Primer C-VAE: An interpretable deep learning primer design method to detect emerging virus variants

Hanyu Wang [1], Emmanuel K. Tsinda *[2], Anthony J. Dunn[4], Francis Chikweto [3], and Alain B. Zemkoho [4]

[1]School of Mathematics and Statistics, University of St Andrews, United Kingdom
[2]Department of Virology, Tohoku University Graduate School of Medicine, Japan
[3]Department of Biomedical Engineering, Tohoku University, Japan
[4]School of Mathematical Sciences, University of Southampton, United Kingdom

December 18, 2025

Abstract

Motivation:

Compared with next-generation sequencing (NGS), polymerase chain reaction (PCR) provides a more cost-effective and faster approach for detecting target organisms in both laboratory and field settings, where primer design is a critical step. In epidemiological surveillance of rapidly mutating viruses, designing effective primers is increasingly challenging. Traditional primer design workflows often require substantial manual intervention and may struggle to produce primers that remain specific across multiple strains within the same viral species (e.g., the Alpha and Delta variants of SARS-CoV-2). Similarly, for organisms with large and highly similar genomes (e.g., *Escherichia coli* and *Shigella flexneri*), designing primers that reliably discriminate between species is important but non-trivial. Therefore, more efficient and scalable primer design methods are needed.

Results:

We introduce Primer C-VAE, a convolutional variational auto-encoder (C-VAE) that integrates a variational auto-encoder (VAE) framework with convolutional neural networks (CNNs), as the core model for identifying sequence variants. We then exploit features learned in the convolutional layers for downstream post-processing to derive candidate regions and generate variant-specific primers. Using SARS-CoV-2 as a case study, Primer C-VAE classified five variants (Alpha, Beta, Gamma, Delta, and Omicron) with 98% accuracy on both the training and test sets, and generated primers for each variant. For most variants, the resulting primers appeared in more than 95% of target sequences and less than 5% of non-target sequences; Omicron showed moderately lower specificity (80% and 20%) owing to its greater genetic diversity. These primers showed good performance in *in silico* PCR tests. For the Alpha, Delta, and Omicron variants, the primer pairs produced amplicons shorter than 200 bp, enabling their use in downstream qPCR assay development. In addition, Primer C-VAE successfully generated effective primers for longer genomes, including *E. coli* and *S. flexneri*.

Conclusion:

Primer C-VAE is an interpretable deep-learning-based primer design approach for developing highly specific primer pairs for target organisms. It provides a flexible, semi-automated, and reliable workflow that is applicable across a range of sequence lengths and degrees of genome completeness. The method also supports downstream quantification applications (e.g., qPCR) and can be applied to a broad range of organisms, including those with large and highly similar genomes.

1 Introduction

Primer design plays a crucial role in modern molecular biology. Through polymerase chain reaction (PCR), it enables a cost-effective and rapid approach for detecting organisms in genetic testing and research. As Bustin

[*] Currently affiliated with Center for Biomedical Innovation, MIT, Cambridge, MA 02139-4307, United States

and Huggett emphasize, ”primers are arguably the single most critical components of any PCR assay” [1]. Because primers determine where amplification begins and ends, their design directly influences assay specificity, sensitivity, and overall efficiency. However, for rapidly mutating viruses such as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), primer design becomes increasingly challenging as new variants continue to emerge. High mutation rates can necessitate the rapid development and updating of primer pairs to detect newly circulating variants, which in turn affects surveillance efforts and can inform clinical and public-health decision-making. Similarly, for closely related bacteria with high genomic similarity, such as *Escherichia coli* (*E. coli*) and *Shigella flexneri* (*S. flexneri*), designing primers that reliably discriminate between species is essential for accurate detection. Therefore, efficient and reliable primer design methods are critical for molecular diagnostics and downstream applications.

Currently, next-generation sequencing (NGS), also referred to as high-throughput sequencing, is widely used to identify emerging viral variants, including SARS-CoV-2 variants [2]. However, implementing NGS typically requires sophisticated equipment and specialized expertise for both data generation and downstream analysis. Sanger sequencing offers a lower-cost alternative, but it remains time-consuming for genomes of approximately 30 kb, such as SARS-CoV-2 [3]. In practice, this approach requires fragmenting the genome into multiple overlapping segments (approximately 700–900 bp each), sequencing each fragment individually, and then reassembling the full sequence using bioinformatics methods. For targeted detection when the organism (or gene) of interest is known, PCR assays provide a more economical and flexible alternative to sequencing-based workflows, which contributes to their widespread use in both research and clinical settings [4, 5]. Conventional PCR assays require a target sequence and a primer pair (forward and reverse) within a standard PCR reagent mixture [6]. The successful performance of these assays depends critically on primers that bind specifically and efficiently to the intended target region.

1.1 Limitations of existing primer design methods

Reliable PCR testing depends critically on primer design, which typically involves a forward and a reverse primer. A standard workflow begins by collecting homologous sequences that cover the target region. Multiple sequence alignment is then used to facilitate visual inspection and to identify conserved regions (typically 18–25 nucleotides) suitable for primer binding. Candidate primers are subsequently evaluated for key thermodynamic properties, including guanine-cytosine (GC) content, melting temperature, and potential secondary structures, to ensure efficient amplification and to minimise primer–dimer formation. The forward primer is designed to anneal to the antisense strand, whereas the reverse primer is designed downstream and anneals to the sense strand of the target DNA (or cDNA derived from RNA). For further details of this conventional methodology, see [7].

In practice, however, primer design is considerably more complex than such schematic descriptions suggest. Rather than being a purely rule-based selection task, it often depends heavily on the designer’s expertise in sequence alignment, specificity assessment, and artefact prevention (e.g., dimers and hairpins). For example, designing primers to detect the SARS-CoV-2 Alpha variant requires identifying an 18–25 nucleotide region within a ˜30,000-nucleotide genome that is both unique to the Alpha lineage and compatible with strict thermodynamic constraints [8]. Achieving adequate specificity typically requires extensive comparison against other SARS-CoV-2 variants and related coronaviruses, which can be time-consuming and cognitively demanding. These requirements, together with the need for sustained attention during manual screening, make the process prone to human error and inconsistency.

To reduce manual effort, tools such as Primer3 and Primer3Plus can rapidly generate candidate primers that satisfy basic design constraints, thereby decreasing repetitive work and some sources of human error. However, primers that meet *in silico* criteria do not necessarily perform well in laboratory PCR experiments. Moreover, the specificity of automatically generated primers to a particular variant cannot be assumed and typically requires additional verification (e.g., database searches and experimental testing).

Another important limitation of commonly used tools is their restriction on input sequence length. For example, Primer3 processes sequences up to 10,000 base pairs, which makes it unsuitable for viruses such as SARS-CoV-2 with genomes exceeding 30,000 bp, and impractical for bacteria such as *E. coli* with genomes on the order of millions of base pairs. As a result, existing pipelines often cannot be applied directly to long genomic sequences without substantial preprocessing or manual intervention.

Taken together, the limitations of existing primer design methods can be summarised as follows:

1. Heavy reliance on specialist expertise and manual screening, resulting in time-intensive workflows that are vulnerable to human error.

2. Limited ability to handle long genomic sequences, restricting applicability across diverse organisms.

3. No inherent guarantee of primer specificity or experimental performance, necessitating additional *in silico* and laboratory validation.

To address these challenges, we propose a deep-learning-based, semi-automated primer design approach. Our method is designed to reduce manual screening, scale to long sequences, and generate more discriminative primers for closely related variants. This yields a streamlined and more robust workflow, which is particularly valuable for the detection of emerging pathogens and rapidly evolving viral lineages.

1.2 Proposed method

In this paper, we propose an interpretable deep-learning approach for primer design that generates both forward and reverse primers using a convolutional variational auto-encoder (C-VAE), which combines a variational auto-encoder (VAE) framework with convolutional neural networks (CNNs). We refer to this method as Primer C-VAE (Convolutional Variational Auto-Encoder for primer design), and show that it addresses several practical limitations of existing primer design workflows.

Primer C-VAE is designed for multiple primer design scenarios, including: (1) designing primers that discriminate between closely related variants within the same viral species (e.g., SARS-CoV-2 Alpha versus Delta), and (2) designing primers that distinguish between organisms with large and highly similar genomes (e.g., *Escherichia coli* and *Shigella flexneri*). In both settings, the goal is to produce primers that are not only thermodynamically feasible but also highly discriminative with respect to the target class.

Our pipeline consists of two main components: forward primer design and reverse primer design. For forward primer design, we first perform preprocessing to collect and validate genomic sequences and to improve data consistency. Specifically, we remove sequences that are incomplete, as well as outliers whose lengths deviate substantially from the dataset distribution (e.g., sequences that are much shorter than the typical length, such as below two-thirds of the mean, or that differ from the mean by more than one-third). We then train a C-VAE model to distinguish the target class from non-target sequences (either other variants of the same virus or sequences from other organisms). To obtain candidate primer regions, we leverage patterns captured by the convolutional layers in the encoder and extract variable-length motifs within the typical primer length range (18–25 bp). These candidates are then filtered using standard thermodynamic criteria and dimer/hairpin checks to identify viable forward primers.

Reverse primer design follows a similar procedure, with one key difference: we use the selected forward primer to locate the downstream target region from which the reverse primer should be designed. These downstream sequences, together with a synthetically generated background dataset that matches their nucleotide composition, are used as inputs to a second C-VAE model. Candidate reverse primers are obtained and filtered using the same thermodynamic and dimer-screening steps, and are paired with the forward primers to form complete primer sets. We further validate primer pairs using Primer-BLAST [9] to reduce off-target amplification and to avoid primer annealing to human genomic sequences as well as closely related microbial genomes. Finally, *in silico* PCR [8] is used to evaluate the specificity and effectiveness of the resulting primer pairs for detecting the target variant(s) or organism(s).

To our knowledge, Primer C-VAE is among the first approaches to employ a VAE-based deep-learning framework for primer design while explicitly supporting variable-length primer candidates and generating both forward and reverse primers within a unified workflow. The method is effective for sequence sub-classification tasks such as discriminating SARS-CoV-2 variants, as well as for separating organisms with large and highly similar genomes, including *E. coli* and *S. flexneri*. Across our experiments, Primer C-VAE achieves strong classification performance and produces highly specific primer pairs. For SARS-CoV-2 variant classification, the model reaches over 98% accuracy, and the generated primer pairs show high specificity: they appear in more than 95% of sequences from the target variant while occurring in less than 5% of sequences from other variants (with Omicron as an exception, where the corresponding values are approximately 80% and 20%). For *E. coli* and *S. flexneri*, the method achieves over 96% accuracy, and the resulting primers exhibit comparable specificity (above 95% in target sequences and below 5% in non-target sequences). A practical advantage of the designed primer pairs for SARS-CoV-2 is that they yield short amplicons (typically < 200 bp), which facilitates downstream assay development across multiple PCR modalities, including conventional PCR, RT-PCR, and qPCR. Overall, Primer C-VAE provides a semi-automated workflow that reduces manual screening effort, scales to long genomic sequences, and improves the reliability of primer design for rapidly evolving pathogens and closely related organisms.

1.3 Related work

Recent advances in computational biology have enabled a wide range of computational strategies for primer design in molecular diagnostics. Primer3, first released in 2000, remains one of the most widely used open-source tools for primer design [10, 11]. It implements established thermodynamic models to estimate oligonucleotide melting temperatures during hybridisation and to assess the stability of potential secondary structures. Although Primer3 is effective for many species-level detection tasks in viral and bacterial assays, it often produces a large set of candidates that still require downstream screening, and it has limited capability for discriminating among closely related variants within the same viral species (e.g., SARS-CoV-2 lineages). In addition, its maximum input length of 10,000 bp restricts direct application to long genomes, including bacterial pathogens such as *Escherichia coli* with genomes on the order of 4.5–5.5 Mb.

To address these limitations, several alternative approaches have been proposed. One strategy uses finite state machines (FSMs) to classify primers into suitable and unsuitable categories [12]. This method can complement Primer3 by prioritising higher-quality candidates from its output. However, it typically requires a sufficiently large and representative primer training set, which can limit applicability to rapidly evolving pathogens such as SARS-CoV-2, where frequent updates or retraining may be necessary.

Genetic algorithm-based approaches provide another direction for primer design [13]. In principle, these methods can overcome fixed input-length constraints and may improve the search for feasible primers compared with rule-based generation. Nevertheless, they have several practical limitations: reliance on stochastic mutation and crossover can lead to instability when mutation patterns are complex or rapidly changing; species- or dataset-specific tuning is often required for hyperparameters such as crossover probability (p_c), mutation probability (p_m), and population size (p); and, like many heuristic optimisation methods, they may converge to local optima rather than globally optimal solutions. Moreover, while such approaches have been used primarily for species-level primer design, they generally do not directly support robust variant-specific detection.

More recent evolutionary algorithm methods have been developed to design both forward and reverse primers for SARS-CoV-2 variant detection [14]. A common characteristic of these approaches is their reliance on pre-identified mutation regions (or “signature” mutations) to guide primer discovery. This dependence can reduce robustness when mutation hotspots shift over time and may limit generalisability to other pathogens when variant-defining mutations are not well characterised. In settings where mutation information is unavailable or incomplete, these methods may be difficult to apply.

Machine learning approaches, such as those in [15], further improve automation by enabling primer generation without explicit reliance on known mutation sites and without strict input-length constraints. Compared with genetic and evolutionary algorithms, these methods can be computationally efficient and more stable across datasets. However, important limitations remain: existing implementations typically generate only forward primers of a fixed length (e.g., 21 bp), still require manual selection of reverse primers based on domain expertise, and often focus on species-level discrimination rather than systematic variant-level classification.

Building on prior learning-based frameworks [15], our method is designed to reduce manual effort in identifying discriminative genomic regions within large sequence collections. It supports long input sequences, enabling analysis of both the ˜30 kb SARS-CoV-2 genome and bacterial genomes exceeding 5 Mb, such as those of *E. coli*. As summarised in Table 16, our approach improves the efficiency of primer design for sequence sub-classification compared with Primer3. A key component of the proposed pipeline is a VAE-based deep learning architecture that supports reverse primer design and enables variable-length primer candidates (18–25 bp) using only the target organism’s genomic sequences as input. This design reduces reliance on mutation annotations and extensive preprocessing, which is particularly useful for rapidly evolving pathogens and microorganisms with large genomes, as well as in scenarios where mutation information is inaccessible or incomplete.

1.4 Outline of the paper

The remainder of this paper is organised as follows. Section 2 describes the proposed Primer C-VAE pipeline and its four sequential computational stages. We begin with genomic data acquisition and bioinformatic preprocessing to construct sequence alignment matrices suitable for neural network input. We then present the convolutional variational auto-encoder architecture used to generate forward-primer candidates, introduce the reverse-primer design procedure, and conclude with validation steps based on BLAST sequence similarity analysis and *in silico* PCR simulation.

Section 3 evaluates the method in two applications. In Section 3.1, we demonstrate variant-specific primer design for SARS-CoV-2 and assess discrimination among the Alpha, Beta, Gamma, Delta, and Omicron variants. In Section 3.2, we extend the evaluation to organisms with substantially larger genomes and design primers that discriminate between the closely related bacterial species *Escherichia coli* and *Shigella flexneri*.

