SYNTHETIC BIOLOGY

Synthetic Biology, 2025, **10(1)**, ysaf007

DOI: https://doi.org/10.1093/synbio/ysaf007
Advance Access Publication 31 March 2025
Research Article

In silico prediction of loop-mediated isothermal amplification using a generalized linear model

Kenshiro Taguchi[1,2], Satoru Michiyuki[2], Takumasa Tsuji[3], Jun'ichi Kotoku[1,3,*]

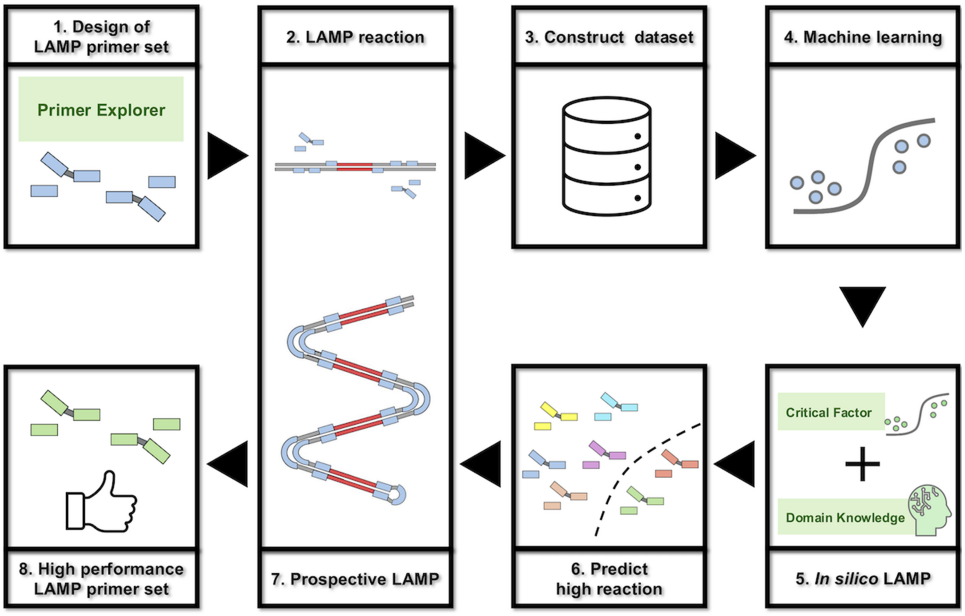
[1]Graduate Degree Program of Health Data Science, Teikyo University, 2-11-1 Kaga, Itabashi-Ku, Tokyo 173-8605, Japan
[2]Fundamental Research Laboratory, Eiken Chemical Co., Ltd, Tochigi 329-0114, Japan
[3]Graduate School of Medical Care and Technology, Teikyo University, 2-11-1 Kaga, Itabashi-Ku, Tokyo 173-8605, Japan
*Corresponding author. Jun'ichi Kotoku Graduate School of Medical Care and Technology, Teikyo University, 2-11-1 Kaga, Itabashi-Ku, Tokyo 173-8605, Japan. E-mail: kotoku@med.teikyo-u.ac.jp.

Abstract

Loop-mediated isothermal amplification (LAMP), a DNA amplification technique under isothermal conditions, provides the important benefits of high sensitivity, specificity, rapidity, and simplicity. Maximizing LAMP features necessitates the design of a complex LAMP primer set (LPS) consisting of four primers for six regions of a given target DNA. Furthermore, the LPS of a given target DNA is designed with LPS design support software such as Primer Explorer. However, even if the design is completed, we still must do many *in vitro* experiments and evaluations. Consequently, designing LPS often fails to achieve high performance, including efficient amplification. For this study, we examined *in silico* LAMP: a generalized linear model to predict DNA amplification from LPS. Using logistic regression with elastic net regularization, we identified factors that strongly affect LPS design. These factors, combined with domain knowledge for LPS design, led to the creation of LAMP kernel variables that are highly essential for high LAMP reaction. *In silico* LAMP, constructed using logistic regression with LAMP kernel variables, allows classification and performance prediction of LPS with an area under the curve of 0.86. These results suggest that a high LAMP reaction can be predicted using LAMP kernel variables and generalized linear regression model. Moreover, an LPS with high performance can be constructed without *in vitro* experimentation.

Keywords: generalized linear model; gene amplification; loop-mediated isothermal amplification (LAMP); machine learning; primer design

Graphical Abstract



Submitted: 26 March 2025; **Received (in revised form)**: 11 March 2025; **Accepted**: 26 March 2025

1. Introduction

Loop-mediated isothermal amplification (LAMP), a technique for the amplification of DNA under isothermal conditions, has been researched and developed actively as a powerful tool for genetic testing [1, 2] because without strict temperature cycles, LAMP has greater or equal efficient amplification and higher specificity compared to those achieved using polymerase chain reaction (PCR). Consequently, LAMP is expected to become the gold standard for genetic testing, alongside PCR.

In fact, PCR requires two primers and a strict reaction temperature cycle for DNA amplification within 2–3 h [3–5]. In contrast, LAMP requires four primers and constant temperature control for DNA amplification within 15 min to 1 h. The LAMP features are achieved by designing a LAMP primer set (LPS) consisting of four primers [F3 primer, forward inner primer (FIP), backward inner primer (BIP), B3 primer] for six regions (F3, F2, F1, B1, B2, and B3) of a given target DNA. Furthermore, by its design, LAMP achieves high amplification efficiency that quickly amplifies a given target DNA [6]. In addition, LAMP without strict temperature cycles allows the design of portable devices for DNA amplification [7]. The portable device is used widely for testing and diagnosing infectious diseases (e.g. malaria, tuberculosis) in developing countries as Point of Care Testing (POCT) [8–10].

Although LAMP presents numerous benefits, designing an LPS of a given target DNA with Primer Explorer, which is LPS design support software [11], can be complex. First, the LPS is designed using DNA features, including thermal stability (ΔG, ΔH), but the designed LPSs do not always achieve high sensitivity and rapidity as gene testing and diagnostic agents. Second, the LPS design to achieve the high sensitivity and rapidity that we aim for is fine-tuned using a seat-of-the-pants and brute-force approach by which we shorten or lengthen sequences of each primer in the designed LPS and shift their positions against a given target DNA. Novel primers (loop primers; LF/LB) and alternative primers for LF/LB have been added to achieve high rapidity [12, 13]. Moreover, an LPS design tool with high specificity [14] has been developed, but it has not yet proved to be a fundamental solution to the complexity of LPS design.

A recently studied predictive model of gene amplification might simplify primer design and optimize more specific primers for target DNA [15]. Using PCR, recurrent neural networks (RNNs), which are used in natural language processing, predicted the amplification of target DNA from PCR primer sets with 70% accuracy [16]. With isothermal amplification techniques, including LAMP, least absolute shrinkage and selection operator (LASSO) regression to predict gene amplification has been constructed under limited reaction conditions using only turn-back primers (in LAMP, FIP, and BIP) and Aac enzyme [17]. Although the regression model is thought to capture trends of isothermal amplification techniques, it is unlikely to reflect the amplification prediction of the LAMP reaction itself. As the first study based on this context, a LAMP amplification prediction model with an F1 score of 0.64 has been reported [18]. However, such models with high accuracy and amplification prediction in the LAMP reaction itself have not yet been established. A LAMP amplification prediction model providing higher accuracy is anticipated.

