dennis A. Bagarozzi, M. Steven Oberste, Heather Hughes, Kanwar Bedi, Dorothie Gerard, Brenique Cameron, and et al. Analysis of the initial lot of the cdc 2019-novel coronavirus (2019-ncov) real-time rt-pcr diagnostic panel. *PLOS ONE*, 16(12), Dec 2021.

[5] Matthew J. MacKay, Anna C. Hooker, Ebrahim Afshinnekoo, Marc Salit, Jason Kelly, Jonathan V. Feldstein, Nick Haft, Doug Schenkel, Subhalaxmi Nambi, Yizhi Cai, and et al. The covid-19 xprize and the need for scalable, fast, and widespread testing. *Nature Biotechnology*, 38(9):1021–1024, Aug 2020.

[6] Kotetsu Kayama, Miyuki Kanno, Naoto Chisaki, Misaki Tanaka, Reika Yao, Kiwamu Hanazono, Gerry Amor Camer, and Daiji Endoh. Prediction of pcr amplification from primer and template sequences using recurrent neural network. *Scientific Reports*, 11(1), Apr 2021.

[7] Tobias Mann, Richard Humbert, Michael Dorschner, John Stamatoyannopoulos, and William Stafford Noble. A thermodynamic approach to pcr primer design. *Nucleic Acids Research*, 37(13), Jun 2009.

[8] Ke Huang, Jilei Zhang, Jing Li, Haixiang Qiu, Lanjing Wei, Yi Yang, and Chengming Wang. Exploring the impact of primer–template mismatches on pcr performance of dna polymerases varying in proofreading activity. *Genes*, 15(2):215, Feb 2024.

[9] Courtney P. Olwagen, Peter V. Adrian, and Shabir A. Madhi. Performance of the biomark hd real-time qpcr system (fluidigm) for the detection of nasopharyngeal bacterial pathogens and streptococcus pneumoniae typing. *Scientific Reports*, 9(1), Apr 2019.

[10] Jingwen Guo, David Starr, and Huazhang Guo. Classification and review of free pcr primer design software. *Bioinformatics*, 36(22–23):5263–5268, Oct 2020.

[11] Tomer Altman. Technical problems with existing cdc covid-19 primers, and an improved set of primers, Mar 2020.

[12] Jian Ye, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, and Thomas L Madden. Primer-blast: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13(1), Jun 2012.

[13] Carmina Angelica Perez-Romero, Lucero Mendoza-Maldonado, Alberto Tonda, Etienne Coz, Patrick Tabeling, Jessica Vanhomwegen, John MacSharry, Joanna Szafran, Lucina Bobadilla-Morales, Alfredo Corona-Rivera, and et al. An innovative ai-based primer design tool for precise and accurate detection of sars-cov-2 variants of concern. *Scientific Reports*, 13(1), Sep 2023.

[14] Hanyu Wang, Emmanuel K. Tsinda, Anthony J. Dunn, Francis Chikweto, and Alain B. Zemkoho. Primer c-vae: An interpretable deep learning primer design method to detect emerging virus variants. *arXiv.org*, Mar 2025.

[15] Eric W Sayers, Jeffrey Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Ryan Connor, Michael Feldgarden, Anna M Fine, Kathryn Funk, Jinna Hoffman, and et al. Database resources of the national center for biotechnology information in 2025. *Nucleic Acids Research*, 53(D1), Nov 2024.

[16] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[17] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[18] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[20] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Anna G. Green, Chang Ho Yoon, Michael L. Chen, Yasha Ektefaie, Mack Fina, Luca Freschi, Matthias I. Gröschel, Isaac Kohane, Andrew Beam, and Maha Farhat. A convolutional neural network highlights mutations relevant to antimicrobial resistance in mycobacterium tuberculosis. *Nature Communications*, 13(1), Jul 2022.