Section 4 discusses the results, compares Primer C-VAE with traditional primer design workflows, analyses the strengths and limitations of the framework, and outlines directions for future improvement. We conclude by summarising the contribution of Primer C-VAE as a computational approach for designing primers for target-specific detection, particularly in settings involving closely related variants and highly similar genomes.

The appendices provide supplementary material, including detailed protocols for data collection and preprocessing, feature extraction and evaluation metrics, workflow flowcharts, BLAST and *in silico* PCR validation results, and comparative analyses with existing primer design tools.

2 Process primer design with Primer C-VAE

Our Primer C-VAE methodology comprises four sequential computational stages for an integrated primer design workflow (Figure 1). Stage I covers genomic data acquisition and bioinformatic preprocessing to construct sequence alignment matrices suitable for neural network input. Stage II generates forward-primer candidates using our convolutional variational auto-encoder architecture, extracting discriminative sequence patterns from the preprocessed data. Stage III performs reverse primer design, using the forward-primer candidates identified in Stage II as anchors to define downstream regions and derive complementary reverse primers that satisfy standard thermodynamic constraints. Stage IV applies validation procedures, including BLAST sequence similarity analysis and *in silico* PCR simulation, to assess primer specificity and amplification performance.

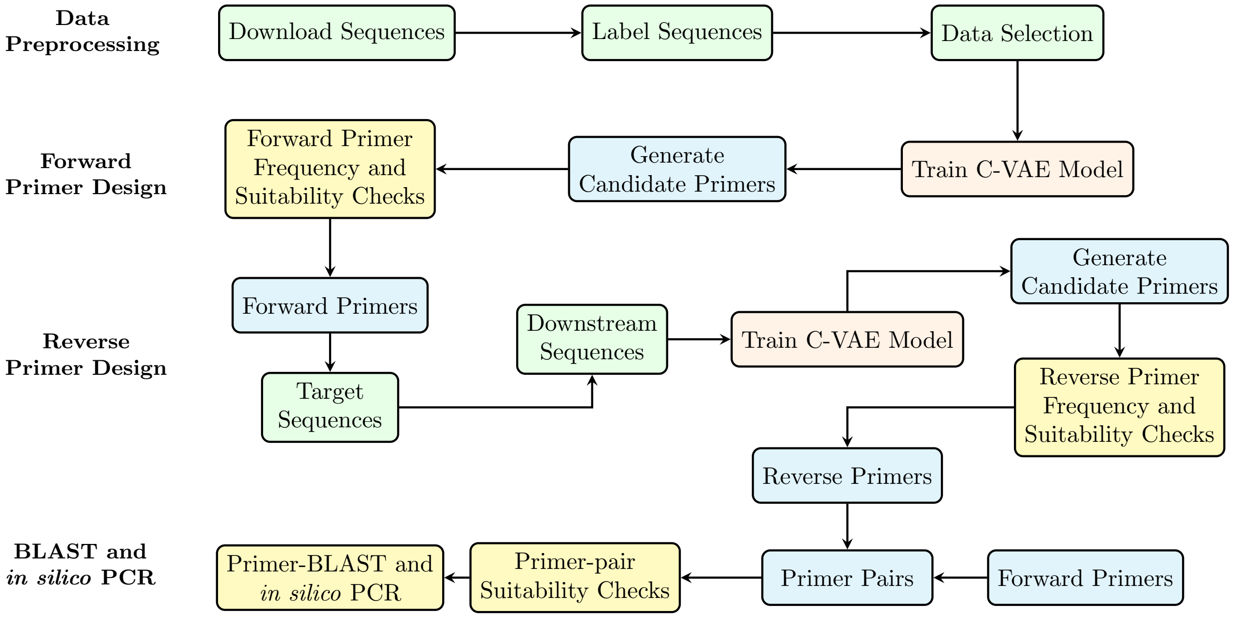


Figure 1: Primer C-VAE primer design workflow.

The Primer C-VAE workflow comprises four stages. **Stage I (Data acquisition and preprocessing)** downloads sequences, assigns labels, and curates datasets for model training. **Stage II (Forward primer design)** trains a C-VAE model, generates candidate forward primers, and filters them using frequency-based screening and thermodynamic suitability checks. **Stage III (Reverse primer design)** extracts downstream regions anchored by selected forward primers, trains a second C-VAE model, generates candidate reverse primers, and applies the same filtering criteria. **Stage IV (Validation)** forms primer pairs, evaluates amplicon size and primer–dimer risk, and assesses specificity using Primer-BLAST followed by *in silico* PCR simulation.

The following sections provide an overview of our computational methodology. We begin with genomic data acquisition and bioinformatic preprocessing, and then detail the primer design framework, including neural-network-based feature extraction and the quantitative metrics used for selecting forward and reverse primer candidates. We conclude by describing the primer-pair validation procedures.

2.1 Stage I: Data Acquisition and Pre-processing

Data acquisition. Genomic sequence data used in this study were obtained from two established repositories: GISAID (Global Initiative on Sharing Avian Influenza Data, [16]) and NCBI (National Center for Biotechnology Information, [17]). These repositories provided SARS-CoV-2 variant genomes and genomic sequences for *Escherichia coli* and *Shigella flexneri*. The sequences were used for multiple computational objectives: training the convolutional variational auto-encoder (C-VAE) for target classification, extracting candidate primer features from the trained network, identifying downstream regions for reverse primer design, and performing specificity checks through comparative sequence analyses. Our dataset includes 8,939 *E. coli* and 5,373 *S. flexneri* sequences from NCBI, together with 473,645 SARS-CoV-2 sequences from GISAID spanning five variants (Alpha, Beta, Gamma, Delta, and Omicron). In total, we analysed 610,000 complete SARS-CoV-2 genomes from GISAID and NCBI; the 473,645 sequences mentioned above refer to the GISAID subset used for model development, while the remaining NCBI sequences were used primarily for independent testing and appearance-rate evaluation. Variant labels follow World Health Organization (WHO) nomenclature, the Pango lineage classification system, and GISAID clade designations. Detailed sample distributions and labels are provided in Appendix Table 6.

Data pre-processing. Our preprocessing pipeline applies organism-specific filtering to improve length consistency. For most organisms, we compute the mean sequence length and remove sequences whose lengths deviate by more than one-third above or below the mean. For viruses with relatively compact and near-complete genomes, such as SARS-CoV-2, we use a less stringent rule and remove only sequences shorter than two-thirds of the mean length to preserve coverage. After filtering, we compute the maximum sequence length (max_vector) within each dataset and standardize all sequences by padding with the nucleotide "N" to obtain a fixed-length representation, yielding $1 \times 1 \times max\_vector$ matrices. In the SARS-CoV-2 dataset, sequences averaged 29–30 kb with a maximum length of 31,079 bp; we removed only sequences shorter than 20 kb and represented each sequence as a $1 \times 1 \times 31{,}079$ matrix for neural network input. Variant-specific average lengths are reported in Appendix Table 7.

To serve as neural network input, each standardized sequence must be converted from categorical nucleotides (A, T, C, G, and N) into numerical values [18]. While common bioinformatic encodings include one-hot and k-mer representations, we use ordinal encoding as the default scheme (Equation 1). This choice is computationally efficient and simple to implement for convolutional architectures, and has been reported to provide competitive performance in similar settings [19]. One-hot encoding is also supported as an alternative when appropriate.

Before model training, each standardized sequence is ordinal-encoded using Equation 1; An example of this transformation is shown in Appendix Figure 13.

$$Y := f(x) := \begin{cases} 0 & \text{if} \quad x = N, \\ 1 & \text{if} \quad x = C, \\ 2 & \text{if} \quad x = T, \\ 3 & \text{if} \quad x = G, \\ 4 & \text{if} \quad x = A. \end{cases} \tag{1}$$

For downstream interpretation, we also implement an inverse transformation that converts numerical outputs back to nucleotide symbols. For classification tasks, taxonomic labels are encoded as integer class indices $y \in \{0, 1, \ldots, C-1\}$ for model training. For example, SARS-CoV-2 variant classification uses $C = 5$ classes (Alpha, Beta, Gamma, Delta, and Omicron), and each sequence is assigned a single class index. When needed for reporting or downstream analysis (e.g., confusion matrices or plots), these indices can be converted to one-hot vectors of dimension 1×5.

Data selection. After preprocessing, we apply a structured data partitioning strategy to support three computational requirements: model training, validation, and testing. During primer design, we additionally select target-class sequences to generate candidate primers and construct reference sequence sets for specificity evaluation. These reference sets support: (i) assessing primer specificity to the target class, (ii) estimating appearance frequencies to reduce off-target amplification risk, and (iii) checking non-complementarity to human genomic sequences and closely related microbial genomes to mitigate cross-reactivity. Accordingly, our data selection procedure includes three steps: (1) splitting data for training/validation/testing; (2) selecting sequences for primer candidate generation; and (3) compiling reference datasets for specificity assessment. A worked example for SARS-CoV-2 variant-specific primer design is provided in Table 1.

2.2 Stage II: Forward Primer Design

Primer C-VAE architecture. We use a dataset of pre-processed genome sequences represented as fixed-length input matrices, $\mathcal{D} = \{seq_1, seq_2, \ldots, seq_n\}$, where each $seq_i \in \mathbb{R}^{1\times1\times max\_vector}$ denotes an encoded genome sequence. The C-VAE model is trained for supervised classification with C classes, where C depends on the application (e.g., $C = 5$ for SARS-CoV-2 variants: Alpha, Beta, Gamma, Delta, and Omicron; and $C = 2$ for *E. coli* versus *S. flexneri*). The model learns latent representations of input sequences, which are used both for sequence reconstruction via the decoder and for class prediction via the classifier head. Figure 2 illustrates the architecture.

The C-VAE model uses a five-layer encoder: two convolutional layers (Conv2D) with ReLU activations, each followed by a max-pooling layer (MaxPool2D), and a final fully connected (FC) layer with ReLU and batch normalization (BN). This convolutional hierarchy progressively extracts sequence features, where earlier layers capture short, motif-like patterns and deeper layers capture higher-level combinations of these patterns. ReLU activations improve optimization behavior and mitigate vanishing gradients during backpropagation [20]. Max-pooling reduces the effective dimensionality while retaining salient features, enabling the FC layer to aggregate multi-scale information for downstream tasks.

Within the variational component, the latent representation z is obtained via the reparameterization trick, using encoder outputs μ and $\log \sigma^2$ to define a Gaussian distribution from which z is sampled. From z, the architecture branches into two pathways: a classification head that predicts sequence labels, and a decoder that reconstructs the input using an approximately inverted architecture.

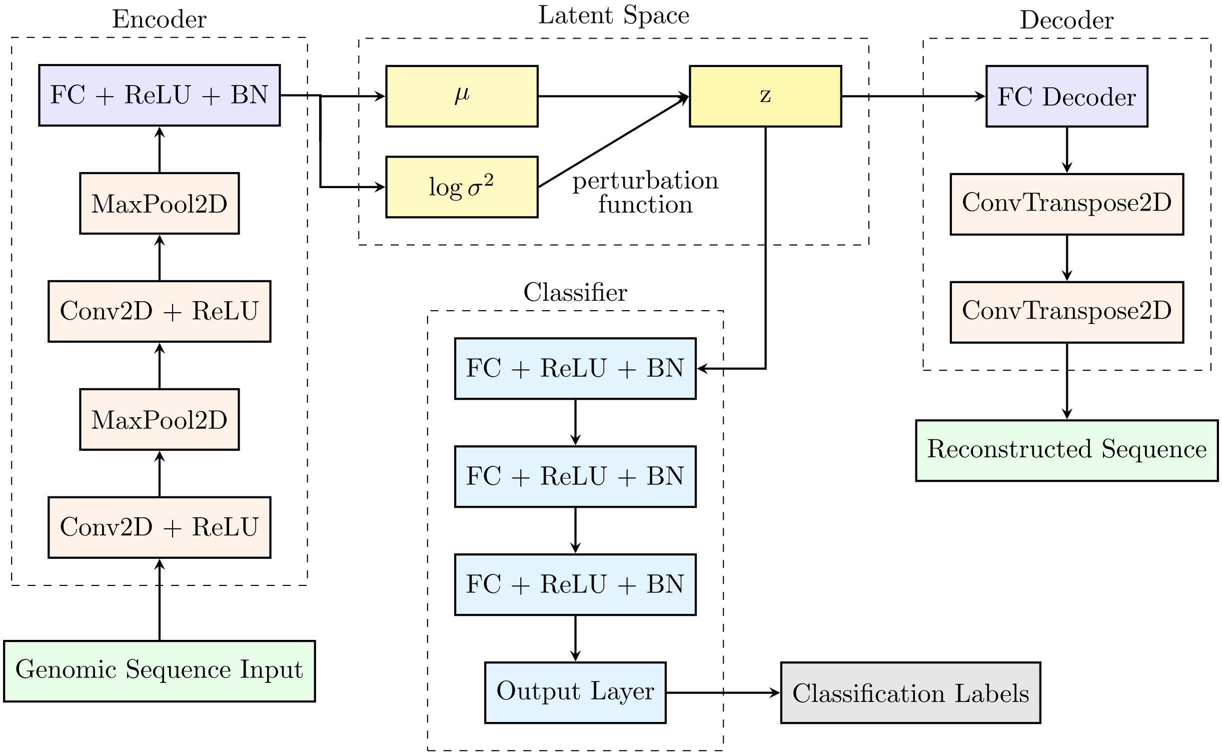


Figure 2: Primer C-VAE architecture

Note: Primer C-VAE consists of three components: (1) a convolutional encoder that extracts hierarchical features from genomic inputs, (2) a variational latent space in which a latent vector z is sampled using the reparameterization trick from the learned parameters μ and $\log\sigma^2$, and (3) two output heads—a classifier for sequence categorization and a decoder for sequence reconstruction. The model is trained end-to-end by jointly optimizing the classification objective and the reconstruction objective, encouraging the latent representation to be both discriminative and informative.

The decoder serves two roles in our pipeline. First, reconstruction provides a regularization signal that encourages the latent representation to preserve information about the input sequence. Second, by comparing reconstructed sequences with the original inputs, we can highlight regions where reconstruction errors are consistently higher across classes, which can indicate genomic loci with increased variability between variants. Although such regions are not guaranteed to be the sole drivers of classification performance, they can provide useful biological cues for primer design by prioritizing candidate segments that differ between target and non-target groups. This dual-objective design therefore supports both discriminative learning and downstream interpretation.