A prediction model with the explanatory power (X-Model) of PCR amplification has been constructed. The free energy of annealing (ΔG) has been identified as an important factor for PCR amplification [19]. Actually, the X-Model is not a black box; it enables us to understand why a prediction was made. Moreover, it has greatly facilitated medical diagnosis [20, 21]. In addition, domain knowledge is often combined when building a prediction model [22]. The model enables us to understand the prediction factor well and to capture desirable characteristics for the prediction. As one might expect, an X-model with higher explanatory power can be built using domain knowledge and can be expected to improve the prediction accuracy [23].

For this study, to reduce the actual time and economic costs associated with *in vitro* experiments, we developed *in silico* LAMP: a generalized linear model to predict DNA amplification from LPS. We used logistic regression with an elastic net penalty to identify important factors affecting high LAMP amplification. Using those identified critical factors and domain knowledge for high LAMP amplification, we created LAMP kernel variables. Subsequently, we built *in silico* LAMP with logistic regression and those variables. Our *in silico* LAMP helps us to understand and design LPSs with high performance without using *in vitro* experimentation.

2. Materials and methods

2.1 Template

Human genomic DNA (Promega K.K.), a full genome, was used as the target DNA template for LAMP reaction. The template solution was prepared, diluted with DNA diluent (5 mM Tris-HCl, 2 μg/ml ssDNA) to 3,300, 330, 33, and 0 pg/5 μl, dispensed in 75 μl aliquots into PCR Plate 96 (Eppendorf AG) using VIAFLO96 (Integra Biosciences AG), and stored at −30°C. The template solutions were treated as 1,000, 100, 10, and 0 copies/5 μl, respectively, based on the molecular weight of the human genomic DNA.

2.2 Target DNA and LAMP primer set

Human chromosome 21 (NC_000021.9) was selected from the latest human reference genome sequence (GRCh38.p13) [24] as the target DNA for LPS. As a feature of LAMP, an LPS is often designed to target sequences of around 500 bp or less. It is not practical to use extremely long sequences of 1000 bp or more as a target. Therefore, sufficiently long sequences of 1000 kinds (980 bp) were extracted randomly from the target DNA. To obtain a unique sequence against full genome (Supplementary Figure S1), Basic Local Alignment Search Tool (BLAST) [25] was used to exclude sequences of 698 kinds that have more than two regions with >90% homology within the genome. From the selected sequences of 302 kinds, 268 were used for LPS design (Fig. 1). Of these (Supplementary Table S1 and S2), LPSs of 96 kinds (F3, B3, FIP, and BIP) were synthesized by Eurofins Genomics K.K. Then 5× LPS solution was also prepared, containing 1.6 μM of FIP, 1.6 μM of BIP, 0.2 μM of F3, and 0.2 μM of B3, dispensed in 8 μl portions into PCR Plate 96 (Eppendorf AG) using VIAFLO96 (Integra Biosciences AG) and stored at −30°C. LAMP reaction was used with them.

2.3 LAMP reaction

We prepared LAMP reaction buffer containing 14 mM tricine, 8.2 mM $MgSO_4$, 1.7 mM each of dNTPs, 1 mM DTT, 0.5% Tween20, a DNA intercalator, and 18.3 U Bst enzyme. After the reaction buffer was prepared in a large volume and then dispensed in 15 μl aliquots into PCR Plate 96 (Eppendorf AG) using VIAFLO96 (Integra Biosciences AG), they were stored at −30°C.

After the template solution was heated at 95°C for 5 min, it was chilled on ice. Then 5 μl of 5× LPS solution was added using VIAFLO96 (Integra AG). After 5 μl of the heated template solution was added (Biomek 4000; Beckman Coulter Inc.) to the reaction buffer, the LAMP reaction was conducted by incubating the mixture at 65°C for 90 min (CFX96; BioRad Laboratories, Inc.) with subsequent heating at 80°C for 10 min to terminate the

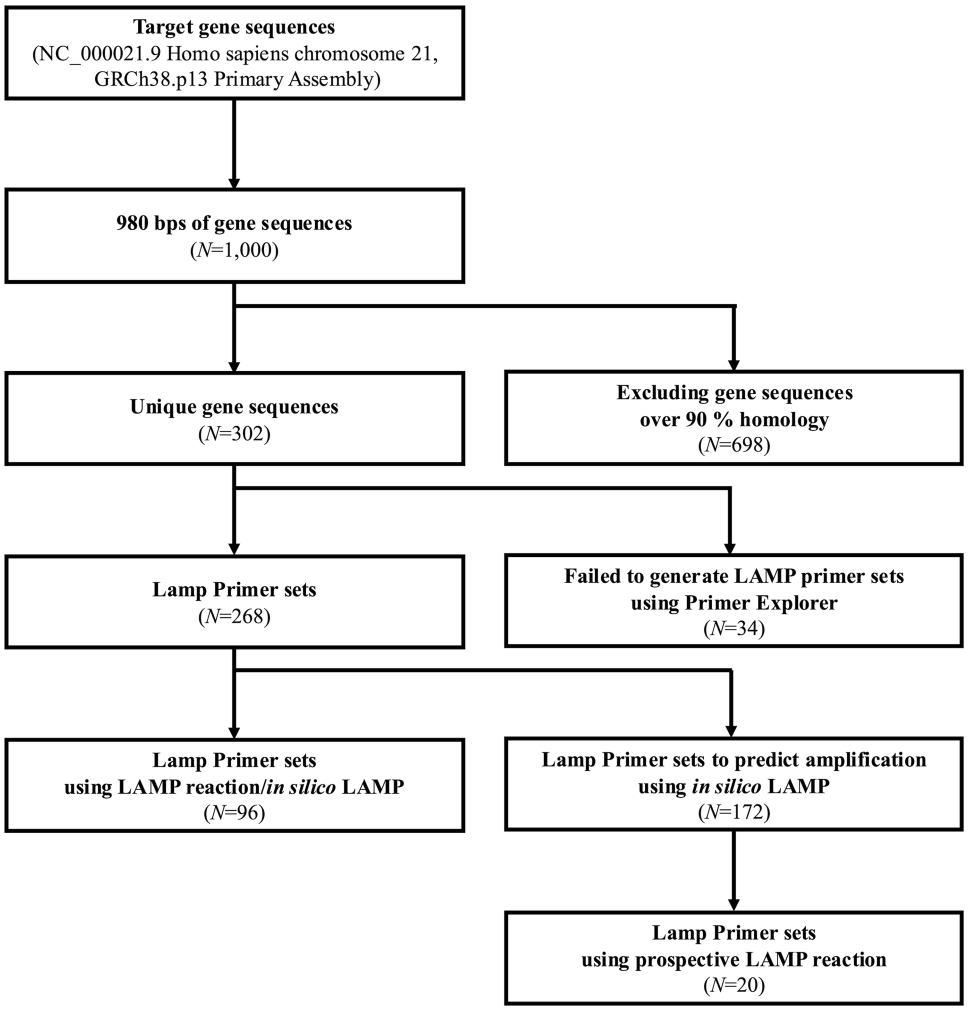


Figure 1. Flowchart showing LAMP primer set creation for the LAMP reaction, *in silico* LAMP, and the prospective LAMP reaction.

reaction. LAMP reactions were monitored in real time according to the fluorescence intensity of the DNA intercalator [26]. The LAMP amplification time (*Tt*) was calculated from the increment of the derivative value of fluorescence at the beginning of the reaction.