[23] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[25] Krithik Ramesh, Sameed M. Siddiqui, Albert Gu, Michael D. Mitzenmacher, and Pardis C. Sabeti. Lyra: An efficient and expressive subquadratic architecture for modeling biological sequences. *arXiv.org*, Mar 2025.

[26] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, Mar 2012.

[27] FDA. Accelerated emergency use authorization (eua) summary of the gene by gene sars-cov-2 detection test. Technical report.

[28] Hong Kong University.

[29] Peihua Niu, Roujian Lu, Li Zhao, Huijuan Wang, Baoying Huang, Fei Ye, Wenling Wang, and Wenjie Tan. Three novel real-time rt-pcr assays for detection of covid-19 virus, Jun 2020.

[30] Japan National Institute of Infectious Diseases.

[31] Thailand Ministry of Public Health.

[32] Joungha Won, Solji Lee, Myungsun Park, Tai Young Kim, Mingu Gordon Park, Byung Yoon Choi, Dongwan Kim, Hyeshik Chang, V. Narry Kim, and C. Justin Lee. Development of a laboratory-safe and low-cost detection protocol for sars-cov-2 of the coronavirus disease 2019 (covid-19). *Experimental Neurobiology*, 29(2):107–119, Apr 2020.

[33] Stephen A. Bustin, Vladimir Benes, Jeremy A. Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Rolf Mueller, Tania Nolan, Michael W. Pfaffl, Gavin L. Shipley, Jo Vandesompele, and Carl T. Wittwer. The miqe guidelines: Minimum information for publication of quantitative real-time pcr experiments. *Clinical Chemistry*, 55(4):611–622, 2009.

[34] Origene Team. qpcr primer pairs.

[35] Steve Lefever, Filip Pattyn, Jan Hellemans, and Jo Vandesompele. Single-nucleotide polymorphisms and other mismatches reduce performance of quantitative pcr assays. *Clinical Chemistry*, 59(10):1470–1480, Oct 2013.

[36] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240. ACM, 2006.

[37] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[38] Applied Biosciences. Power sybr® green rna-to-ct™1-step kit protocol (pn 4391003c).

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 8024–8035, 2019.

A Technical Appendices and Supplementary Material

Table 3: Feature definitions and methods for numerical features.

Feature	Description / Calculation method
Forward primer length	Length of the forward primer.
Forward primer melting temperature	Calculated melting temperature of the forward primer.
Forward primer GC content	Fraction of GC content in forward primer.
Forward primer indels	Number of gaps in forward primer/target alignment.
Forward primer mismatches	Number of mismatches in the forward primer/target alignment.
Reverse primer length	Length of the reverse primer.
Reverse primer melting temperature	Calculated melting temperature of the reverse primer.
Reverse primer GC content	Fraction of GC content in reverse primer.
Reverse primer indels	Number of gaps in reverse primer/target alignment.
Reverse primer mismatches	Number of mismatches in the reverse primer/target alignment.
Product length	Length of predicted amplicon.
Product melting temperature	Calculated melting temperature of the amplicon.

Table 4: Classification Model Test Set Performance Statistics

Architecture	Input	Accuracy	Precision	Recall	F1 Score	AUC
L2 Regularization	Features	0.797	0.787	0.797	0.783	0.854
L1 Regularization		0.806	0.798	0.806	0.797	0.860
Elastic Net (L1+L2)		0.806	0.798	0.806	0.797	0.860
RF		0.788	0.782	0.788	0.784	0.845
GB		0.809	0.801	0.809	0.801	0.865
MLP		0.819	0.816	0.819	0.817	0.878
SVM		0.806	0.797	0.806	0.795	0.860
KNN		0.794	0.793	0.794	0.794	0.843
CNN	Full Seq. (1.4 kb)	0.626	0.803	0.633	0.708	0.671
Transformer		0.717	0.717	0.910	0.834	0.600
LSTM		0.727	0.725	0.917	0.854	0.612
Lyra		0.841	0.898	0.878	0.887	0.904
CNN	Core Seq. (56 bp)	0.793	0.882	0.821	0.850	0.847
Transformer		0.830	0.853	**0.923**	0.886	0.890
LSTM		0.853	**0.910**	0.882	0.896	0.921
Lyra		0.849	0.877	0.918	0.897	0.915
Lyra + MLP	Core Seq. & Features	**0.860**	0.900	0.905	**0.903**	**0.930**