2.2.1 Model Training

During preprocessing, input genomic sequences are standardized to a uniform length using the maximum sequence length within each dataset as the reference. Each sequence is represented as a $1 \times 1 \times max\_vector$ tensor to ensure compatibility with the C-VAE architecture. After data curation, the C-VAE model is trained to learn discriminative genomic features that separate target classes from non-target classes. In the SARS-CoV-2 experiment, this is formulated as a five-class classification task, where each sequence is assigned to one of

$$\mathcal{C} = \{\text{Alpha, Beta, Gamma, Delta, Omicron}\}$$

Training uses the Adam optimizer [21] with an adaptive learning rate, together with a multi-class classification objective ($\mathcal{L}_{class}$). Specifically, the classifier head outputs logits $s \in \mathbb{R}^5$, and $\mathcal{L}_{class}$ is defined as the categorical cross-entropy loss:

$$\mathcal{L}_{class} = -\sum_{k=1}^{5} y_k \log p_k, \quad \text{where } p = \text{softmax}(s).$$

In practice, $\mathcal{L}_{class}$ is implemented using PyTorch's `CrossEntropyLoss`, which takes logits as input and applies the softmax operation internally for numerical stability [22].

The training objective combines multiple loss terms. The reconstruction loss ($\mathcal{L}_{recon}$), defined as the mean squared error between the input sequence and its reconstruction, encourages the latent representation to preserve information about the input. In addition, the Kullback–Leibler divergence loss ($\mathcal{L}_{KL}$) acts as a regularizer that encourages the approximate posterior to remain close to a standard normal prior, given by:

$$\mathcal{L}_{KL} = -0.5 \sum (1 + \log \sigma^2 - \mu^2 - \sigma^2).$$

The overall loss function is:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta \cdot \mathcal{L}_{KL} + \lambda_{class} \cdot \mathcal{L}_{class} + \lambda_{reg} \cdot \mathcal{L}_{reg}.$$

Here, $\mathcal{L}_{reg}$ denotes an L2 regularization term that mitigates overfitting by penalizing large parameter magnitudes. The hyperparameters β, λ_{class}, and λ_{reg} serve as weighting coefficients that control the relative contributions of each term. This combined objective encourages the model to learn latent representations that are both informative for reconstruction and discriminative for variant classification, which supports downstream primer candidate identification.

2.2.2 Feature Extraction and Forward Primer Generation

After training the C-VAE model, we extract candidate forward primers using four feature-identification strategies. Three strategies derive candidate regions from activation patterns in the convolutional encoder, whereas the fourth leverages differences between the input sequence and the decoder reconstruction.

The first convolutional layer in the encoder is designed to capture short, motif-like patterns and consists of 12 filters with a $1 \times N$ kernel, where N is set to match the desired primer length. In practice, we generate candidate regions within the typical primer range (18–25 nt) by applying post-processing to the activation maps of this first convolutional layer using three encoder-based methods: (1) Pooling, (2) Top-k, and (3) Mix. In addition, a decoder-based method, (4) Reconstruction, identifies informative positions by analyzing discrepancies in the reconstructed sequence.

Pooling method. This method is inspired by max-pooling, but with an important modification: instead of retaining only the maximum value, we record the position of the maximum activation within each pooling window. Concretely, for each filter activation map, we apply a custom max-pooling operation that stores argmax indices (Figure 3). The nucleotide positions corresponding to these maxima are written to a position file. This procedure is applied independently to all 12 filters to capture a diverse set of candidate positions.

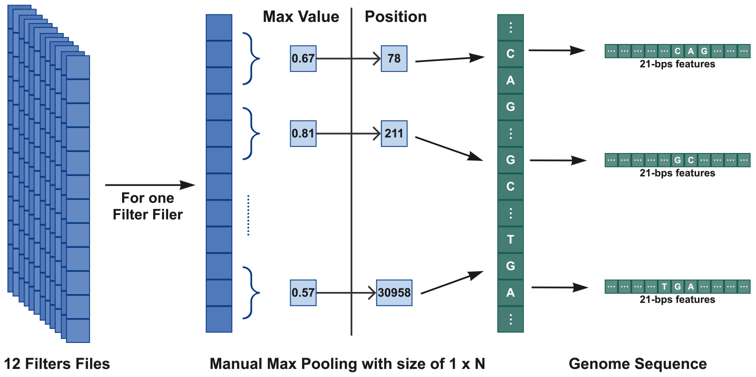


Figure 3: Feature extraction from filter activation maps by simulating max-pooling with index tracking.

Top-k method. Unlike pooling, the Top-k method performs global selection without predefined pooling windows. Given a user-specified parameter k, we identify the k largest activation values across the full activation map and record their positions. We selected k via empirical analysis on the SARS-CoV-2 Alpha and Delta datasets, evaluating $k \in \{75, 125, 175, 250\}$, corresponding to approximately 0.25%, 0.50%, 0.58%, and 0.83% of a ˜30,000-nt genome. As reported in Appendix Table 9, $k = 175$ (0.58%) provided a good trade-off between computational cost and primer generation performance, while maintaining high appearance rate within the target variant.

Mix method. The Mix method combines the Pooling and Top-k strategies. We first apply pooling to partition the activation map into windows, and then record the top-k positions within each window. Based on empirical testing, we use a pooling window size of 1×500 and record the top 10 positions per window.

Reconstruction method. This method uses the decoder reconstruction to identify candidate positions. Although the reconstructed sequence has the same length as the input, it may differ at specific nucleotide positions. Rather than treating these discrepancies as errors, we use them as a signal: positions with consistently higher reconstruction divergence can indicate regions that are less well captured by the shared representation or that differ systematically between classes. Such regions may be informative for primer design, for example by highlighting loci that vary between target and non-target groups or loci with elevated variability within variant populations. We therefore record positions with the highest divergence between the input and the reconstructed sequence (Figure 4).

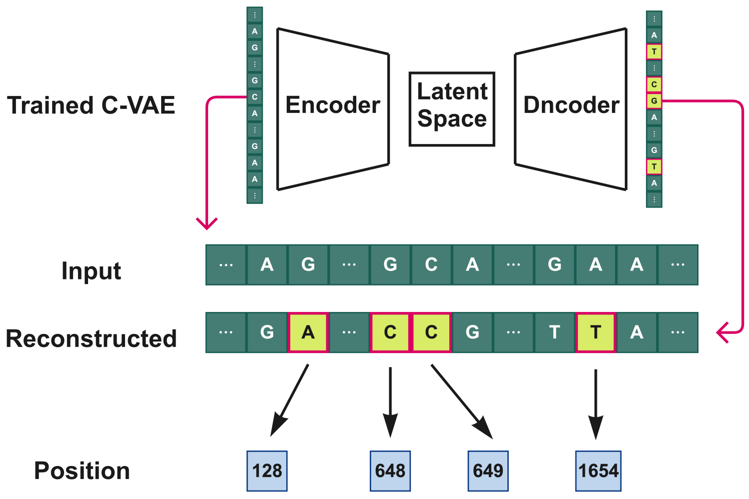


Figure 4: Feature extraction from reconstructed sequences by identifying nucleotide divergence.

Our forward-primer design procedure begins by using the extracted positions as anchors for primer construction. For each selected position, we retrieve the corresponding nucleotide in the input sequence and extend to include flanking bases to reach a user-defined primer length $L \in [18, 25]$ nt. For example, when $L = 25$, we construct a 25-nt primer by taking the anchor nucleotide plus 12 nt upstream and 12 nt downstream. For even lengths, we extend asymmetrically using $\lfloor (L-1)/2 \rfloor$ bases on one side and $\lceil (L-1)/2 \rceil$ on the other to obtain exactly L nucleotides.

Each candidate primer is then filtered using standard quality criteria, including thermodynamic properties, dimer/hairpin risk, and appearance rate in the target dataset. As shown in Figure 5, candidates must satisfy established primer design constraints [24, 25] before being retained as viable forward primers:

1. Length of 18–25 nt;
2. GC content of 40–60%;
3. 1–2 G/C bases at both the 5′ and 3′ ends (GC clamp);
4. Melting temperature (Tm) of 45–60°C (calculated with Primer3; see also [26]);
5. For downstream pairing, forward and reverse primers should have a Tm difference within 5°C;
6. No strong self-complementarity (e.g., no more than five consecutive complementary bases), assessed using IDT OligoAnalyzer [27].

To assess specificity, we compute the appearance rate of each candidate primer in target and non-target sequence sets. Primers that fail to bind to the target class or that match multiple loci can reduce assay performance and increase the risk of non-specific amplification [28]. Unless otherwise stated, the appearance rate is computed using an **exact-match** rule (0 mismatches): a primer is counted as present in a sequence only if its nucleotide string (or its reverse complement) appears as an exact contiguous substring. This strict filter provides a conservative first-pass specificity screen; potential mismatch-tolerant binding and off-target amplification are further evaluated in Stage IV via thermodynamic analysis and *in silico* PCR tools.

2.3 Stage III: Reverse Primer Design

In conventional PCR assay design, the reverse primer is placed downstream of the forward primer on the reference (5′ →3′) sequence, and the distance between the two primer binding sites determines the amplicon

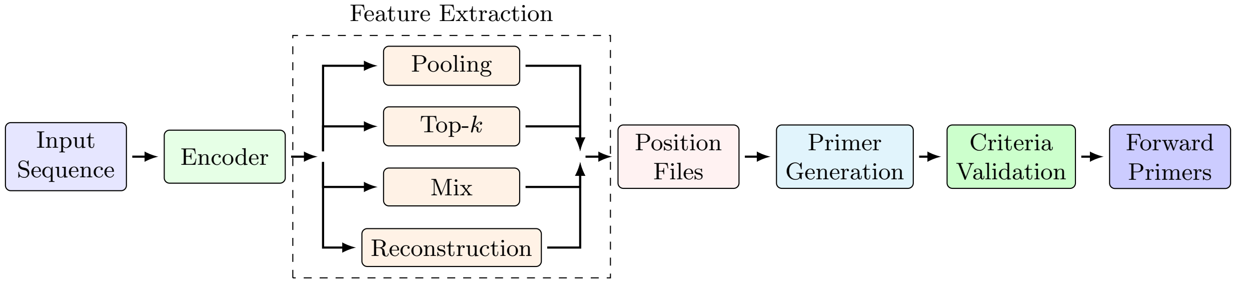


Figure 5: Computational workflow for feature extraction and forward primer design. Candidate positions are identified by four feature extraction methods and recorded in position files, which guide primer construction. Candidate primers are then filtered using thermodynamic and specificity criteria.

length. For example, when a short product is desired (e.g., ∼200 bp for qPCR), the reverse primer is selected such that the resulting amplicon length falls within the target range.

Building on this principle, we develop a deep-learning-assisted procedure for reverse primer design. Unlike forward primer generation, reverse primer design in our pipeline is conditioned on two inputs: (1) validated forward primers and (2) the corresponding target-class genome sequences. Using the forward primers as anchors, we restrict the search space to biologically relevant downstream regions, which improves efficiency and reduces unnecessary scanning of the full genome.

2.3.1 Data pre-processing for generating the downstream dataset

Reverse primer design requires target-class sequences and validated forward primers. We first locate each validated forward primer within each target sequence and then define the downstream region in which the reverse primer binding site must lie, i.e., the segment extending from the end of the forward primer to the 3′ terminus on the reference sequence. This reflects the standard convention that sequences are indexed in the 5′ →3′ direction, and reverse primers are chosen downstream of the forward primer to yield the desired amplicon length.

For each target sequence containing a validated forward primer, we partition the sequence into three regions (Figure 6):

1. Upstream region: from the 5′ terminus to the nucleotide immediately preceding the forward primer;
2. Forward primer region: the forward primer binding segment;
3. Downstream region: from the last nucleotide of the forward primer to the 3′ terminus.

For reverse primer design, we retain only the downstream region and construct a downstream dataset. This targeted preprocessing step confines the candidate search space to biologically valid locations and substantially reduces computational complexity.

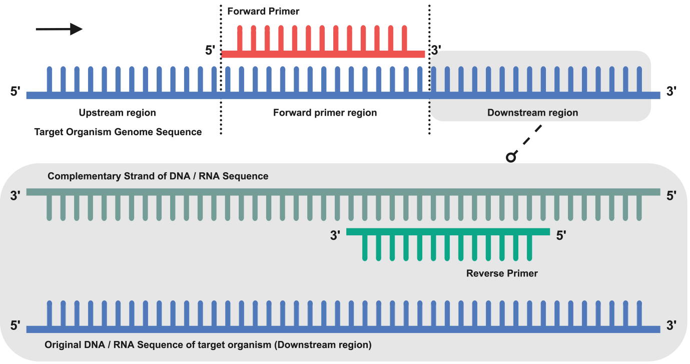

Forward Primer
5'
3'
5'
3'
Upstream region
Forward primer region
Downstream region
Target Organism Genome Sequence
Complementary Strand of DNA / RNA Sequence
3'
5'
3'
5'
Reverse Primer
5'
3'
Original DNA / RNA Sequence of target organism (Downstream region)

Figure 6: Functional segmentation of target sequences for reverse primer design. Each sequence is partitioned into upstream, forward primer, and downstream regions. The start position and length of the downstream region depend on the binding location of the validated forward primer in that sequence.

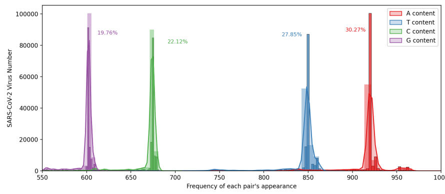


Figure 7: Nucleotide composition analysis of downstream regions in the target dataset.

2.3.2 Generation of Synthetic Downstream Data

We adapt the C-VAE framework used in Stage II to reverse primer design. A key challenge is that the downstream dataset constructed for a given target variant (or target class) contains only one biological label, which provides no explicit contrast for supervised discrimination. Without a meaningful negative class, the model may learn features that satisfy the training objective without capturing class-specific biological constraints.

To introduce a contrasting label, we construct a synthetic reference dataset and formulate reverse primer learning as a binary classification task: real downstream sequences versus synthetic downstream sequences. The synthetic sequences are generated to match the nucleotide composition of the real downstream dataset. Specifically, we estimate the empirical A/T/C/G frequencies from authentic downstream regions (Figure 7) and then sample random sequences that preserve these mononucleotide proportions. This provides a controlled negative class that is statistically similar at the nucleotide-frequency level but does not reflect biological structure shaped by evolutionary and functional constraints.

For clarity, consider the SARS-CoV-2 Delta example with the validated forward primer `CTACCGCAATGGCTTGTCTTG`. Our procedure consists of four steps: (1) select Delta sequences; (2) extract downstream regions based on the forward primer location; (3) compute nucleotide composition statistics; and (4) generate a synthetic downstream dataset that matches these statistics.

2.3.3 Model Training and Feature Extraction

The C-VAE architecture and loss formulation remain the same as in Stage II, but the training data and labels differ. We train the model to discriminate authentic downstream sequences from synthetic sequences under this binary classification setup. Although synthetic sequences match the overall nucleotide composition of real downstream regions, they lack higher-order patterns and biological constraints present in genuine genomes. Consequently, the trained model can learn features that reflect biologically structured sequence patterns rather than nucleotide frequencies alone. These learned features are then used to prioritize candidate regions for reverse primer generation in the downstream dataset.