2.4 Data set construction

An LPS has factors such as length, melting temperature (T_m), GC content, the free energy of annealing (ΔG), the enthalpy (ΔH), and folding energy (m_{fold}) [27], which are important factors of local regions such as the 3′ and 5′ ends for isothermal amplification [17]. Design regions of an LPS for target DNA were segmented (Fig. 2). Those factors in each region were used as explanatory variables (Supplementary Table S3). These explanatory variables were standardized when creating prediction models.

The response variable is designated as being of high amplification (High-amp) or low amplification (Low-amp) based on whether it satisfied the LAMP condition. Then the response variable and condition can be formulated as presented below.

$$y_i = \begin{cases} High\ amp,\ if\ the\ LPS_i\ satisfy\ the\ LAMP\ condition \\ Low\ amp,\ otherwise \end{cases} \quad (1)$$

$$LAMP\,condition := condition\,1\,|\,(condition\,2\,\&\,condition\,3) \quad (2)$$

$$condition\,1 := \text{mean of}\,Tt_i\,(10\,\text{copies}) < 55 \quad (3)$$

$$condition\,2 := \text{mean of}\,Tt_i\,(1{,}000,\,100,\,10\,\text{copies}) \leq 60 \quad (4)$$

$$condition\,3 := AUC_i \geq 0.95 \quad (5)$$

There in, y_i is the response variable of individual $i \in \{1, \ldots, N\}$ and is satisfied by the LAMP condition. Additionally, LPS_i and Tt_i denote the LPS and the *Tt* of individual $i \in \{1, \ldots, N\}$, respectively. Also, AUC_i was calculated from Receiver Operating Characteristic (ROC) analysis using positive samples (1,000, 100, and 10 copies) and negative samples (0 copies) of each LPS_i.

2.5 Statistical analysis

The LAMP explanatory variables used for constructing the dataset were continuous variables expressed as the mean and standard deviation (SD). Differences in the LAMP explanatory variables were analyzed using analysis of variance. All statistical analyses were conducted using R (ver. 4.4.0) [28]. Results for which the *P*-value was <.05 were inferred as significant.

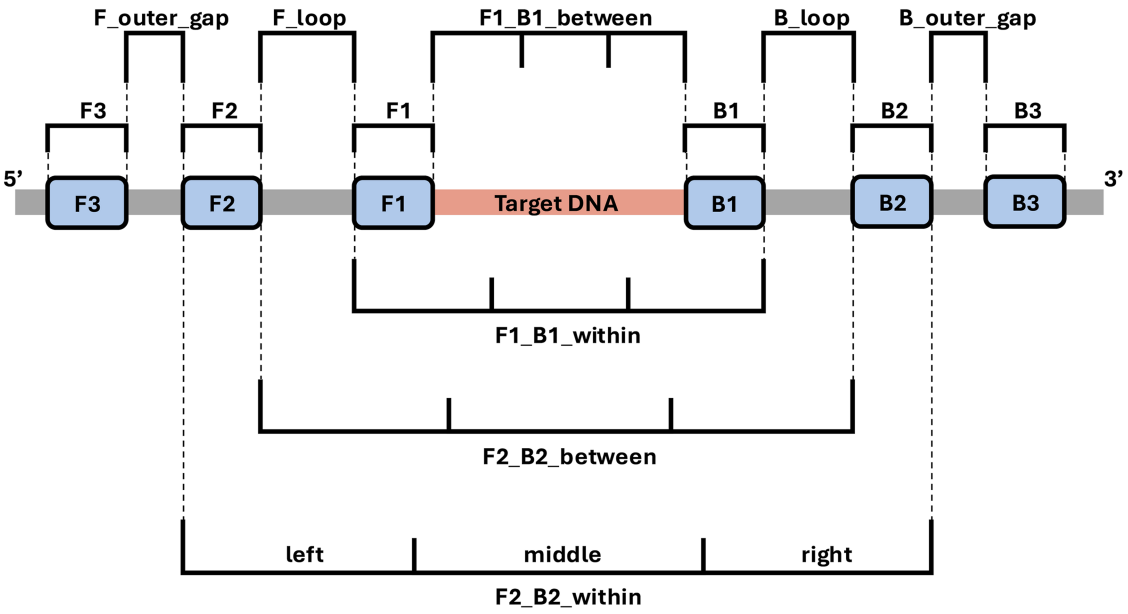


Figure 2. Segmented design regions of an LPS.

2.6 Generalized linear regression model and penalty term

We created a generalized linear regression model (GLM) [29] for predicting LAMP amplification. To estimate the LAMP amplification status (High-amp or Low-amp) as categorical outcomes, to select variables, and to avoid overfitting, we used a penalized logistic regression model on GLM. In the framework of GLM, logistic regression models can be represented by a random component, a systematic component, and a link function. In the random component, the response variable y_i follows a binomial distribution, the probability of which is π_i. In the systematic component, the linear predictor, given as a linear combination of the explanatory variables x_i, the parameters as partial regression coefficient w and intercept b, which can be formulated as

$$\eta_i = x_i w + b, \tag{6}$$

where x_i represents general and segmented features corresponding to each LPS_i ($i \in N$). In the link function, we use the logit link function, which is formulated as

$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{7}$$

This function linking η_i to π_i is given as

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{8}$$

Solving for π_i, the logistic regression model is given as

$$\pi_i = \frac{1}{1+\exp(-\eta_i)}. \tag{9}$$

The elastic net penalty ($P_{Elastic}$), a hybrid of LASSO and Ridge, combines L1 and L2 penalties (P_{L1} and P_{L2}) [30–32]. We fit w and b using maximum likelihood estimation. Maximization of the likelihood corresponds to the minimization of the negative logarithm of the likelihood L. Furthermore, loss function l including the $P_{Elastic}$ is defined as presented below.

$$l = -\frac{1}{N}L + \alpha P_{Elastic} \tag{10}$$

$$L = \sum_{i=1}^{N} B_i L_i \tag{11}$$

$$L_i = y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \tag{12}$$

$$B_i|_{y_i=g} = \frac{N}{2n_g} \tag{13}$$

$$P_{Elastic} = P_{L1} + P_{L2} \tag{14}$$

$$P_{L1} = W_{L1}\|w\|_1 \tag{15}$$

$$P_{L2} = \frac{(1-W_{L1})}{2}\|w\|_2 \tag{16}$$

In those equations, α stands for the weight of $P_{Elastic}$, $B_i|_{y_i=g}$ denotes the weight of L_i of individuals i belonging to categorical outcome $g \in \{Highamp, Lowamp\}$, n_g expresses the number belonging to categorical outcome g [33]. Also, W_{L1} is of [0, 1]: the weight of $\|w\|_1$ is the L1 norm; and $1 - W_{L1}$ the weight of $\|w\|_2$ is the L2 norm.