Table 5: Regression Model Test Set Performance Statistics

Architecture	Input	Performance			Time (s)	
		R2	MAE	RMSE	Training	Inference
Linear	Features	0.327	0.231	0.285		
Ridge		0.328	0.232	0.285		
RF		0.349	0.222	0.280		
SVR		0.357	0.211	0.278		
MLP		0.375	0.214	0.274		
GBR		0.393	0.217	0.270		
CNN	Full seq. (1.4 kb)	0.500	0.180	0.245	4.246	0.424
Transformer		0.050	0.289	0.338	38.296	3.227
LSTM		0.022	0.290	0.343	6.291	0.624
Lyra		0.633	0.157	0.210	22.875	1.823
CNN	Core seq. (56 bp)	0.499	0.180	0.246	1.412	**0.127**
Transformer		0.360	0.218	0.278	2.461	0.225
LSTM		0.594	0.157	0.221	**1.023**	0.128
Lyra		0.618	0.161	0.214	3.244	0.229
Lyra & MLP		**0.679**	**0.141**	**0.197**	3.449	0.274

Table 6: Number of variants used for each virus in the evaluation.

Virus	Number of variants (n)
CCHFV	119
CHIKV	1514
EBOVZ	135
HCV	868
LASV	98
mpox-Ia/Ib	551
mpox-IIb	2000
WNV	800
YFV	811
ZIKV	196

A.1 Origene Benchmark Dataset

To assess external generalizability of our models, we constructed a benchmark dataset from Origene qPCR primers for human genes. Origene is a commercial provider whose primers are extensively experimentally validated, with most products cited more than three times in literature. Each primer pair targets a human gene.

In addition, we incorporated findings from Huang et al. (2024) [8], who systematically tested 111 primer–target combinations under varying numbers, types, and positions of mismatches across two master mixes. Their study demonstrated that three or more mismatches at the $3'$ end of the primer drastically reduce PCR efficiency, while up to eight mismatches at the $5'$ end remain largely tolerable (maintaining $>80\%$ of the efficiency of a perfect match). We used these experimentally established rules to generate artificial positive and negative examples.

The resulting benchmark dataset comprised 33,860 true pairs and 38,168 false pairs:

- **True set** ($n = 33{,}860$)**:**
  1. Origene primer pairs with their intended targets ($n = 16{,}930$).
  2. The same targets with up to eight random mutations introduced at the $5'$ end of the primer-binding sites ($n = 16{,}930$).
- **False set** ($n = 38{,}168$)**:**
  1. Origene primer pairs paired with off-target human genes, excluding homologous mappings ($n = 6{,}893$).

2. Origene primer pairs paired with viral genomic sequences ($n = 14{,}345$).
3. Intended targets with 3–5 random mutations introduced at the $3'$ end of the primer-binding sites ($n = 16{,}930$).

This benchmark therefore combines experimentally validated commercial primers with biologically motivated synthetic perturbations, providing a rigorous and diverse test bed for evaluating model generalizability beyond our synthetic PCR dataset.

A.2 Wetlab Validation Protocol

To quantitatively assess the performance of PRIMERCAST primers against CDC-standard primers, we performed RT-qPCR experiments.

One-step RT-qPCR was performed using the Power SYBR™ Green master mix (Applied Biosystems, Waltham, MA, USA). SARS-CoV-2 RNA (Twist Biosciences, South San Francisco, CA, USA) were added at a concentration of 10^3 copies into each on-target reaction. Human lung total transcripts (ThermoFisher, Waltham, MA, USA) were added at a concentration of 10^6 copies to each off-target reaction.