Reverse primer candidate extraction. After training on the real-versus-synthetic downstream datasets, we extract candidate reverse primers from the downstream regions using the same four feature extraction strategies described in Stage II (Pooling, Top-k, Mix, and Reconstruction). Concretely, we apply the encoder-based methods to the first convolutional-layer activation maps of downstream inputs and record high-activation nucleotide positions in position files, and we apply the Reconstruction method by identifying positions with high divergence between the downstream input and its reconstruction. These positions serve as anchors for candidate construction, identical to forward primer generation: for a desired primer length $L \in [18, 25]$ nt, we extend around each anchor position within the downstream sequence to obtain an L-nt candidate. Candidates that cannot be formed due to proximity to the downstream boundary are discarded.

Because reverse primers anneal to the complementary strand, each extracted candidate sequence is converted to its reverse complement before thermodynamic screening and pairing with the corresponding forward primer. This reuse of the Stage II extraction procedures ensures that forward and reverse primer candidates are generated in a consistent, model-driven manner while restricting reverse primer search to biologically valid downstream regions.

2.4 Stage IV: *in silico* PCR and Primer-BLAST Validation

Following reverse primer design, all candidate primer pairs undergo *in silico* validation using standard virtual PCR simulation protocols [8]. This step allows us to assess expected amplification behavior and primer specificity before laboratory testing. Reverse primer candidates are generated using the same C-VAE architecture and the same four feature extraction strategies as in forward primer design. For validation, we use three complementary tools: (1) FastPCR [29] for thermodynamic analysis and *in silico* PCR; (2) Unipro UGENE [30] for thermodynamic analysis and *in silico* PCR; and (3) Primer-BLAST [9] for genome-scale specificity assessment.

We evaluate each primer pair against two criteria: (i) successful amplification of the intended target with an amplicon length within the desired size range, and (ii) no detectable off-target amplification products when tested against the selected reference genomes. Primer pairs are considered *in silico* validated only if they satisfy both criteria consistently across all three tools. Candidates that pass this *in silico* screening are then prioritized for downstream experimental evaluation. This multi-tool validation strategy provides an additional quality-control layer and helps reduce the number of unsuitable primer pairs carried forward to wet-lab testing. Throughout this paper, the term ”validation” refers to to computational screening (does not constitute wet-lab validation) including *in silico* validation via Primer-BLAST, FastPCR, and UGENE, unless explicitly stated otherwise.

3 Numerical Experiment and Results

3.1 Experiment 1: SARS-CoV-2 Emerging Variant Primer Design

Although the acute public health emergency has subsided, SARS-CoV-2 continues to evolve, giving rise to new lineages that may warrant continued surveillance [31–33]. According to World Health Organization (WHO) surveillance data, more than 700 million COVID-19 cases had been reported globally by October 2024, with reported deaths exceeding 7 million [34]. The recent detection of the XEC lineage—a recombinant derived from the KS.1.1 and KP.3.3 lineages—in multiple European countries and the United Kingdom [35] further highlights the ongoing need for robust molecular detection systems capable of rapidly identifying novel genomic signatures.

	Training set	Validation set	Test set	Generated primers	Calculated appearance
Source	GISAID	GISAID	NCBI	GISAID	GISAID and NCBI
Alpha	2,000	2,000	2,000	1,000 (or)	5,000
Beta	2,000	2,000	2,000	1,000 (or)	5,000
Gamma	2,000	2,000	2,000	1,000 (or)	5,000
Delta	2,000	2,000	2,000	1,000 (or)	5,000
Omicron	2,000	2,000	2,000	1,000 (or)	5,000
Other Taxa	0	0	0	0	3,640
Total Number	10,000	10,000	10,000	1,000	28,640

Table 1: Data selection for the C-VAE model, generated and calculated appearance rate of forward primers.

The study analyzed 610,000 complete SARS-CoV-2 genome sequences obtained from the GISAID and NCBI repositories. To ensure balanced representation across variants, we performed stratified random sampling from GISAID and selected 4,000 sequences per variant. These 20,000 GISAID sequences were then split into a training set (2,000 per variant; 10,000 total) and a validation set (2,000 per variant; 10,000 total), yielding an equal class distribution (1:1:1:1:1) in both splits (Table 1). During training, we used a mini-batch size of 50, corresponding to 200 training batches per epoch for the 10,000-sequence training set, which helps accommodate memory constraints [36] and reduces the impact of ordering effects during optimization [37]. We additionally constructed an independent test set composed exclusively of NCBI sequences (2,000 per variant; 10,000 total), with lineage labels assigned using the PANGOLIN classification system [38].

The proposed C-VAE achieved strong discriminative performance, with classification accuracy exceeding 98% across the five SARS-CoV-2 variants on both the validation set and the independent test set. Figure 8 summarizes the results using a confusion matrix. Figure 9 visualizes the embeddings from the final network layer using a t-SNE projection [39].

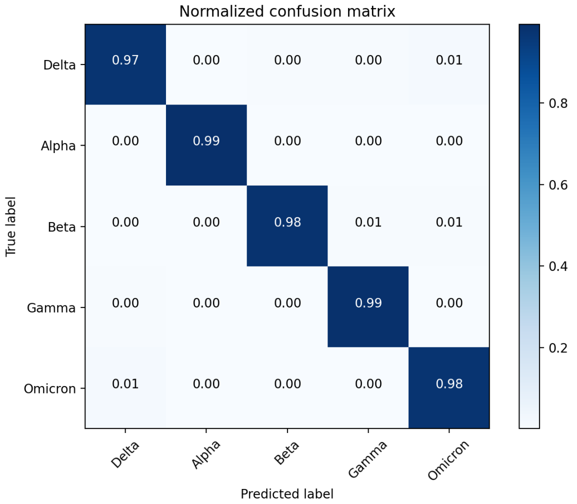


Figure 8: Confusion matrix for five-class SARS-CoV-2 variant classification using the C-VAE model (2,000 sequences per variant in Test set).

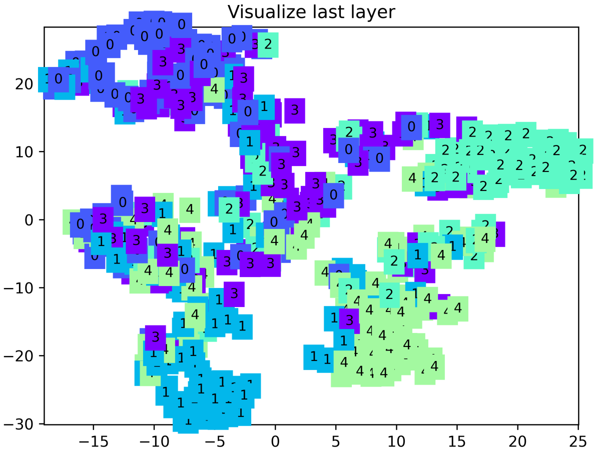


Figure 9: t-SNE visualization of the final-layer embeddings for five SARS-CoV-2 variants. Each point represents one sequence and is colored by variant label.

The standardized genome sequences were processed by the encoder's convolutional layers, producing activation maps for 12 filters. After comparing four feature extraction strategies, we selected the Pooling method due to its favorable computational efficiency while maintaining primer discovery performance. Primer appearance rates were quantified using the datasets summarized in Table 1 and Appendix Table 8. The resulting forward primers showed high target specificity: they appeared in more than 95% of sequences from the target variant and in less than 5% of sequences from non-target variants. Omicron was a notable exception, with an appearance rate of approximately 80% in target sequences and below 20% in other variants, consistent with its higher genetic diversity and ongoing evolution [35]. For Omicron, we therefore applied a relaxed appearance-rate threshold to retain sufficiently specific primers without resorting to lineage-specific primer design.

Detailed results are reported in Table 10 (the *Homo sapiens* genome), Table 11 (non-*Homo sapiens* hosts), and Table 12 (other taxa; Appendix). These validated forward primers were then used as anchors for downstream-region extraction in Stage III, as illustrated in Appendix Figure 14. For the complete pipeline, including *in silico* PCR validation and reverse primer design, see Figure 1 and Appendix Figure 15, respectively. After filtering candidates by standard primer design constraints and appearance-frequency thresholds, we obtained variant-specific forward primers: 66 (Alpha), 23 (Beta), 59 (Gamma), 52 (Delta), and 69 (Omicron).

For reverse primer design, we adapted the C-VAE pipeline to account for the downstream-sequence setting and the additional constraints of reverse primers. Specifically, we trained an independent C-VAE model for each validated forward primer and compared four feature extraction strategies (Pooling, Top-k, Mix, and Reconstruction). Performance was evaluated by the number of viable reverse primer candidates (and primer pairs) produced after applying the downstream filtering and screening criteria. The effectiveness of these strategies varied by variant. For the Alpha variant (Appendix Table 13), the Top-k, Mix, and Reconstruction methods produced substantially more viable candidates than Pooling. For the Delta variant (Appendix Table 14), the Top-k method yielded the highest extraction efficiency. Based on these evaluations, we selected the Top-k method as the default strategy for reverse primer extraction. We then computed the occurrence-frequency distributions of the resulting reverse primers for each variant (Appendix Table 15).

Final primer-pair selection was performed using thermodynamic constraints and complementarity screening. We first filtered candidates by GC content and melting temperature (Tm). We then removed primers with strong self-complementarity (self-dimers) and excluded forward–reverse combinations with substantial cross-complementarity (heterodimers). Finally, we required the Tm difference between paired primers to be within 5°C to support robust amplification. Applying these criteria yielded 1,478 computationally validated primer pairs across the five variants, summarized in Table 2.

	Forward Primer Number	Reverse Primer Number	Amplicon Size <200 bp	Amplicon Size <500 bp	Amplicon Size <1,000 bp
Alpha	66	400	6	14	66
Beta	23	18	0	0	1
Gamma	59	272	0	49	66
Delta	52	457	33	106	154
Omicron	69	331	23	26	50

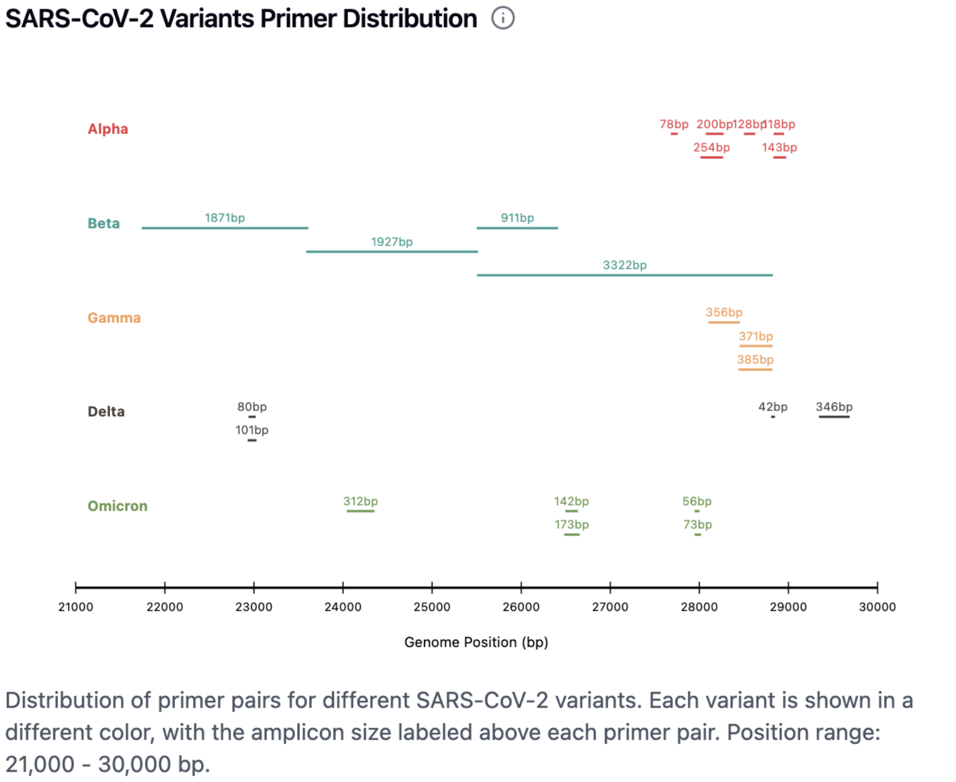
Table 2: Numbers of generated forward and reverse primers and computationally validated primer pairs by amplicon-length threshold for each SARS-CoV-2 variant. Reverse primers are generated conditional on each validated forward primer, and primer pairs are formed only from compatible forward–reverse combinations.

Primer-BLAST After computational design and thermodynamic screening, we performed Primer-BLAST [9] as an additional specificity check within SARS-CoV-2. Specifically, Primer-BLAST searches were restricted to SARS-CoV-2 sequences only. This step was used to verify that each primer pair maps to the intended locus and does not produce unintended matches elsewhere within SARS-CoV-2 genomes. Cross-reactivity to non-SARS-CoV-2 taxa was assessed separately via the appearance-frequency analysis described above. The detailed search settings and representative alignment outputs are provided in Appendix Figures 16 and 17, respectively.

***in silico* PCR** Primer pairs that passed all design and specificity criteria were then evaluated using *in silico* PCR on two independent platforms: FastPCR [29] and Unipro UGENE [30]. Candidate forward and reverse primers were assessed in batch mode in both tools. The results indicate that primer pairs produced by our pipeline amplify the intended variant-specific regions with high specificity. The *in silico* PCR outputs include genomic binding coordinates, melting temperature (Tm), predicted complementarity, and expected amplicon length. Representative results are shown in Appendix Figures 18 and 19. Based on these validation steps, we report a curated set of 22 primer pairs in Table 3, with their genomic binding positions summarized in Figure 10.