2.7 Prediction models

With W_{L1}, the penalized logistic regression model was constructed (EN-Model). From the explanatory variables sparsified in the EN-Model, and using these partial regression coefficients, which were non-zero, the logistic regression model was constructed (SL_{EN}) (Table 1).

Table 1. Prediction models and comparison models

	W_{L1}	Penalty	Notes
All explanatory variables			
L-Model	1	P_{L1}	Penalized logistic regression
R-Model	0	P_{L2}	Penalized logistic regression
EN-Model	[0, 1]	$P_{Elastic}$	Penalized logistic regression
Sparsified explanatory variables			
SL_L	-	-	Logistic regression
SL_R	-	-	Logistic regression
SL_{EN}	-	-	Logistic regression
LAMP kernel variables			
EN-Model	[0, 1]	$P_{Elastic}$	Penalized logistic regression
SL_{EN} (*in silico* LAMP)	-	-	Logistic regression

2.8 Comparison models

With W_{L1} of 0 or 1, where the value of $P_{Elastic}$ is equal to P_{L2} or P_{L1}, the penalized logistic regression model was built (R-Model or L-Model). From the explanatory variables sparsified in the R-Model or L-Model, and using these partial regression coefficients, which were non-zero, the logistic regression model was built (SL_R, SL_L) (Table 1).

2.9 Critical variable identification and predictive model optimization

To select explanatory variables, we performed nested cross-validation [34]. After we selected 12 out of 96 LPSs for validation in eight-fold outer cross-validation and used the remaining 84 LPSs for model fitting, we used 14 out of 84 LPSs for validation in six-fold inner cross-validation and used the remaining 70 LPSs for optimizing hyperparameters (α, W_{L1}). Hyperparameters were optimized using Bayesian optimization to maximize the balanced accuracy, which can be formulated as the following.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \quad (17)$$

In that equation, TP, TN, FP, and FN denote True Positive, True Negative, False Positive, and False Negative, respectively [35]. The *P*-values of the estimated partial regression coefficients are based on the Wald statistic. We used Python (ver. 3.10.4) and employed the statsmodels package (ver. 0.14.2) [36] for the penalized logistic regression model and optuna package (ver. 3.1.1) [37] for Bayesian optimization.

2.10 Creation of LAMP kernel variable and *in silico* LAMP

We created LAMP kernel variables considering that local ΔH explains the energy of DNA synthesis more simply, including insertion or extension events [38], and considering that domain knowledge is useful to build models with higher explanatory power and prediction accuracy [23]. The LAMP kernel variables were created with ratios of GC contents to enthalpies, the GC contents of F2_GC and B2_GC, which are identified as critical factors for High-amp in the EN-Model using all explanatory variables, and the enthalpies of F2_dH and B2_dH, which play an important role in DNA synthesis from domain knowledge. The SL_{EN} with LAMP kernel variables is also called *in silico* LAMP (Table 1).

2.11 Prospective LAMP reaction

Using *in silico* LAMP, we predicted the amplification statuses of LPSs of 172 kinds in a LAMP reaction. In all, LPSs of 20 kinds, with LPSs of 10 kinds each selected randomly from the LPS predicted to be High-amp and Low-amp, were used for LAMP reaction to conduct an experiment.

3. Results

3.1 Properties of dataset

The explanatory variables of the 127 features per LPS were obtained by removing the features including missing values from the 172 features per LPS (Supplementary Table S3). Table 2 shows the distribution of the explanatory variables sparsified in the EN-Model. Of the LPSs of 96 kinds in the dataset, 18 (18.8%) were labeled as High-amp. For each of the GC contents and ΔH values of an LPS region (F3, F2, F1, B1, B2, and B3), a significant difference ($P < .05$) was found between High-amp and Low-amp in the LAMP reaction (Table 2 and Supplementary Table S4). Low-amp of LPS in the dataset had F2_dH mean of -170.9 ± 17.0 kcal/mol and B2_dH mean of -162.2 ± 17.0 kcal/mol. High-amp of LPS in the dataset was associated with a higher F2_dH mean of -154.5 ± 15.4 kcal/mol and B2_dH mean of -149.5 ± 6.9 kcal/mol. In addition, the Low-amp of LPS in the dataset had F2_GC mean of $40.1 \pm 9.1\%$ and B2_ GC mean of $42.6 \pm 8.8\%$. High-amp of LPS in the dataset was associated with a higher F2_GC mean of $51.6 \pm 9.1\%$ and B2_GC mean of $52.6 \pm 5.5\%$.

3.2 Partial regression coefficients of EN-model and SL_{EN}

Figure 3 presents the EN-Model and SL_{EN} with all explanatory variables and LAMP kernel variables for partial regression coefficients and *P*-value counts. The EN-Model was built to select variables involved in the response variables by sparsifying all explanatory variables. The explanatory variables, including F2_GC and B2_GC, were selected. These partial regression coefficients were not 0 in either model in the nested cross-validation. Other partial regression coefficients were 0 in all models in the nested cross-validation (Supplementary Fig. S2). Next, of those selected explanatory variables, SL_{EN} was built to identify factors that can be shown statistically as influencing the response variables. Critical factors with significant ($P < .05$) partial regression coefficients in either model in nested cross-validation, including F2_GC and B2_GC, were identified.

We created LAMP kernel variables with the identified critical factors and domain knowledge to build a prediction model with higher explanatory power. First, we built the EN-Model using the LAMP kernel variables to confirm whether it is involved in the response variables, or not. Their partial regression coefficients were non-zero, confirming their involvement in the response variables. Additionally, we built SL_{EN} to confirm that the LAMP kernel variables influence the response variables. The LAMP kernel variables with significant ($P < .05$) partial regression coefficients in all models in nested cross-validation were observed.

3.3 Comparison of models and classification performance

Figure 4 presents a confusion matrix of the proposed models evaluated using balanced accuracy. For *in silico* LAMP, the numbers of correct answers and rates of Low-amp and High-amp were, respectively, 50 (64.0%) and 15 (83.0%). The model achieved average balanced accuracy of 73.8% for classification.