One-step RT-qPCR was set up as follows according to manufacturer protocol [38].

- 10.0 µL 2× Power SYBR™ Green RT-PCR Mix
- 0.16 µL 125× RT Enzyme Mix
- 0.4 µL Forward primer (10 uM)
- 0.4 µL Reverse primer (10 uM)
- RNA template
- Nuclease-free water to 20.0 µL

Reactions were set up on ice, mixed gently, and briefly centrifuged before loading into a 96-well optical qPCR plate. Each sample was run in technical triplicate.

Thermal cycling was performed on a QuantStudio 6 instrument (Applied Biosystems, Waltham, MA, USA) with the following parameters. An annealing temperature of 55 °C was chosen to maximize compatibility across all primer sets.

1. Reverse transcription: 48 °C for 30 min
2. Initial denaturation: 95 °C for 10 min
3. 40 cycles of:
   - Denaturation: 95 °C for 15 s
   - Annealing/extension: 55 °C for 1 min (fluorescence acquisition)
4. Melt curve analysis: 60–95 °C, increments of 0.3 °C with continuous fluorescence monitoring

Primer sequences are as follows, derived from sources [27, 28, 29, 30, 31, 32].

Table 7: Existing standard SARS-CoV-2 primers and PRIMERCAST primers

Source	Forward (5' to 3')	Reverse (5' to 3')
US CDC N1	GACCCCAAAATCAGCGAAAT	TCTGGTTACTGCCAGTTGAATCTG
US CDC N2	TTACAAACATTGGCCGCAAA	GCGCGACATTCCGAAGAA
University of Hong Kong	TAATCAGACAAGGAACTGATTA	CGAAGGTGTGACTTCCATG
China CDC	GGGGAACTTCTCCTGCTAGAAT	CAGACATTTTGCTCTCAAGCTG
Japan National Institute of Infectious Diseases	AAATTTTGGGGACCAGGAAC	TGGCAGCTGTGTAGGTCAAC
Thailand NIH	CGTTTGGTGGACCCTCAGAT	CCCCACTGCGTTCTCCATT
Korea Institute for Basic Science	CAATGCTGCAATCGTGCTAC	GTTGCGACTACGTGATGAGG
PrimerCast_1	GCATTACGTTTGGTGGACCC	TGAACCAAGACGCAGTATTA
PrimerCast_2	AATGCACCCCGCATTACGTT	TGGACTGCTATTGGTGTTAA
PrimerCast_3	TCCTCATCACGTAGTCGCAA	TTACCAGACATTTTGCTCTC
PrimerCast_4	CGCAACAGTTCAAGAAATTC	AAGCTGGTTCAATCTGTCAA
PrimerCast_5	CCAAGGTTTACCCAATAATA	GTTAATTGGAACGCCTTGTC
PrimerCast_6	AGACCTTAAATTCCCTCGAG	AGTTCCTAGGTAGTAGAAAT
PrimerCast_7	AGACGGCATCATATGGGTTG	TTGGCAATGTTGTTCCTTGA
PrimerCast_8	CAAATTTCAAAGATCAAGTC	TAGGCTCTGTTGGTGGGAAT
RNaseP (control)	AGATTTGGACCTGCGAGCG	GAGCGGCTGTCTCCACAAGT

A.3 Compute Resources

Models were trained using a single NVIDIA Tesla T4 card. The model was implemented and computed on a 4 core, Xeon E5-2673 virtual machine with 26 GB of RAM and a NVIDIA Tesla T4 card.

Training times and inference times for the regression model are presented in Table 5. Similar training and inference times were observed in the classification models, which share much of the architecture with the regression models. The total compute time needed to run the training scripts averaged around 4-6 hours.

A.4 Model Training and Optimal Model Parameters

For numerical features, input features were normalized with the scikit-learn StandardScaler function [21], and training was performed until convergence or 1000 epochs, whichever came first.