Primers (5' to 3')	GC content	Tm (°C)	Position	Amplicon Size
Alpha Variant				
F - AGGAGCTATAAAATCAGCACC	42.86%	49.60	27680->27700	78 bps
R - TCGATGCACTGAATGGGTGAT	47.62%	53.89	27737<-27757	
F - TCAACTCCAGGCAGCAGTAAAC	50.00%	54.93	28834->28855	118 bps
R - CAAACATTTTGCTCTCAAGCTG	40.91%	51.34	28930<-28951	
F - TTCAACTCCAGGCAGCAGTAA	47.62%	52.40	28500->28520	128 bps
R - GGCCTTTACCAAACATTTTGC	42.86%	50.45	28607<-28627	
F - AATTCAACTCCAGGCAGCAGTAAAC	44.00%	56.04	28830->28855	143 bps
R - CCTTGTTGTTGTTGGCCTTTACCAA	44.00%	56.60	28948<-28973	
F - CCATTCAGTGCATCGATATCGG	50.00%	53.59	28073->28097	200 bps
R - CTGATTTTGGGGTCCATTTAGA	40.91%	50.11	28251<-28272	
F - GAGCTATAAAATCAGCACC	42.11%	45.40	28012->28031	254 bps
R - TTGGGGTCCATTTAGAGACAT	42.86%	50.31	28245<-28266	
Beta Variant				
F - TCATAGCGCTTCCAAAATC	42.11%	47.81	25503->25522	911 bps
R - AGACCAGAAGATCAAGAACTCTAG	41.67%	51.29	26390<-26414	
F - GTTTGCTAACCCTGTCCTACCAT	47.83%	54.33	21740->21762	1,871 bps
R - CTACACCAAGTGACATAGTGTAG	43.48%	50.36	23588<-23610	
F - CTACACTATGTCACTTGGTGTA	40.91%	49.16	23588->23609	1,927 bps
R - AAGCGCTATGAAAAACAGCAAG	40.91%	52.69	25493<-25514	
F - TCATAGCGCTTCCAAAATC	42.11%	47.81	25504->25522	3,322 bps
R - CTACTGCTGCCTGGAGTTG	57.89%	52.15	28807<-28825	
Gamma Variant				
F - GCCAGAAACCTAAATTGGGTA	42.86%	49.96	28102->28122	356 bps
R - CATCTCGACTGCTATTGGTGT	47.62%	52.05	28437<-28457	
F - CGAGATGACCAAATTGGCTAC	47.62%	51.34	28451->28471	371 bps
R - TTAGAGCTGCCTGGAGTTGAA	47.62%	53.08	28801<-28821	
F - ACACCAATAGCAGTCGAGATG	47.62%	52.05	28437->28457	385 bps
R - TTAGAGCTGCCTGGAGTTGAA	47.62%	53.08	28801<-28821	
Delta Variant				
F - TCAACTCCAGGCAGCAGTATG	52.38%	54.20	28807->28827	42 bps
R - CATTCTAGCAGGAGAAGTTCC	47.62%	50.00	28828<-28848	
F - GGTAGCAAACCTTGTAATGGT	42.86%	50.30	22940->22960	80 bps
R - CCATTAGTGGGTTGGAAACCA	47.62%	52.19	22999<-23019	
F - TCTATCAGGCCGGTAGCAAAC	52.38%	54.02	22929->22949	101 bps
R - GTAACCAACACCATTAGTGGG	47.62%	50.72	23009<-23029	
F - AGGCTTATGAAACTCAAGCCT	42.86%	51.34	29342->29362	346 bps
R - AGTGGCCTCGGTGAAAATGTG	52.38%	55.31	29667<-29687	
Omicron Variant				
F - CACTCCGCATTACGTTTGGTG	52.38%	54.48	27946->27966	56 bps
R - ACCATTCTGGTTACTGCCAGT	47.62%	53.32	27981<-28001	
F - ACTCCGCATTACGTTTGGTGG	52.38%	55.42	27947->27967	73 bps
R - TTGTTTTGATCGCGCCCCACC	57.14%	58.53	27999<-28019	
F - CTCCTTGAAGAATGGAACCT	45.00%	48.79	26495->26515	142 bps
R - TTAAAGTTACTGGCCATAACAGCC	41.67%	53.36	26613<-26637	
F - GAGCTTAAAAAGCTCCTTGAAG	40.91%	49.90	26483->26505	173 bps
R - GCAGCAAGCACAAAACAAGTT	42.86%	53.28	26635<-26656	
F - GTGCACAAAAGTTTAACGGCCT	45.45%	52.38	24042->24063	312 bps
R - TATGGTTGACCACATCTTGAAG	40.91%	50.23	24332<-24353	

Table 3: Primer pairs successfully validated via in-silico PCR for each SARS-CoV-2 virus variant detection.


SARS-CoV-2 Variants Primer Distribution

Distribution of primer pairs for different SARS-CoV-2 variants. Each variant is shown in a different color, with the amplicon size labeled above each primer pair. Position range: 21,000 - 30,000 bp.

Figure 10: Visualization of primer pair binding positions in the target genome sequence.

3.2 Experiment 2: Primer Design for *E. coli* and *S. flexneri*

Escherichia coli (*E. coli*) and *Shigella flexneri* (*S. flexneri*) are clinically important bacterial species with distinct implications for human health. *E. coli* is commonly a commensal member of the human intestinal microbiome, whereas *S. flexneri* is an established pathogen associated with shigellosis and a substantial global burden of gastrointestinal disease and foodborne transmission [40,41]. Rapid and reliable identification of these organisms is therefore important for clinical decision-making and public health surveillance.

Although PCR-based assays are widely used to detect *E. coli* and *S. flexneri* due to their high analytical sensitivity and specificity [42,43], designing primers that robustly discriminate between these two species remains challenging. This difficulty is driven by substantial genomic similarity between the species, together with considerable within-species genetic diversity. To address this discrimination problem, we adapted Primer C-VAE for differential detection of *E. coli* and *S. flexneri* using full-length genomes.

Our analysis included 8,939 complete genome sequences for *E. coli* and 5,373 for *S. flexneri*. For model development and evaluation, we constructed balanced datasets from NCBI by sampling 1,000 sequences per species for training, 1,000 per species for validation, and 1,000 per species for testing (Table 4). Using this setup, the optimised C-VAE achieved classification accuracy exceeding 97% on both the validation set and the independent test set. Detailed performance metrics are reported in Table 4, and the corresponding confusion matrix is shown in Figure 11.

	Training set	Validation set	Test set	Generated primers	Calculated appearance
Source	NCBI	NCBI	NCBI	NCBI	NCBI
E. coli	1,000	1,000	1,000	800 (or)	1,500
S. flexneri	1,000	1,000	1,000	800 (or)	1,500
Total Number	2,000	2,000	2,000	1,600	3,000

Table 4: Data selection for the C-VAE model and datasets used to generate and compute appearance rates of forward primers for *E. coli* and *S. flexneri*.

However, because most *E. coli* strains are non-pathogenic commensals and only a subset of lineages are associated with virulence, we prioritized primer-pair reporting for *S. flexneri* in this experiment. The genomes of *E. coli* and *S. flexneri* are approximately 4.5–5.5 Mb, which is substantially larger than the SARS-CoV-2 genome (˜30 kb). This increase in sequence length and genomic complexity poses additional computational

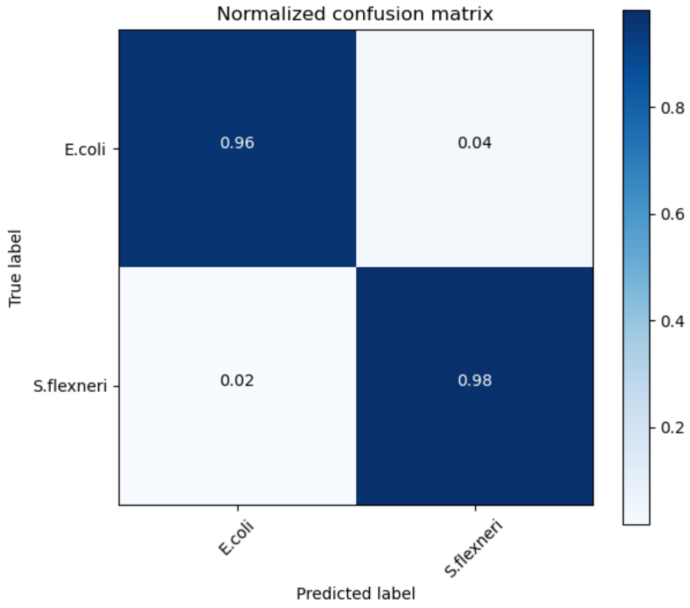
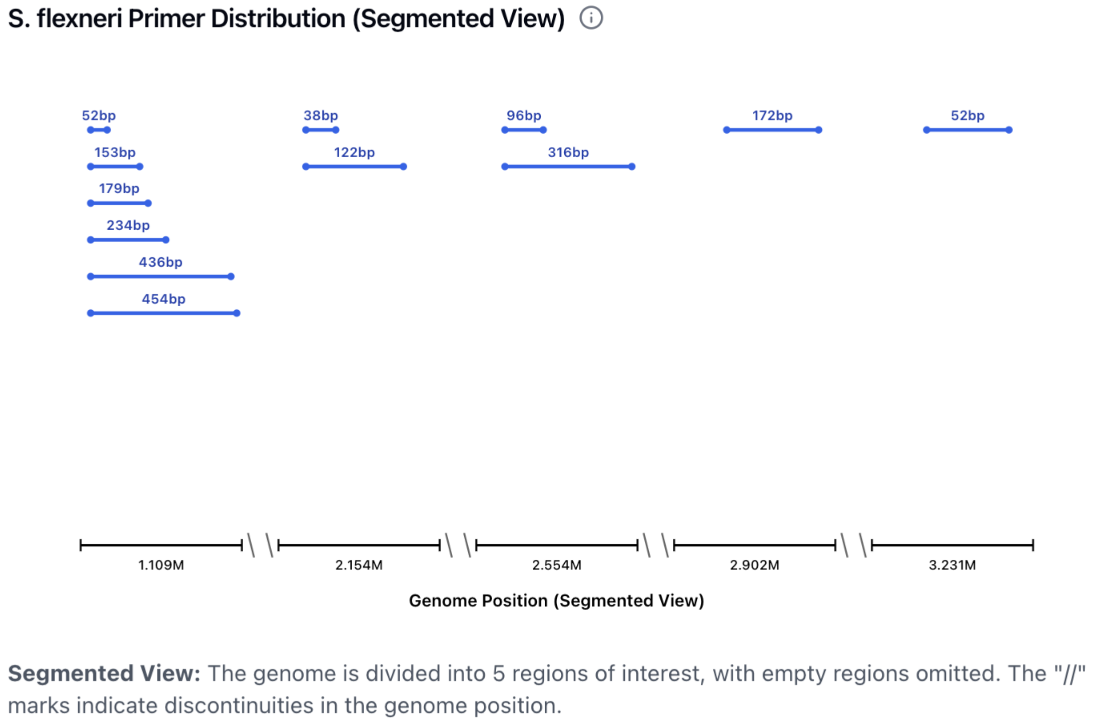


Figure 11: Confusion matrix from cross-validation for the C-VAE model based on 1,000 sequences of each E. coli and S. flexneri.

demands for our C-VAE implementation, requiring greater memory and longer processing time than the SARS-CoV-2 experiments. Despite these challenges, the proposed pipeline produced highly specific primer candidates. The optimised *S. flexneri* primer pairs are reported in Table 5, and their genomic binding coordinates are summarized in Figure 12.

Primers (5' to 3')	GC content	Tm (°C)	Position	Amplicon Size
F -GAGCTGATGGCTTCATCCAGA	52.38%	57.74	2153834->2153854	38 bps
R - GCCGATCCCCTGAAAGC	64.71%	55.63	2153855<-2153871	
F - GTGACGCTGTAGATGATACGT	47.62%	55.53	3230534->3230554	52 bps
R - AAAACCGTCTGAAAAGCCGCA	47.62%	59.49	3230565<-3230585	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	52 bps
R - TATTGCTGGGTAATCAGGCGT	47.62%	57.38	1109062<-1109082	
F - GCGCGGTTTTAATGAAGAAGA	42.86%	55.39	2554171->2554191	96 bps
R - TCACGACGCATCAGATGATGC	52.38%	59.02	2554246<-2554266	
F - GAGCTGATGGCTTCATCCAGA	52.38%	57.75	2153834->2153854	122 bps
R - GTGACTAACGGCAGCGGTAAG	57.14%	59.23	2153935<-2153955	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	153 bps
R - AGCGATCCTTGATAAAGCAGG	47.62%	56.02	1109163<-1109183	
F - AGCTCAAGCAACAATTTACGC	42.86%	55.92	2902098->2902118	172 bps
R - TTAAAGCTCTTGCCGCAGAGG	52.38%	58.83	2902249<-2902269	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	179 bps
R - CAAGGTTCCAGCCATCCATTG	52.38%	57.41	1109190<-1109209	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	234 bps
R - AACATTTGCTGATGTTGACGA	38.1%	54.57	1109244<-1109264	
F - GCGCGGTTTTAATGAAGAAGA	42.86%	55.39	2554171->2554191	316 bps
R - TTGTGCCTGTAATGTGGTGCC	52.38%	59.31	2554466<-2554486	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	436 bps
R - CTACGGTGCTGATTATCGCCT	52.38%	57.89	1109446<-1109466	
F - GCTTTAGTATCGACTTGCTGA	42.86%	53.67	1109031->1109051	454 bps
R - ATACCCTGGTGATTGCCACTA	47.62%	56.38	1109464<-1109484	

Table 5: Primer pairs selected from S. flexneri with amplicon sizes ranging from 0-500 bps, sorted by amplicon size.



Segmented View: The genome is divided into 5 regions of interest, with empty regions omitted. The "//" marks indicate discontinuities in the genome position.

Figure 12: Visualization of primer pair binding positions in the target genome sequence.

4 Discussion

We propose a VAE-based deep-learning workflow for flexible-length primer design that supports target-specific detection across diverse organisms. The approach addresses a common practical challenge in primer development: identifying highly discriminative primer regions in settings with substantial genetic heterogeneity, such as SARS-CoV-2. In current practice, variant-oriented primer design for SARS-CoV-2 often relies on manual screening of large collections of full-length genomes to identify variant-defining alterations (e.g., deletions [44] or characteristic mutations [45, 46]). Such workflows are time-consuming and typically require substantial domain expertise. In addition, widely used automated tools such as Primer3 and Primer3Plus [10, 11] impose practical constraints for long-genome applications (e.g., a maximum input length below 10,000 bp; Appendix Table 16). These constraints limit direct application to organisms with large genomes, including *E. coli* and *S. flexneri* (4.5–5.5 Mb), and may require truncation or preprocessing that can discard potentially informative regions. By contrast, Primer C-VAE is designed to reduce manual screening effort and support long sequences. As shown in Table 16 (Delta example: hCoV-19/Indonesia/JK-GS-FKUINIHRD-0489/2022), our workflow improves the efficiency of identifying candidate primer regions from complete genomes.

Across our experiments, *in silico* PCR validation indicates that primer pairs produced by the proposed pipeline achieve high specificity for the intended targets. The C-VAE encoder learns discriminative sequence representations that support sequence classification and guide primer candidate extraction, while the decoder provides an auxiliary reconstruction objective that regularizes representation learning. Differences between the input and reconstructed sequences can also help highlight regions that vary systematically across classes, offering a degree of interpretability that is useful for downstream primer design. Together, these components demonstrate a practical integration of deep learning and primer design workflows for variant- and species-level discrimination.

Several limitations and opportunities for improvement remain. First, the current pipeline does not incorporate degenerate bases, which could improve robustness to within-class variability and represents an important extension for future work. Second, although we apply thermodynamic and complementarity screening (e.g., GC content, melting temperature, and dimer checks) prior to *in silico* PCR, these steps still require multiple filters and tool-based evaluations; further automation and tighter integration of these criteria could streamline the workflow. Third, the computational cost of training remains non-trivial, particularly for reverse primer design where separate models are trained conditional on each validated forward primer and additional hyperparameter tuning may be required. Fourth, a limitation of this study is the absence of wet-lab experimental validation. Although our *in silico* PCR validation using multiple established tools (FastPCR, Unipro UGENE, and Primer-BLAST) provides strong computational evidence for primer specificity, experimental confirmation of amplification performance remains an important step before clinical or field deployment; future work will focus on laboratory validation of selected primer pairs. Finally, the method depends on having sufficient genome sequences for training; for organisms with limited publicly available complete genomes (e.g., Human astrovirus, HAstV), direct application may be constrained. Addressing these challenges—for example through more efficient training strategies, stronger sharing of information across primer-specific models, and data-efficient learning—is a promising direction for future research.