Table 2. LAMP primer set characteristics of non-zero variables in the EN-Model

Variables	Low amplification (N = 78)	High amplification (N = 18)	Total (N = 96)	P-value
F3 region				
F3_Tm, °C	54.19 (1.36)	55.81 (1.80)	54.49 (1.58)	<.01
F3_5_dG, kcal/mol	−6.49 (0.91)	−7.12 (1.06)	−6.61 (0.97)	.01
F2 region				
F2_len, bp	22.23 (2.35)	19.89 (2.03)	21.79 (2.46)	<.01
F2_GC, %	40.10 (9.12)	51.57 (9.12)	42.25 (10.13)	<.01
F2_dG, kcal/mol	−32.74 (1.96)	−31.38 (1.63)	−32.48 (1.96)	.01
F2_dH, kcal/mol	−170.93 (17.01)	−154.51 (14.37)	−167.86 (17.69)	<.01
F1 region				
F1_Tm, °C	59.54 (1.40)	61.21 (1.74)	59.85 (1.60)	<.01
F1_GC, %	45.76 (8.68)	54.47 (9.24)	47.39 (9.38)	<.01
B1 region				
B1_5_dG, kcal/mol	−6.42 (0.82)	−7.04 (1.11)	−6.54 (0.91)	.01
B1_5_dH, kcal/mol	−32.43 (1.08)	−33.40 (1.95)	−32.61 (1.33)	<.01
B2 region				
B2_len, %	21.09 (2.33)	19.28 (0.96)	20.75 (2.25)	<.01
B2_GC, %	42.56 (8.78)	52.58 (5.46)	44.44 (9.12)	<.01
B2_dH, kcal/mol	−162.21 (16.99)	−149.51 (6.88)	−159.83 (16.35)	<.01
B2_5_dG, kcal/mol	−6.48 (0.71)	−6.85 (0.71)	−6.55 (0.72)	.05
B2_5_dH, kcal/mol	−32.56 (1.17)	−33.09 (1.15)	−32.66 (1.18)	.08
B2_3_dG, kcal/mol	−6.27 (0.80)	−6.62 (0.54)	−6.33 (0.77)	.08
B3 region				
B3_len, bp	20.37 (2.15)	19.06 (1.06)	20.12 (2.05)	.01
B3_dH, kcal/mol	−156.96 (15.75)	−147.96 (8.25)	−155.27 (15.03)	.02
B3_3_dG, kcal/mol	−6.42 (0.86)	−7.12 (1.06)	−6.55 (0.93)	<.01
Loop region				
F_loop_len, bp	22.79 (6.87)	26.17 (5.32)	23.43 (6.71)	.05
F_loop_Tm, °C	55.20 (9.48)	64.41 (8.79)	56.93 (9.99)	<.01
F_loop_dG, kcal/mol	−34.32 (10.72)	−42.15 (9.30)	−35.79 (10.87)	.01
F_loop_5_dG, kcal/mol	−6.42 (0.75)	−6.99 (0.68)	−6.52 (0.77)	<.01
Segment region				
F1_B1_within_middle_Tm, °C	52.80 (6.10)	56.42 (7.36)	53.48 (6.47)	.03
F2_B2_between_fold_dG, kcal/mol	0.11 (0.99)	−0.07 (0.94)	0.07 (0.98)	.5
F2_B2_within_c_fold_dG, kcal/mol	−0.10 (1.14)	−0.19 (0.92)	−0.12 (1.09)	.74

Data presented the mean (standard deviation) of each variable.

Figure 5 shows the ROC analysis results obtained when evaluating the predictive ability of high-performance LPS in nested cross-validation. Figure 5a and b show the ROC analysis of EN-Model/SL_{EN} using all explanatory variables, the sparsified explanatory variables, and LAMP kernel variables. Importantly, those models were improved when using the LAMP kernel variables [area under the curve (AUC): from 0.782 to 0.833, and from 0.708 to 0.858].

Table 3 presents results of the LPS classification in terms of sensitivity, specificity, balanced accuracy, and AUC. The best AUC was achieved by *in silico* LAMP, yielding 9.7% greater AUC than the EN-Model with all explanatory variables.

3.4 Results of prospective LAMP reaction using *in silico* LAMP

To verify the effects of LPS design using *in silico* LAMP, we performed the prospective LAMP reaction using an LPS predicted by *in silico* LAMP. We present the confusion matrix and the ROC analysis for the prospective LAMP reaction in Fig. 5a and b. The LPS predicted using *in silico* LAMP produced balanced accuracy of 0.725 ± 0.017 and AUC of 0.811 ± 0.007, with no deviation from the balanced accuracy and AUC at the model evaluation, thereby validating the usefulness of *in silico* LAMP in prospective experiments.

4. Discussion

To construct an LPS with high performance without using *in vitro* large-scale experiments, we created LAMP kernel variables, built with *in silico* LAMP by application of their variables and GLM to the LAMP reaction. Actually, LPS design is becoming increasingly complex to achieve the desired performance of the LAMP reaction in terms of, e.g. high amplification and rapidity. Moreover, LPS design necessitates extensive screening by experimentation to achieve higher degrees of performance. This *in silico* LAMP facilitates the design of LPSs with high amplification and rapidity.

After randomly extracting target sequences of 1000 kinds having sufficient length (980 bp) from chromosome 21, we selected unique target sequences of 302 kinds from the human genome. We designed LPSs of 96 kinds and a LAMP dataset showing their amplification status and their characteristics. This operation was done to prevent the LPSs from binding to genome regions other than chromosome 21 in the LAMP reaction. However, although the LPSs were designed with unique target sequences, we did not confirm the specificity of these LPSs in the actual reaction. Therefore, these LPSs have the slightest potential to bind to multiple locations in the genome. In addition, the LAMP reaction was overestimated when actually binding to multiple locations. From our analysis of the LAMP dataset, we were able to confirm almost all the specific factors for efficient amplification in the LAMP reaction