All neural models were implemented in Pytorch, and trained for 50 epochs using the AdamW optimizer [39] with a learning rate of 10^{-3} with binary cross entropy (BCE) loss with a batch size of 64 [19], [40].

For both scalar feature and sequence encoders classifiers, model optimization was performed by hyperparameter search and performance was validated with 5-fold cross-validation. Unless otherwise noted, all reported results correspond to the best-performing architecture as selected based on validation set performance.

Optimal parameters for the LSTM classifier were determined to be a unidirectional 1-layer, 25 hidden unit model at learning rate of 10^{-3} and a weight decay of 0.01.

For both numerical and sequence regressors, hyperparameters were optimized via grid search. Training was performed using the AdamW optimizer with mean squared error (MSE) loss [39], [19]. For neural models, we performed limited hyperparameter sweeps over learning rate ($\{10^{-4}, 10^{-3}\}$) and weight decay ($\{0, 0.01\}$). Batch size was fixed at 64 across all experiments. Unless otherwise noted, reported results correspond to the best-performing configuration selected on the validation set.

Optimal parameters for the Lyra+MLP hybrid regressor were determined to be a 128-dimension MLP, 64-deminsional Lyra, and 32 dimensional combining layer at a learning rate of 10^{-3} and weight decay of 0.01.

Table 8: Model hyperparameters used in training.

Model Type	Model	Hyperparameter	Values
Scalar Feature	L1/L2/L1+L2 Logistic Regression	Regularization Strength	logarithmic in $[10^{-5}, 10^{5}]$
	L1 + L2 Logistic Regression	L1/L2 Mix Ratio	linear in $[0, 1]$
	Random Forest, Gradient Boosting	Estimators	logarithmic in $[10^{1}, 10^{3}]$
	Gradient Boosting	Learning Rate	logarithmic in $[10^{-4}, 10^{-1}]$
	Multi-Layer Perceptron	Hidden Layer Size	$[(50), (100), (50, 100)]$
		Regularization Strength	logarithmic in $[10^{-5}, 1]$
	Support Vector Machine	Regularization Strength	logarithmic in $[10^{-3}, 10^{3}]$
	k-Nearest Neighbors	Nearest Neighbors	linear in $[1, 20]$
		Weighting	Uniform, Distance
Sequence Encoder	Transformer	Heads	linear in $[2, 32]$
		Layers	linear in $[2, 16]$
		Internal Dimensions	linear in $[2, 16]$
	LSTM	Hidden Layer Size	linear in $[1, 50]$
		Layers	$[1, 5, 10]$
		Bidirectionality	True, False
	Lyra	Model Dimension	logarithmic in $[8, 128]$
Hybrid	Lyra + MLP	Lyra Model Dimension	logarithmic in $[8, 128]$
		MLP Model Dimension	logarithmic in $[8, 128]$

A.5 Societal Impacts

Due to the highly specialized nature of the dataset and of the models, the use case of these models would be utilized in designing more specific and sensitive primers for use in diagnostic applications. We envision using this model would greatly speed the detection of emerging pathogens by providing a reliable and accurate test for such a pathogen. This would be useful for designing better primers for other routine molecular biology tasks, such as cloning, sequencing, etc. In this capacity, the model would be providing a service to speed up both diagnostic and experimental work, helping further scientific progress. However, it is possible that certain parties would not have access to the resources needed to run this model. To address this inequity, we plan to both release the source code and dataset for this model as well as hosting a web interface for public use.

A.6 Model Limitations

The process of PCR has many factors that influence the efficiency beyond the base pairing interaction between primer and target. Buffer composition, polymerase activity, etc. can all affect the efficacy of the reactions. While the search space for reaction conditions is too large to be efficiently explored via wetlab experiments, the usage of multiple different master mix kits showed good correlation in primer performance despite significant differences in composition. This indicates sequence matching between primer and target plays a much more significant role in the sensitivity of the reaction compared to master mix differences.