Acknowledgement

The work of Hanyu Wang is funded by China Scholarship Council with Reference 202308060184. The work of Anthony J. Dunn is jointly funded by Decision Analysis Services Ltd and EPSRC through the studentship with Reference EP/R513325/1. The work of Alain B. Zemkoho is supported by the EPSRC grant EP/V049038/1 and The Alan Turing Institute under the EPSRC grants EP/N510129/1 and EP/W037211/1.

Conflict of interest statement

The authors declare that there is no conflict of interest.

5 Hardware and Software Environments

All experiments were implemented in Python and evaluated in the following hardware and software environments.

Windows/Linux (GPU-accelerated)

- Operating system: Windows 11 Pro 23H2 and Ubuntu 22.04.5 LTS
- CPU: 13th Gen Intel(R) Core(TM) i7-13700 @ 2.10 GHz
- RAM: 32.0 GB
- GPU: NVIDIA GeForce RTX 4070 Ti
- Python: 3.9.7 (64-bit)
- PyTorch: 2.5.0 with CUDA 12.4
- IDE: PyCharm 2024.2.3 (Professional Edition)

macOS (CPU-only)

We additionally tested the pipeline on macOS without GPU acceleration (CPU-only execution).

- Operating system: macOS Sonoma 14.4
- CPU: Apple M1 (8 cores: 4 performance + 4 efficiency)
- GPU: Apple M1 7-core GPU
- RAM: 8.0 GB
- Python: 3.12.4
- PyTorch: 2.5.0
- IDE: PyCharm 2024.2.3 (Professional Edition)

Data availability

We gratefully acknowledge the authors responsible for obtaining the specimens and genetic sequence data generated and shared via the GISAID Initiative (https://doi.org/10.55876/gis8.220628xf), and that we used for the research presented in this paper. All the codes used and generated in the course of the work presented in this article are available in the following GitHub page: https://github.com/harrywang9917/Primer_C-VAE

References

[1] Stephen Bustin and Jim Huggett. qPCR primer design revisited. *Biomolecular Detection and Quantification*, 14:19–28, 2017.

[2] World Health Organization. Genomic sequencing of SARS-CoV-2: a guide to implementation for maximum impact on public health. https://www.who.int/publications/i/item/9789240018440, 2021. Accessed: 2024-10-11.

[3] Illumina, Inc. Key differences between next-generation sequencing and sanger sequencing. https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/advantages/ngs-vs-sanger.html, Oct 2024. Accessed: 2024-10-11.

[4] Illumina, Inc. Advantages of next-generation sequencing vs. qPCR. https://www.illumina.com/science/technology/next-generation-sequencing/beginners/advantages/ngs-vs-qpcr.html, Oct 2024. Accessed: 2024-10-11.

[5] Theodore Johnson, Tanner Bishoff, Kaleb Kremsreiter, Austin Lebanc, and Macario Camacho. Diagnostic testing for COVID-19: systematic review of meta-analyses and evidence-based algorithms. *The Medical Journal*, (PB 8-21-01/02/03):50–59, 2021.

[6] Todd C Lorenz. Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. *Journal of Visualized Experiments*, (63):e3998, 2012.

[7] Stephen Bustin, Reinhold Mueller, and Tania Nolan. Parameters for successful PCR primer design. In *Methods in Molecular Biology*, volume 2065, pages 5–22, 2020.

[8] Matej Lexa, Jakub Horak, and Bretislav Brzobohaty. Virtual PCR. *Bioinformatics*, 17(1):192–193, 2001.

[9] 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):1–11, 2012.

[10] Triinu Koressaar and Maido Remm. Enhancements and modifications of primer design program Primer3. *Bioinformatics*, 23(10):1289–1291, 2007.

[11] 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):e115–e115, 2012.

[12] Daniel Ashlock, Kenneth Bryden, Steven Corns, Patrick Schnable, and Tsui-Jung Wen. Training finite state classifiers to improve PCR primer design. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, page 4385, 2004.

[13] Jain-Shing Wu, Chungnan Lee, Chien-Chang Wu, and Yow-Ling Shiue. Primer design using genetic algorithm. *Bioinformatics*, 20(11):1710–1717, 2004.

[14] Alejandro Lopez Rincon, Carmina A Perez Romero, Lucero Mendoza Maldonado, Eric Claassen, Johan Garssen, Aletta D Kraneveld, and Alberto Tonda. Design of specific primer sets for SARS-CoV-2 variants using evolutionary algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 982–990, 2021.

[15] Alejandro Lopez-Rincon, Alberto Tonda, Lucero Mendoza-Maldonado, Daphne GJC Mulders, Richard Molenkamp, Carmina A Perez-Romero, Eric Claassen, Johan Garssen, and Aletta D Kraneveld. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports*, 11(1):1–11, 2021.

[16] Yuelong Shu and John McCauley. GISAID: global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), 2017.

[17] National Center for Biotechnology Information Bethesda (MD): National Library of Medicine (US). National center for biotechnology information (ncbi)[internet]. Available from: https://www.ncbi.nlm.nih.gov/, 1988. Accessed: 2024-10-06.

[18] Jinny X Zhang, Boyan Yordanov, Alexander Gaunt, Michael X Wang, Peng Dai, Yuan-Jyue Chen, Kerou Zhang, John Z Fang, Neil Dalchau, Jiaming Li, et al. A deep learning model for predicting next-generation sequencing depth from DNA sequence. *Nature Communications*, 12(1):1–10, 2021.

[19] Allen Chieng Hoon Choong and Nung Kion Lee. Evaluation of Convolutionary Neural Networks Modeling of DNA Sequences using Ordinal versus one-hot Encoding Method. *bioRxiv*, page 186965, 2017.

[20] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[21] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Jonathon Shlens. Notes on Kullback-Leibler Divergence and likelihood. *arXiv preprint arXiv:1404.2000*, 2014.

[24] CW Dieffenbach, TM Lowe, and GS Dveksler. General concepts for PCR primer design. *PCR Methods and Applications*, 3(3):S30–S37, 1993.

[25] Addgene: how to design primers. https://www.addgene.org/protocols/primer-design/, 2019. Accessed: 2024-10-30.

[26] Rosario San Millán Joseba Bikandi. Melting temperature (tm) calculation. http://insilico.ehu.es/tm.php?formula=basic, July 2015. Accessed: 2024-10-30.

[27] Integrated DNA Technologies. OligoAnalyzer Tool - Primer analysis and Tm calculator. https://eu.idtdna.com/pages/tools/oligoanalyzer. Accessed: 2024-10-22.

[28] Livia Schrick and Andreas Nitsche. Pitfalls in PCR troubleshooting: expect the unexpected? *Biomolecular Detection and Quantification*, 6:1–3, 2016.

[29] Ruslan Kalendar, Bekbolat Khassenov, Yerlan Ramankulov, Olga Samuilova, and Konstantin I Ivanov. FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics*, 109(3-4):312–319, 2017.

[30] Konstantin Okonechnikov, Olga Golosova, Mikhail Fursov, and UGENE team. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics*, 28(8):1166–1167, 2012.

[31] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, and Guo-Wei Wei. Emerging vaccine-breakthrough SARS-CoV-2 variants. *ACS Infectious Diseases*, 8(3):546–556, 2022. PMID: 35133792.

[32] Paul A Christensen, Randall J Olsen, S. Wesley Long, Sishir Subedi, James J Davis, Parsa Hodjat, Debbie R Walley, Jacob C Kinskey, Matthew Ojeda Saavedra, Layne Pruitt, Kristina Reppond, Madison N Shyer, Jessica Cambric, Ryan Gadd, Rashi M Thakur, Akanksha Batajoo, Regan Mangham, Sindy Pena, Trina Trinh, Prasanti Yerramilli, Marcus Nguyen, Robert Olson, Richard Snehal, Jimmy Gollihar, and James M Musser. Delta variants of SARS-CoV-2 cause significantly increased vaccine breakthrough COVID-19 cases in Houston, Texas. *The American Journal of Pathology*, 192(2):320–331, 2022.

[33] Xuemei He, Weiqi Hong, Xiangyu Pan, Guangwen Lu, and Xiawei Wei. SARS-CoV-2 Omicron variant: Characteristics and prevention. *MedComm*, 2(4):838–845, 2021.

[34] World Health Organization. COVID-19 cases — WHO COVID-19 dashboard. https://data.who.int/dashboards/covid19/cases, Oct 2024. Accessed: 2024-10-11.

[35] Prerna Arora, Christine Happle, Amy Kempf, Inga Nehlmeier, Metodi V Stankov, Alexandra Dopfer-Jablonka, Georg M N Behrens, Stefan Pöhlmann, and Markus Hoffmann. Impact of JN.1 booster vaccination on neutralisation of SARS-CoV-2 variants KP.3.1.1 and XEC. *bioRxiv*, 2024.10.04.616448, 2024. Available from: https://www.biorxiv.org/content/10.1101/2024.10.04.616448v1.

[36] Fengxiang He, Tongliang Liu, and Dacheng Tao. Control batch size and learning rate to generalize well: theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, 32:1143–1152, 2019.

[37] Tal Schuster, Darsh J Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. *arXiv preprint arXiv:1908.05267*, 2019.

[38] Áine O'Toole, Emily Scher, Anthony Underwood, Ben Jackson, Verity Hill, John T McCrone, Rachel Colquhoun, Chris Ruis, Khalil Abu-Dahab, Ben Taylor, Corin Yeats, Louis du Plessis, Daniel Maloney, Nathan Medd, Stephen W Attwood, David M Aanensen, Edward C Holmes, Oliver G Pybus, and Andrew Rambaut. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2), 2021. veab064.

[39] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems*, 15:1–8, 2002.

[40] Matthew A Croxen, Robyn J Law, Roland Scholz, Kristie M Keeney, Marta Wlodarska, and B. Brett Finlay. Recent advances in understanding enteric pathogenic Escherichia coli. *Clinical Microbiology Reviews*, 26(4):822–880, 2013.

[41] Elaine Scallan, Robert M Hoekstra, Frederick J Angulo, Robert V Tauxe, Marc-Alain Widdowson, Sharon L Roy, Jeffery L Jones, and Patricia M Griffin. Foodborne illness acquired in the United States–major pathogens. *Emerging Infectious Diseases*, 17(1):7–15, 2011.

[42] Dirk Wildeboer, Linda Amirat, Robert G Price, and Ramadan A Abuknesha. Rapid detection of Escherichia coli in water using a hand-held fluorescence detector. *Water Research*, 44(8):2621–2628, 2010.

[43] Ying Chen, Linyan Zhang, Ling Xu, Xinjian Guo, Huan Yang, Linlin Zhuang, Ying Li, Zhenzhen Wang, and Bing Gu. Rapid and sensitive detection of Shigella flexneri using fluorescent microspheres as label for immunochromatographic test strip. *Annals of Translational Medicine*, 7(20):565, 2019.

[44] Chantal BF Vogels, Mallery I Breban, Isabel M Ott, Tara Alpert, Mary E Petrone, Anne E Watkins, Chaney C Kalinich, Rebecca Earnest, Jessica E Rothman, Jaqueline Goes de Jesus, Ingra Morales Claro, Giulia Magalhães Ferreira, Myuki AE Crispim, Brazil-UK CADDE Genomic Network, Lavanya Singh, Houriiyah Tegally, Ugochukwu J Anyaneji, Network for Genomic Surveillance in South Africa, Emma B Hodcroft, Christopher E Mason, Gaurav Khullar, Jessica Metti, Joel T Dudley, Matthew J MacKay, Megan Nash, Jianhui Wang, Chen Liu, Pei Hui, Steven Murphy, Caleb Neal, Eva Laszlo, Marie L Landry, Anthony Muyombwe, Randy Downing, Jafar Razeq, Tulio de Oliveira, Nuno R Faria, Ester C Sabino, Richard A Neher, Joseph R Fauver, and Nathan D Grubaugh. Multiplex qPCR discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biology*, 19(5):e3001236, 2021.

[45] Mamdouh Sibai, Hannah Wang, Priscilla SW Yeung, Malaya K Sahoo, Daniel Solis, Kenji O Mfuh, Chun-Hong Huang, Fumiko Yamamoto, and Benjamin A Pinsky. Development and evaluation of an RT-qPCR for the identification of the SARS-CoV-2 Omicron variant. *Journal of Clinical Virology*, 148:105101, 2022.

[46] Hannah Wang, Jacob A Miller, Michelle Verghese, Mamdouh Sibai, Daniel Solis, Kenji O Mfuh, Becky Jiang, Naomi Iwai, Marilyn Mar, ChunHong Huang, Fumiko Yamamoto, Malaya K Sahoo, James Zehnder, and Benjamin A Pinsky. Multiplex SARS-CoV-2 genotyping reverse transcriptase PCR for population-level variant screening and epidemiologic surveillance. *Journal of Clinical Microbiology*, 59(8):e00859–21, 2021.

Appendix A Data collection and pre-processing

This appendix provides supplementary details on data collection and preprocessing. Table 6 summarizes the SARS-CoV-2 variant metadata used in this study, including WHO labels, Pango lineages, GISAID clades, sample counts, and the corresponding class labels used in our experiments. Table 7 reports the average genome length for each SARS-CoV-2 variant. Figure 13 illustrates an example of sequence standardization and encoding. Table 8 lists the SARS-CoV-2 and non-SARS-CoV-2 coronavirus sequences collected from GISAID and NCBI that were used to compute the appearance rates of forward and reverse primers.

WHO label	Pango lineage	GISAID clade	Number of samples	Label
Alpha	B.1.1.7+Q.*	GRY	119,077	1
Beta	B.1.351	GH/501Y.V2	27,782	2
Gamma	P.1	GR/501Y.V3	48,588	3
Delta	B.1.617.2+AY.*	G/478K.V1	142,815	4
Omicron	B.1.1.529+BA.*	GR/484A	135,383	0

Table 6: SARS-CoV-2 variants collected from GISAID for C-VAE model development.
Note: SARS-CoV-2 variants are described under multiple nomenclature systems. This table reports the WHO label, Pango lineage, and GISAID clade for each variant, together with the sample counts and the class labels used in this study.

Variant	Average length (bp)	Label
Alpha	29,769	1
Beta	29,774	2
Gamma	29,770	3
Delta	29,766	4
Omicron	29,748	0

Table 7: Average genome sequence length for each SARS-CoV-2 variant.
Note: The average genome lengths of the five variants are very similar. In our dataset, the longest sequence was a Delta genome with length 31,079 bp.