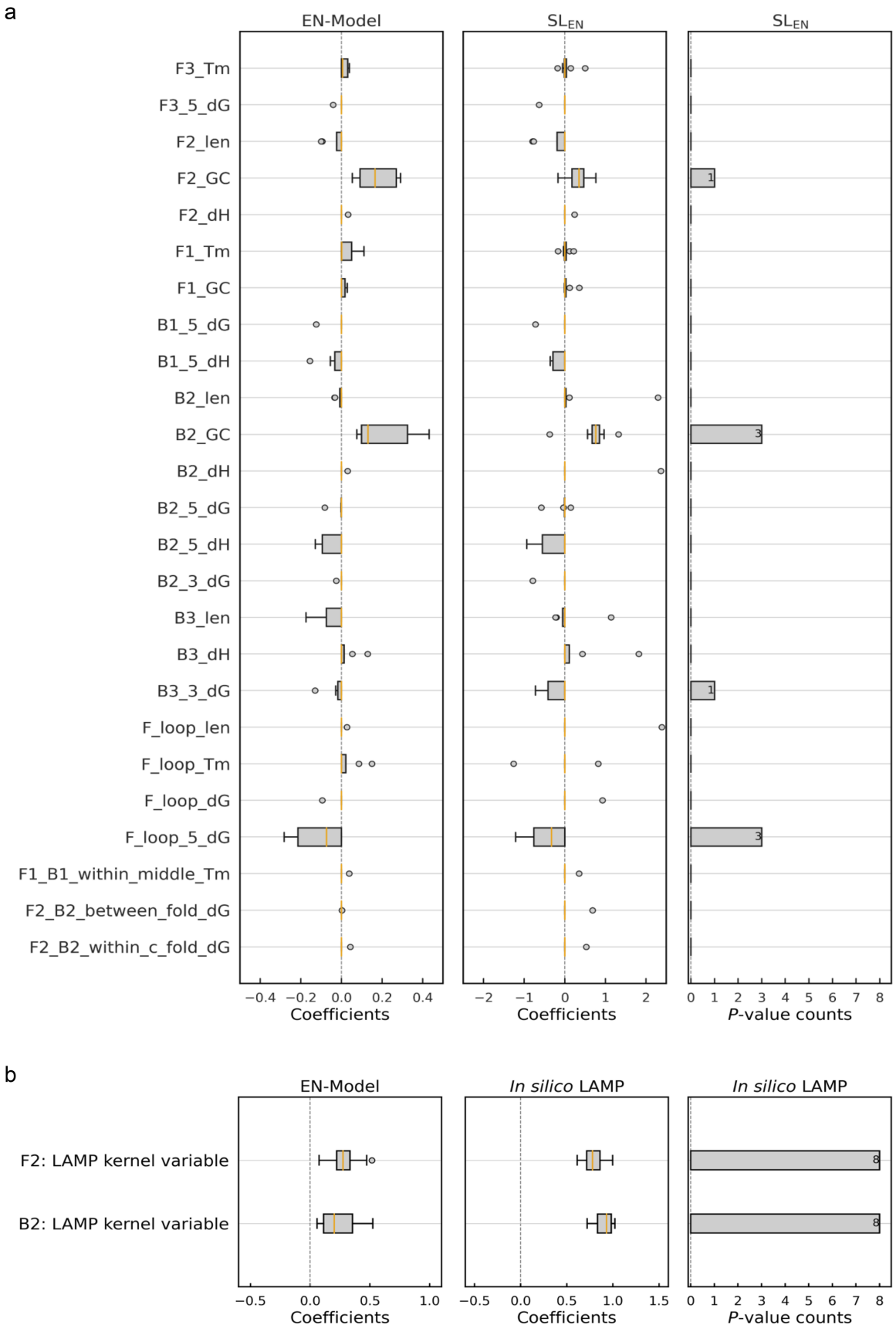


Figure 3. Evaluation in the EN-Model and the SL_{EN}. (a) EN-Model and the SL_{EN} using all explanatory variables. (b) EN-Model and the SL_{EN} (*in silico* LAMP) using LAMP kernel variables are shown with comparison of partial regression coefficients not sparsified in nested cross-validation of the EN-Model in the left panel, comparison of partial regression coefficients in nested cross-validation of the SL_{EN} in the center panel, and counts of significant differences ($P < .05$) for partial regression coefficients in nested cross-validation of the SL_{EN} in the right panel.

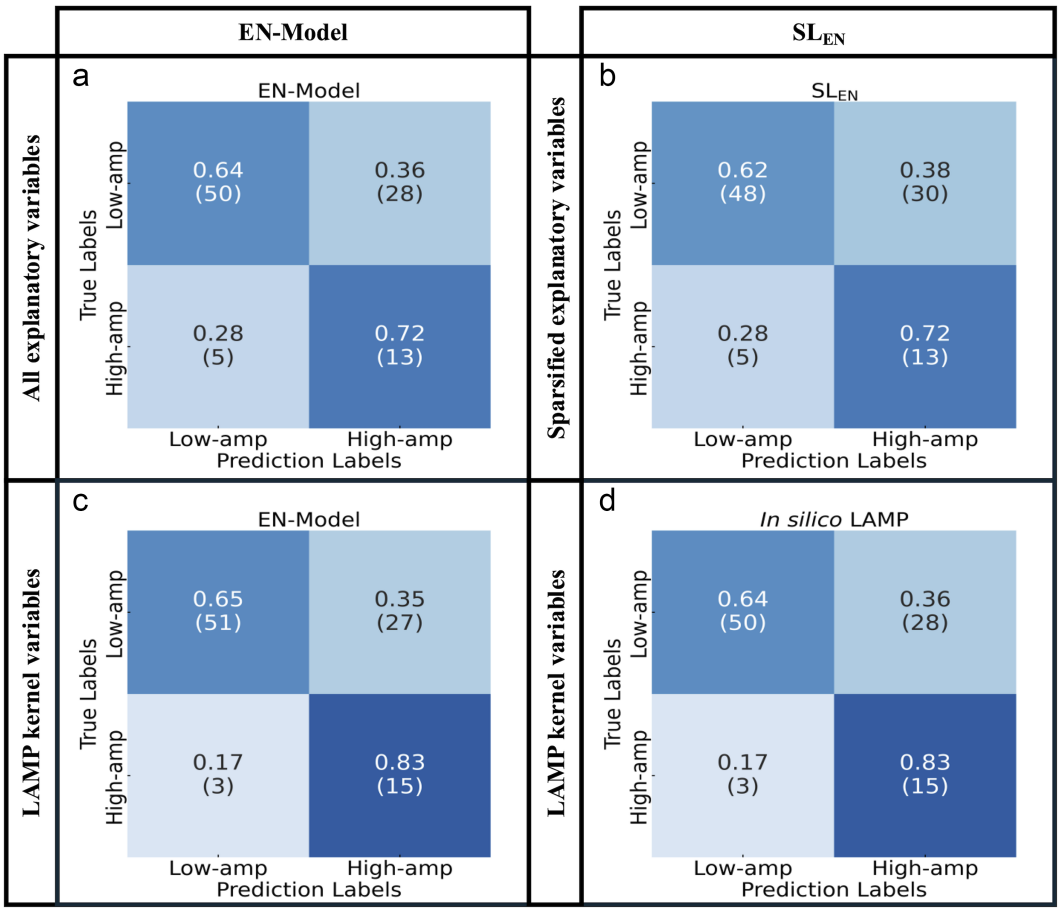


Figure 4. Confusion matrixes of High/Low-amplification classification in the EN-Model and the SL_{EN}. (a) EN-Model with all explanatory variables, (b) SL_{EN} with sparsified explanatory variables, (c) EN-Model with LAMP kernel variables, and (d) *in silico* LAMP: SL_{EN} with LAMP kernel variables.