Example of sequence standardization and ordinal encoding:

1) Original sequences with different lengths:
5'- AGTCAGCATCTCATGTGCGAGTCCTGACGCTGACTAGC -3' (38 bp)
5'- ATCTCATGTGCGAACGCTGACTAGAAAATCCAAAAAANNNNNNA -3' (45 bp)
5'- CGCTGACTAGAAAATCCAAAAAANNNCGTTTACTTCGANNN -3' (41 bp)
5'- NNNNAGTCAGCATCTCATGTGTCCTGACGCTGACTAG -3' (37 bp)

2) Standardized sequences (padded with N to the maximum length):
5'- AGTCAGCATCTCATGTGCGAGTCCTGACGCTGACTAGCNNNNNNN -3' (45 bp)
5'- ATCTCATGTGCGAACGCTGACTAGAAAATCCAAAAAANNNNNNA -3' (45 bp)
5'- CGCTGACTAGAAAATCCAAAAAANNNCGTTTACTTCGANNNNNNN -3' (45 bp)
5'- NNNNAGTCAGCATCTCATGTGTCCTGACGCTGACTAGNNNNNNNN -3' (45 bp)

3) Ordinal encoding:
Encoded: 4 3 2 1 4 3 1 4 2 1 2 1 4 2 3 2 3 1 3 4 3 2 1 1 2 3 4 1 3 1 2 3 4 1 2 4 3 1 0 0 0 0 0 0 0
Encoded: 4 2 1 2 1 4 2 3 2 3 1 3 4 4 1 3 1 2 3 4 1 2 4 3 4 4 4 4 2 1 1 4 4 4 4 4 4 0 0 0 0 0 0 0 4
Encoded: 1 3 1 2 3 4 1 2 4 3 4 4 4 4 2 1 1 4 4 4 4 4 4 0 0 0 1 3 2 2 2 4 1 2 2 1 3 4 0 0 0 0 0 0 0
Encoded: 0 0 0 0 4 3 2 1 4 3 1 4 2 1 2 1 4 2 3 2 3 2 1 1 2 3 4 1 3 1 2 3 4 1 2 4 3 1 0 0 0 0 0 0 0

Figure 13: Example of sequence standardization and ordinal encoding. Sequences are padded with "N" to a fixed length and then mapped to integers using A=4, G=3, T=2, C=1, and N=0.

Coronavirus species	**Source**	**Host**	**Number of samples**
SARS-CoV-2 (all five variants)	GISAID	*Homo sapiens*	58,547
SARS-CoV-2 (all five variants)	NCBI	*Homo sapiens*	78,692
SARS-CoV-2	GISAID	*Manis javanica*	19
SARS-CoV-2	GISAID	*Rhinolophus affinis*	1
SARS-CoV-2	GISAID	*Rhinolophus*	1
SARS-CoV-2	GISAID	*Canis*	29
SARS-CoV-2	GISAID	*Felis catus*	51
MERS-CoV	NCBI	*Homo sapiens*	738
HCoV-OC43	NCBI	*Homo sapiens*	1,311
HCoV-NL63	NCBI	*Homo sapiens*	634
HCoV-229E	NCBI	*Homo sapiens*	446
HCoV-HKU1	NCBI	*Homo sapiens*	404
SARS-CoV-P2	NCBI	*Homo sapiens*	1
SARS-CoV-HKU-39849	NCBI	*Homo sapiens*	2
SARS-CoV-GDH-BJH01	NCBI	*Homo sapiens*	1
HAstV-BF34	NCBI	*Homo sapiens*	2

Table 8: Coronavirus sequences used to compute primer appearance rates.
Note: In addition to SARS-CoV-2 variant genomes, we included SARS-CoV-2 sequences from non-human hosts and other coronavirus species to evaluate cross-reactivity when computing the appearance rates of generated primers.

Appendix B Feature extraction and evaluation

This appendix provides supplementary results for feature extraction and primer evaluation. Table 9 reports the effect of different Top-k values on the number of primer pairs obtained and the resulting amplicon-length ranges. Tables 10 and 11 summarize forward primer appearance rates evaluated on *Homo sapiens* and non-*Homo sapiens* hosts, respectively. Table 12 reports forward primer appearance rates across other coronavirus taxa. Tables 13 and 14 present the numbers of reverse primer candidates generated per forward primer under the four extraction methods (Pooling, Top-k, Mix, and Reconstruction) for the Alpha and Delta variants, as described in Section 2.3. Finally, Table 15 reports reverse primer appearance rates evaluated on *Homo sapiens* hosts, analogous to the forward-primer analysis in Table 10.

Variant / metric	**Top-75**	**Top-125**	**Top-175**	**Top-250**
Alpha: total primer pairs	60	94	136	171
Amplicon length < 1,000 bp	7	11	17	17
Amplicon length < 500 bp	7	10	16	15
Amplicon length < 300 bp	0	0	0	0
Amplicon length < 200 bp	0	0	0	0
Delta: total primer pairs	290	483	623	736
Amplicon length < 1,000 bp	0	1	4	2
Amplicon length < 500 bp	0	0	0	0
Amplicon length < 300 bp	0	0	0	0
Amplicon length < 200 bp	0	0	0	0

Table 9: Effect of the Top-k parameter on the number of generated primer pairs and their amplicon-length distribution for the Alpha and Delta variants.
Note: The performance of the Top-k feature extraction method depends on the choice of k.

Forward primer (5′–3′)	**SARS-CoV-2** (Alpha)	**SARS-CoV-2** (Beta)	**SARS-CoV-2** (Gamma)	**SARS-CoV-2** (Delta)	**SARS-CoV-2** (Omicron)
Dataset	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI
Host	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*
Number of sequences	5,000	5,000	5,000	5,000	5,000
Alpha Variant					
TACTAATGATAACACCTCAAG	0.9928	0.0038	0.0004	0.0	0.0
CAATTTGGCAGAGACATTGAT	0.9928	0.0004	0.0004	0.0	0.0
TCAAACTGTCAAACCTGGTAA	0.9926	0.001	0.0008	0.0002	0.0002
CTTTTCAAACTGTCAAACCTG	0.9924	0.001	0.0008	0.0002	0.0002
Beta Variant					
CGAACAAACTAAAATGTCTGA	0.0014	0.9832	0.0482	0.0304	0.0016
GCTTAGGGTTGATACAGCCAA	0.0142	0.9756	0.0134	0.0068	0.0056
TAGGGTTGATACAGCCAATCC	0.0142	0.975	0.0132	0.0068	0.0056
AGGGTTGATACAGCCAATCCT	0.0142	0.975	0.0132	0.0068	0.0054
Gamma Variant					
TGTGGTAAACAAGCTACACAA	0.0	0.0004	0.9958	0.0	0.0
GTGGTAAACAAGCTACACAAT	0.0	0.0004	0.9958	0.0	0.0
GTGTGGTAAACAAGCTACACA	0.0	0.0004	0.9954	0.0	0.0
ACACAATATCTAGTACAACAG	0.0002	0.0004	0.9934	0.0	0.0
Delta Variant					
GATACTAGTTTGTCTGGTTTT	0.0016	0.0382	0.0032	0.998	0.1762
AGTTTGTCTGGTTTTAAGCTA	0.0016	0.0382	0.0032	0.998	0.1766
TATGGTTGATACTAGTTTGTC	0.0046	0.0426	0.0082	0.9974	0.1764
TGGTTGATACTAGTTTGTCTG	0.0024	0.0408	0.0056	0.9972	0.1766
Omicron Variant					
GCGCTTCCAAAATCATAACTC	0.0004	0.0002	0.0002	0.0004	0.8344
TCACACCGGAAGCCAATATGG	0.0002	0.0	0.0	0.0002	0.833
AATAACAGTCACACCGGAAGC	0.0002	0.0	0.0	0.0002	0.8328
AGAGATAGGTACGTTAATAGT	0.0034	0.0004	0.0006	0.0002	0.8318

Table 10: Appearance frequencies of selected forward primers across SARS-CoV-2 variants, evaluated on 5,000 *Homo sapiens* host sequences per variant (combined GISAID and NCBI datasets).
Note: Values report the fraction of sequences in which each forward primer occurs within the corresponding variant-specific dataset. Primers are designed to be variant-discriminative; consequently, they show high occurrence in the target variant and low occurrence in non-target variants.

Forward Primer (5′–3′)	Alpha Pooling Method	Alpha Top method	Alpha Mix Method	Alpha Recon Method
TTGGCAGAGACATTGATGACA	30	96	58	63
TCTTATGGGTTGGGATTATCC	48	61	115	67
TTGCACGTCTTGACAAAGTTG	37	60	52	43
TCCTTGCACGTCTTGACAAAG	40	51	43	46
TGCACGTCTTGACAAAGTTGA	34	63	38	33
TTTGGCAGAGACATTGATGAC	47	84	56	56
CACAACACATTTGTGTCTGGT	26	53	38	28
CTTGCACGTCTTGACAAAGTT	42	57	49	50
CACACAACACATTTGTGTCTG	21	54	36	44
ACACAACACATTTGTGTCTGG	32	61	44	31
CCTTGCACGTCTTGACAAAGT	33	58	38	56
GCACGTCTTGACAAAGTTGAG	32	22	45	32
Average Number	35.166	52.000	51.000	45.75

Table 13: Numbers of reverse primer candidates generated under four feature extraction methods for each Alpha-variant forward primer.
Note: The number of reverse primer candidates varies with the feature extraction strategy. For each validated Alpha forward primer (left column), we report the number of reverse primer candidates produced by Pooling, Top-k, Mix, and Reconstruction.

Forward primer (5′–3′)	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2	SARS-CoV-2
Dataset	GISAID	GISAID	GISAID	GISAID	GISAID
Host	*Manis javanica*	*Rhinolophus affinis*	*Rhinolophus*	*Canis*	*Felis catus*
Number of sequences	19	1	1	29	51
CTCAGACTAATTCTCGTCGGC	0.0	0.0	0.0	0.0	0.0
ACTAATTCTCGTCGGCGGGCA	0.0	0.0	0.0	0.0	0.0
TCAGACTAATTCTCGTCGGCG	0.0	0.0	0.0	0.0	0.0
AGACTAATTCTCGTCGGCGGG	0.0	0.0	0.0	0.0	0.0
CAGACTAATTCTCGTCGGCGG	0.0	0.0	0.0	0.0	0.0
AACTCCAGGCAGCAGTATGGG	0.0	0.0	0.0	0.0	0.0
ACTCCAGGCAGCAGTATGGGA	0.0	0.0	0.0	0.0	0.0
CTCCAGGCAGCAGTATGGGAA	0.0	0.0	0.0	0.0	0.0
CCAGGCAGCAGTATGGGAACT	0.0	0.0	0.0	0.0	0.0
AGGCAGCAGTATGGGAACTTC	0.0	0.0	0.0	0.0	0.0

Table 11: Appearance frequencies of selected forward primers in SARS-CoV-2 sequences from non-*Homo sapiens* hosts (GISAID).

Note: This table reports appearance frequencies for a subset of Delta forward primers evaluated on SARS-CoV-2 genomes from non-human hosts (as listed in the header). The primers were designed using human-host Delta sequences, and no matches were observed in these non-human host datasets (all reported frequencies are 0.0).

Forward primer (5′–3′)	MERS-CoV	HCoV-OC43	HCoV-NL63	HCoV -229E	HCoV -HKU1	SARS-CoV -P2	SARS-CoV -HKU-39849	SARS-CoV -GDH-BJH01	HAstV -BF34
Dataset	NCBI	NCBI	NCBI	NCBI	NCBI	NCBI	NCBI	NCBI	NCBI
Host: HS (*Homo sapiens*)	HS	HS	HS	HS	HS	HS	HS	HS	HS
Sequence Number	738	1,311	634	446	404	1	2	1	2
CTCAGACTAATTCTCGTCGGC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ACTAATTCTCGTCGGCGGGCA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TCAGACTAATTCTCGTCGGCG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AGACTAATTCTCGTCGGCGGG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CAGACTAATTCTCGTCGGCGG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AACTCCAGGCAGCAGTATGGG	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ACTCCAGGCAGCAGTATGGGA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CTCCAGGCAGCAGTATGGGAA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
CCAGGCAGCAGTATGGGAACT	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AGGCAGCAGTATGGGAACTTC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 12: Appearance frequencies of selected Delta forward primers evaluated on *Homo sapiens* coronavirus genomes from non-SARS-CoV-2 taxa (NCBI). **Note:** This table reports appearance frequencies for a subset of forward primers designed for the SARS-CoV-2 Delta variant, evaluated against *Homo sapiens* sequences from other coronavirus species. No matches were observed in these non-SARS-CoV-2 datasets (all reported frequencies are 0.0), supporting the specificity of the primers with respect to the examined taxa.

Forward Primer ($5'$–$3'$)	Delta	Delta	Delta	Delta
	Pooling Method	Top method	Mix Method	Recon Method
GATCACCGGTGGAATTGCTAC	16	68	29	39
GAATTGCTACCGCAATGGCTT	15	55	25	28
ACTCAGACTAATTCTCGTCGG	29	62	35	32
CTACCGCAATGGCTTGTCTTG	18	49	23	43
TGGAATTGCTACCGCAATGGC	17	60	29	23
CTCAGACTAATTCTCGTCGGC	27	70	41	26
ATTGCTACCGCAATGGCTTGT	20	60	24	27
AGACTCAGACTAATTCTCGTC	27	60	42	32
TAGATTTTGTTCGCGCTACTG	18	53	30	33
CCGGTGGAATTGCTACCGCAA	16	70	23	48
ACCGCAATGGCTTGTCTTGTA	17	74	22	32
GGTGGAATTGCTACCGCAATG	16	63	28	45
CTCCTTTAGATTTTGTTCGCG	26	65	16	25
TCACCGGTGGAATTGCTACCG	7	60	24	2
ACCGGTGGAATTGCTACCGCA	18	54	24	38
ATCACCGGTGGAATTGCTACC	20	59	25	43
TACCGCAATGGCTTGTCTTGT	22	67	18	27
GGAATTGCTACCGCAATGGCT	14	66	28	31
CCGCAATGGCTTGTCTTGTAG	21	61	31	30
GTGGAATTGCTACCGCAATGG	15	64	22	25
TTGCTACCGCAATGGCTTGTC	16	66	31	43
GCTACCGCAATGGCTTGTCTT	10	66	27	22
TCAGACTCAGACTAATTCTCG	27	67	47	65
AATTGCTACCGCAATGGCTTG	15	60	29	22
CAGACTCAGACTAATTCTCGT	18	77	39	34
CGGTGGAATTGCTACCGCAAT	14	57	24	25
TGCTACCGCAATGGCTTGTCT	19	72	16	36
GACTCAGACTAATTCTCGTCG	26	74	44	26
Average Number	18.714	63.536	28.429	32.214

Table 14: Numbers of reverse primer candidates generated under four feature extraction methods for each Delta-variant forward primer.

Note: The number of reverse primer candidates varies with the feature extraction strategy. For each validated Delta forward primer (left column), we report the number of reverse primer candidates produced by Pooling, Top-k, Mix, and Reconstruction.