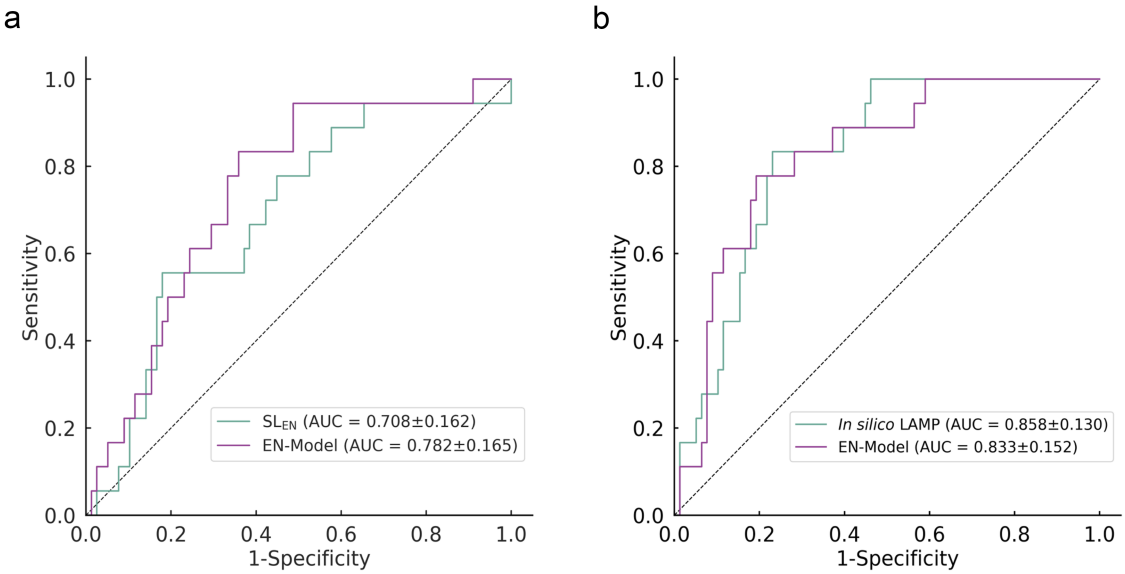


Figure 5. Classification accuracy of the EN-Model and the SL_{EN} classified by High/Low-amplification. AUC, area under the curve. (a) EN-Model with all explanatory variables and SL_{EN} with the sparsified explanatory variables. (b) EN-Model with LAMP kernel variables, *in silico* LAMP: SL_{EN} with LAMP kernel variables.

(Table 2). More specifically, we identified the GC contents and ΔH means in the LPS region (F3, F2, F1, B1, B2, and B3) as the great and significant differences ($P < .05$) between high and low amplification in the LAMP reaction. In PCR requiring a thermal cycle, ΔG contributed significantly to PCR amplification [19]. However, the LAMP reaction is isothermal, suggesting that the energy

Table 3. Comparison of sensitivity, specificity, balanced accuracy, and AUC by learned model

	Sensitivity	Specificity	Balanced Accuracy	AUC
All explanatory variables				
L-Model	0.833 ± 0.220	0.615 ± 0.066	0.724 ± 0.111	0.796 ± 0.154
R-Model	0.792 ± 0.217	0.629 ± 0.095	0.710 ± 0.132	0.722 ± 0.189
EN-Model	0.792 ± 0.217	0.668 ± 0.122	0.730 ± 0.106	0.782 ± 0.165
Sparsified explanatory variables				
SL_L	0.750 ± 0.264	0.617 ± 0.118	0.683 ± 0.128	0.756 ± 0.154
SL_R	0.875 ± 0.331	0.207 ± 0.077	0.541 ± 0.171	0.521 ± 0.181
SL_{EN}	0.688 ± 0.256	0.603 ± 0.115	0.645 ± 0.135	0.708 ± 0.162
LAMP kernel variables				
EN-Model	0.833 ± 0.220	0.653 ± 0.103	0.743 ± 0.130	0.833 ± 0.152
***in silico* LAMP**	$\mathbf{0.833 \pm 0.220}$	$\mathbf{0.643 \pm 0.107}$	$\mathbf{0.738 \pm 0.119}$	$\mathbf{0.858 \pm 0.130}$

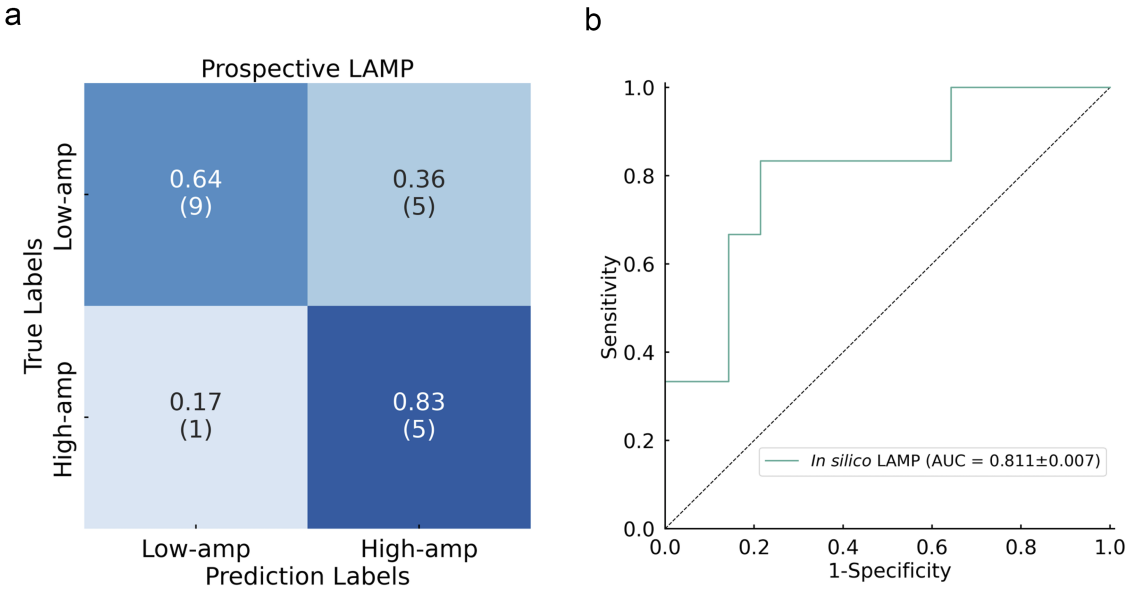


Figure 6. Results of prospective LAMP reaction using *in silico* LAMP: (a) confusion matrixes and (b) ROC curve. AUC, area under the curve.

of the DNA itself (ΔH) [39] makes a major contribution to LAMP amplification.

To identify critical factors for the efficient amplification of the LAMP reaction, we selected explanatory variables using the EN-Model. Among the 127 explanatory variables, the analysis results revealed that the GC content in F2 and B2 showed high partial regression coefficients, meaning that those variables are important for high amplification in LAMP (Fig. 3a). Considering that F2 and B2 are first annealed to a target DNA in the LAMP reaction [1], F2_GC and B2_GC are reasonable variables and are consistent with the LAMP amplification theory.

However, the LAMP reaction using Bst enzyme should incorporate consideration of high or low GC content [40]. In F2 and B2 of high GC content, the enzyme facilitates F2 and B2 to anneal to the single strands of a target DNA very efficiently, but the enzyme struggles at F3 and B3 to unwind the double strands. In contrast, for F2 and B2 of low GC content, F2 and B2 have reduced efficient annealing to the single strands of a target DNA, thereby limiting efficient LAMP amplification. Although DNA synthesis is a complex process, local ΔH, as simple as overall ΔH per base pair, has affected the insertion or extension event and the overall reaction energy [38].

We created new variables, such as the local ΔH as LAMP kernel variables, using domain knowledge of ΔH and the identified critical factors of GC content for high LAMP amplification, which indicate ΔH of F2 and B2 per GC content of F2 and B2. Next, we built *in silico* LAMP using LAMP kernel variables and logistic regression. In nested cross-validation, the median value of the partial regression coefficients was non-zero, confirming a significant difference ($P < .05$) (Fig. 3b). These results confirmed that our LAMP kernel variables, such as local ΔH, functioned as important factors for LAMP reaction.