Reverse primer (5′–3′)	SARS-CoV-2 (Alpha)	SARS-CoV-2 (Beta)	SARS-CoV-2 (Gamma)	SARS-CoV-2 (Delta)	SARS-CoV-2 (Omicron)
Dataset	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI	GISAID and NCBI
Host	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*	*Homo sapiens*
Sequence Number	5,000	5,000	5,000	5,000	5,000
Alpha Variant					
CATCAATGTCTCTGCCAAATTG	0.9936	0.0004	0.0008	0.0002	0.0
ACCAGACACAAATGTGTTGTGT	0.9928	0.0008	0.0018	0.0004	0.0
CAGACACAAATGTGTTGTGTG	0.992	0.0008	0.0018	0.0004	0.0
ACAGCATCAGTAGTGTCATCA	0.9918	0.0004	0.0008	0.0002	0.0
Beta Variant					
ACAGGGTTAGCAAACCTCTT	0.0	0.9638	0.0	0.0006	0.0
CTACTGCTGCCTGGAGTTG	0.0016	0.9594	0.0004	0.0006	0.0008
GTTCGTTTAGACCAGAAGATCAAG	0.0002	0.9546	0.0	0.0004	0.0
GGTTATGATTTTGGAAGCGCTA	0.0006	0.953	0.0178	0.0012	0.0004
Gamma Variant					
AATTTGGTCATCTCGACTG	0.0	0.0	0.9918	0.0002	0.0
TGGTCATCTCGACTGCTATTGGTGT	0.0	0.0	0.9914	0.0002	0.0
GCCAATTTGGTCATCTCGAC	0.0	0.0	0.9912	0.0002	0.0
TGAACCGTCGATTGTGTGAA	0.0	0.0006	0.985	0.0002	0.0
Delta Variant					
CATTGCGGTAGCAATTCCA	0.0006	0.0002	0.0	0.9952	0.165
AGCGCGAACAAAATCTAAAGGA	0.001	0.003	0.0006	0.9934	0.1634
GTAGCGCGAACAAAATCTAAAGGAG	0.0008	0.003	0.0006	0.9932	0.163
GACGAGAATTAGTCTGAGTCTGAT	0.0032	0.0004	0.0004	0.9896	0.1624
Omicron Variant					
ATTGTGCCAACCACCATAGAA	0.0038	0.0008	0.002	0.0036	0.834
GGTGTGACTGTTATTGCCTGACCA	0.0	0.0	0.0	0.0	0.8292
TAACGTACCTATCTCTTCCGAA	0.0032	0.0	0.001	0.0	0.8286
CACCTGTGCCTTTTAAACCATTG	0.0	0.0	0.0	0.0002	0.828

Table 15: Appearance frequencies of selected reverse primers across SARS-CoV-2 variants, evaluated on 5,000 *Homo sapiens* host sequences per variant (combined GISAID and NCBI datasets).
Note: Values report the fraction of sequences in which each reverse primer occurs within the corresponding variant-specific dataset. Reverse primers are designed to be variant-discriminative; consequently, they show high occurrence in the target variant and low occurrence in non-target variants.

Appendix C Flowcharts of the methodology

This section summarizes the complete workflows for forward and reverse primer design. Figure 14 presents the forward primer design pipeline, from data collection and preprocessing through feature extraction and screening to the final set of candidate forward primers. Figure 15 illustrates the subsequent reverse primer design workflow, which is performed after forward primer selection. For an overview of the full end-to-end approach, see Figure 1.

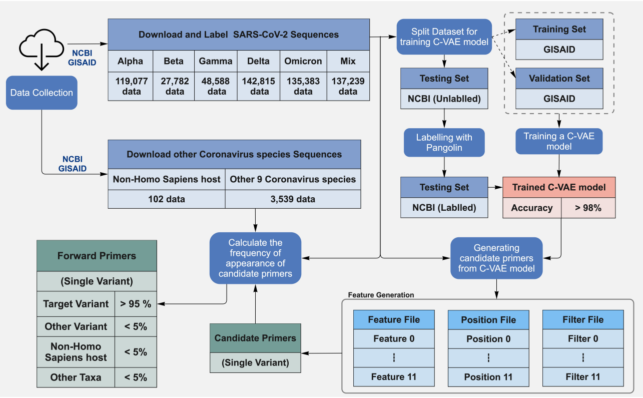


Figure 14: Flowchart of the forward primer design workflow.

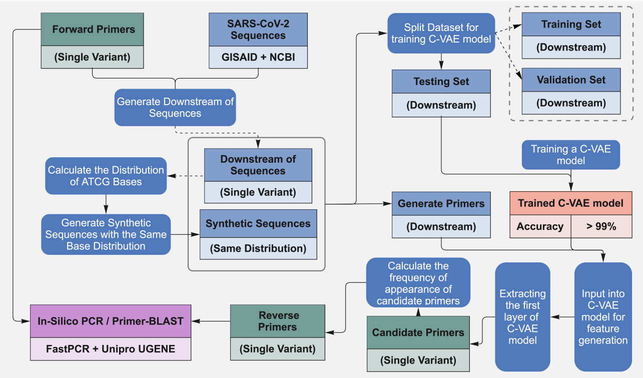
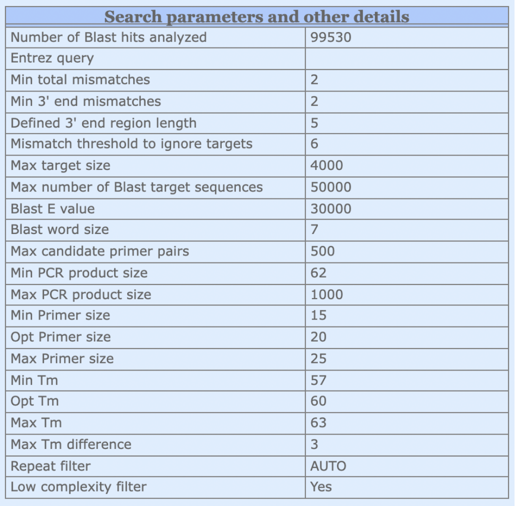


Figure 15: Flowchart of the reverse primer design workflow.

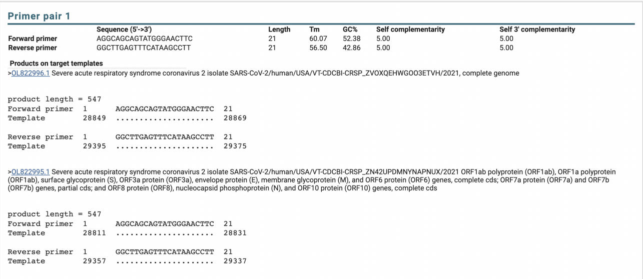
Appendix D BLAST and *in silico* PCR for primer validation

This section provides representative outputs from Primer-BLAST, which we used to verify primer-pair specificity within SARS-CoV-2. Figures 16 and 17 show the Primer-BLAST search settings and an example of the resulting alignments, respectively.



Search parameters and other details	
Number of Blast hits analyzed	99530
Entrez query	
Min total mismatches	2
Min 3' end mismatches	2
Defined 3' end region length	5
Mismatch threshold to ignore targets	6
Max target size	4000
Max number of Blast target sequences	50000
Blast E value	30000
Blast word size	7
Max candidate primer pairs	500
Min PCR product size	62
Max PCR product size	1000
Min Primer size	15
Opt Primer size	20
Max Primer size	25
Min Tm	57
Opt Tm	60
Max Tm	63
Max Tm difference	3
Repeat filter	AUTO
Low complexity filter	Yes

Figure 16: Primer-BLAST search settings used for primer-pair validation.
Note: Primer-BLAST was run on the NCBI website with the search restricted to SARS-CoV-2 sequences. Key parameters were specified to ensure a consistent and valid specificity assessment.



Primer pair 1

	Sequence (5'->3')	Length	Tm	GC%	Self complementarity	Self 3' complementarity
Forward primer	AGGCAGCAGTATGGGAACTTC	21	60.07	52.38	5.00	5.00
Reverse primer	GGCTTGAGTTTCATAAGCCTT	21	56.50	42.86	5.00	5.00

Products on target templates

>OL822996.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/VT-CDCBI-CRSP_ZVOXQEHWGOO3ETVH/2021, complete genome

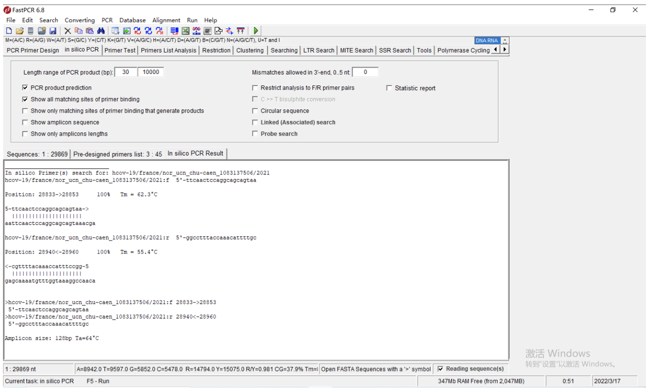
```
product length = 547
Forward primer  1      AGGCAGCAGTATGGGAACTTC  21
Template        28849  .....................  28869

Reverse primer  1      GGCTTGAGTTTCATAAGCCTT  21
Template        29395  .....................  29375
```

>OL822995.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/VT-CDCBI-CRSP_ZN42UPDMNYNAPNUX/2021 ORF1ab polyprotein (ORF1ab), ORF1a polyprotein (ORF1ab), surface glycoprotein (S), ORF3a protein (ORF3a), envelope protein (E), membrane glycoprotein (M), and ORF6 protein (ORF6) genes, complete cds; ORF7a protein (ORF7a) and ORF7b (ORF7b) genes, partial cds; and ORF8 protein (ORF8), nucleocapsid phosphoprotein (N), and ORF10 protein (ORF10) genes, complete cds

```
product length = 547
Forward primer  1      AGGCAGCAGTATGGGAACTTC  21
Template        28811  .....................  28831

Reverse primer  1      GGCTTGAGTTTCATAAGCCTT  21
Template        29357  .....................  29337
```

Figure 17: Example Primer-BLAST output for primer-pair specificity assessment.
Note: The output lists matching SARS-CoV-2 accessions identified by Primer-BLAST and reports the predicted amplicon length for each match under the specified settings.

In this section, Figures 18 and 19 show representative *in silico* PCR results obtained using FastPCR and Unipro UGENE, respectively.



FastPCR 6.8

Length range of PCR product (bp): 30 10000

Mismatches allowed in 3'-end, 0..5 nt: 0

☑ PCR product prediction
☑ Show all matching sites of primer binding
☐ Show only matching sites of primer binding that generate products
☐ Show amplicon sequence
☐ Show only amplicons lengths
☐ Restrict analysis to F/R primer pairs
☐ C >> T bisulphite conversion
☐ Circular sequence
☐ Linked (Associated) search
☐ Probe search
☐ Statistic report

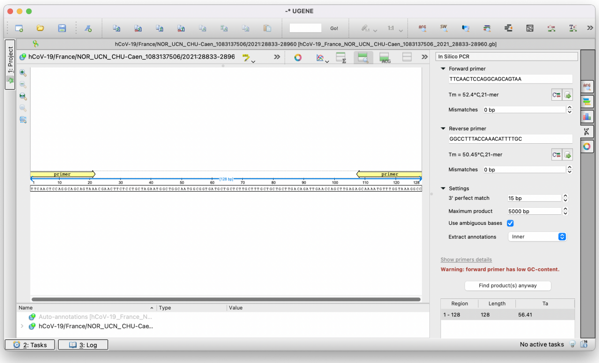
Sequences: 1 : 29869 | Pre-designed primers list: 3 : 45 | In silico PCR Result

```
In silico Primer(s) search for: hcov-19/france/nor_ucn_chu-caen_1083137506/2021
hcov-19/france/nor_ucn_chu-caen_1083137506/2021:f  5'-ttcaactccaggcagcagtaa

Position: 28833->28853     100%   Tm = 62.3°C

5-ttcaactccaggcagcagtaa->
  ||||||||||||||||||||
aattcaactccaggcagcagtaaacga

hcov-19/france/nor_ucn_chu-caen_1083137506/2021:r  5'-ggcctttaccaaacattttgc

Position: 28940<-28960     100%   Tm = 55.4°C

<-cgttttacaaaccatttccgg-5
  ||||||||||||||||||||
gagcaaaatgtttggtaaaggccaaca


>hcov-19/france/nor_ucn_chu-caen_1083137506/2021:f 28833->28853
 5'-ttcaactccaggcagcagtaa
>hcov-19/france/nor_ucn_chu-caen_1083137506/2021:r 28940<-28960
 5'-ggcctttaccaaacattttgc

Amplicon size: 128bp Ta=64°C
```

1 : 29869 nt | A=8942.0 T=9597.0 G=5852.0 C=5478.0 R=14794.0 Y=15075.0 R/Y=0.981 CG=37.9% Tm= | Open FASTA Sequences with a '>' symbol | ☑ Reading sequence(s)

Current task: in silico PCR F5 - Run | 347Mb RAM Free (from 2,047MB) | 0:51 | 2022/3/17

Figure 18: Representative *in silico* PCR results for selected primer pairs evaluated using FastPCR.



hCoV-19/France/NOR_UCN_CHU-Caen_1083137506/2021:28833-28960 [hCoV-19_France_NOR_UCN_CHU-Caen_1083137506_2021_28833-28960.gb]

In Silico PCR

Forward primer: TTCAACTCCAGGCAGCAGTAA

Tm = 52.4°C,21-mer

Mismatches: 0 bp

Reverse primer: GGCCTTTACCAAACATTTTGC

Tm = 50.45°C,21-mer

Mismatches: 0 bp

Settings

3' perfect match: 15 bp

Maximum product: 5000 bp

Use ambiguous bases: ☑

Extract annotations: Inner

Show primers details

Warning: forward primer has low GC-content.

Find product(s) anyway

Region	Length	Ta
1 - 128	128	56.41

UGENE
1: Project
primer
primer
Name
Type
Value
Auto-annotations [hCoV-19_France_N...
hCoV-19/France/NOR_UCN_CHU-Cae...
2: Tasks
3: Log
No active tasks

Figure 19: Representative *in silico* PCR results for selected primer pairs evaluated using Unipro UGENE.

Appendix E Comparison with Primer3Plus

This appendix compares Primer C-VAE with Primer3Plus. Table 16 summarizes primer pairs generated by our deep-learning pipeline (including both forward and reverse primers) and contrasts them with primer pairs generated using Primer3Plus under the corresponding settings.

	Primer C-VAE (No practical upper limit on input length)		Primer3Plus (Up to 10,000 bp per input)	
Forward primers				
Total (initial)	3,626 primers		30 primers	
GC-content filtering	2,725 primers	↓24.85%	30 primers	↓0%
Appearance-frequency filtering	29 primers	↓98.94%	0 primers	↓100%
Reverse primers				
Total (initial)	53,777 primers		30 primers	
GC-content filtering	34,955 primers	↓35%	30 primers	↓0%
Appearance-frequency filtering	357 primers	↓98.98%	0 primers	↓100%
Primer pairs				
Primer design rules	143 primer pairs		0 primer pairs	

Table 16: Comparison between Primer C-VAE and Primer3Plus for primer generation and filtering outcomes. **Note:** Primer3Plus limits the input sequence length to 10,000 bp and returns up to 10 primer pairs per input (10 forward and 10 reverse). For SARS-CoV-2 (˜30,000 bp), the genome must therefore be split into three segments and processed separately, yielding 30 candidate primers of each type before downstream filtering.