Our *in silico* LAMP performed well in predicting LPSs with high amplification and rapidity, compared within *silico* LAMP, EN-Model, L-Model, R-Model, SL_{EN}, SL_L, and SL_R, showing balanced accuracy and AUC (Figs 4, 5 and Table 3). *In silico* LAMP can achieve classification with higher accuracy than either the EN-Model, L-Model, R-Model, SL_{EN}, SL_L, or SL_R (Table 3). This finding suggests that LAMP kernel variables are important for high LAMP amplification, allowing for interpretable predictive models, and in turn allowing for highly robust LPS design.

This finding suggests that LAMP kernel variables, which are based on critical features of LPSs designed from sufficiently long target sequences, are important for high LAMP amplification, allowing for interpretable predictive model. Furthermore, the standard LPS design for Primer Explorer is performed in the range where the distance between F3 primer and B3 primer is a maximum of 220 bp or a minimum of 120 bp. Therefore, for general

LPS designs where the target sequence is <1000 bp, highly robust LPSs can be designed.

Additionally, we tested our *in silico* LAMP using independent LPSs and conducting a prospective LAMP experiment. Results showed balanced accuracy of 0.740 with no deviation from the balanced accuracy at *in silico* LAMP evaluation (Fig. 6). The findings suggest that there are additional unidentified critical variables for high LAMP amplification because 1 of 20 kinds of LPS was found to be a FN (Fig. 6a). In contrast, because our *in silico* LAMP, which was also trained and optimized to have more false positives than false negatives, half of the LPSs that *in silico* LAMP predicted as High-amp were observed to be High-amp in the prospective LAMP reaction. However, the probability distribution of the True Labels in the prospective experiment (Fig. 6a) is similar to the probability distribution at the time of *in silico* LAMP construction (Fig. 4d). Consequently, our *in silico* LAMP was demonstrated as useful to avoid unnecessary screening of LPSs having a low potential of being High-amp.

Recently, Endoh *et al.* [41] also studied prediction of LAMP amplification. The prediction model was built with an extra trees classifier that maximizes a modified F1 value, which is changed to precision and recall to sensitivity and specificity, using AutoML. A notable feature of this method is that explanatory variables are created by compression of the sequence order information of the LAMP primers and target templates. The compression can reduce computational resources considerably when we apply so many LAMP primers and genetic big-data to build a prediction model.

Our *in silico* LAMP is similar in this regard, finally creating only two very critical predictors of high LAMP amplification. However, our model applies the physical property of LAMP primers and does not apply sequence order information or target templates. By contrast, the model presented by Endoh *et al.* can apply the sequence order information. It is hoped that the physical properties will be applied eventually.

In fact, both models have potential predictors that have not yet been controlled. In addition, considering the fact that 5 of the predicted 10 High-amps were Low-amps in the prospective experiment, a multimodal model should be built with physical properties and sequence order information to improve the future prediction accuracy and robustness.

More notably, the template which was used differs between our study and that reported by Endoh *et al.* [41]. We used full-genome as the target sequence in our predictive model of LAMP amplification, whereas Endoh *et al.* [41] used synthetic sequences. Those sequences entail distinct benefits and tradeoffs. With a synthetic sequence, we can design each LPS against each synthetic sequence one-to-one and confirm a purely LAMP reaction, but preparing exactly the same number of copies of those templates is difficult. With full-genome, we can design each LPS on the full-genome itself and confirm the versatility and practicality LAMP reaction, but if the genome that we used had variants against the reference genome, then the LAMP reaction might not proceed because of the high specificity characteristic of LAMP. Although both target sequences include benefits and tradeoffs, we selected the full-genome as our target sequence because we emphasized creation of more generalizable predictors and a practical reduction in the number of *in vitro* experiments.

Despite these promising findings, this study has several limitations. First, general speaking, as the target sequence and LPS design regions become shorter rather than longer, LAMP is more likely to amplify. Therefore, although we use target sequences of sufficient length (980 bp), our *in silico* LAMP might not directly apply to uncommon, extremely short target sequences and the LPSs designed for them.

Second, constructing a dataset from a single chromosome 21 might bias the results compared to those of a dataset including other chromosomes. Additionally, using small datasets, the results require learning and validation using larger datasets from other various chromosomes.

Third, critical variables different from LAMP kernel variables might be necessary for High-amp in large datasets. Therefore, in future work, with the construction of a dataset with all chromosome regions, our model, including *in silico* LAMP, will be updated and validated. We expect to be able to create more robust variables, to build a more robust model, and to elucidate the LAMP reaction and LPS design further.

In summary, we constructed LPS and LAMP datasets and built EN-Model to identify critical variables influencing high amplification in the LAMP reaction. We generated LAMP kernel variables from domain knowledge and the identified critical variables for high amplification in the LAMP reaction. Using LAMP kernel variables and a logistic regression model, we built *in silico* LAMP with high classification accuracy. Our *in silico* LAMP facilitates the design of LPSs with high amplification and rapidity in LAMP reaction. We look forward to development sufficient that LAMP kernel variables might also be applied to other isothermal amplification methods.

Acknowledgments

We thank Yifei Li for helpful discussions related to this study.

Author contributions

K.T.: Conceptualization, Data Curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft. S.M.: Data Curation, Investigation, Resources, Writing—review & editing. T.T.: Supervision, Writing—Review & Editing. J.K.: Project Administration, Supervision, Writing—review & editing, Formal analysis, Methodology.

Supplementary data

Supplementary Data is available at *SYNBIO* online.

Conflict of interest: K.T. and S.M. are employees of Eiken Chemical Co., Ltd. The authors declare that they have no conflict of interest related to this report or the study it describes.

Funding

This work was partly supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grants (Nos 22H05108 and 24K10918) and Japan Science and Technology Agency (JST) ERATO Grant Number JPMJER2102.

Data availability

The data underlying the research described herein were provided by Eiken Chemical Co., Ltd. under license and by permission. Data will be shared on request to the corresponding author with the permission of Eiken Chemical Co., Ltd.

References

1. Notomi T, Okayama H, Masubuchi H *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res* 2000;**28**:E63. https://doi.org/10.1093/nar/28.12.e63

2. Nagamine K, Watanabe K, Ohtsuka K *et al*. Loop-mediated isothermal amplification reaction using a nondenatured template. *Clin Chem* 2001;**47**:1742–43. https://doi.org/10.1093/clinchem/47.9.1742
3. White TJ, Arnheim N, Erlich HA. The polymerase chain reaction. *Trends Genet* 1989;**5**:185–89. https://doi.org/10.1016/0168-9525(89)90073-5
4. Gibbs RA. DNA amplification by the polymerase chain reaction. *Anal Chem* 1990;**62**:1202–14. https://doi.org/10.1021/ac00212a004
5. Vosberg HP. The polymerase chain reaction: an improved method for the analysis of nucleic acids. *Hum Genet* 1989;**83**:1–15. https://doi.org/10.1007/BF00274139
6. Soroka M, Wasowicz B, Rymaszewska A. Loop-mediated isothermal amplification (LAMP): the better sibling of PCR?. *Cells* 2021;**10**:1931.
7. Tomita N, Mori Y, Kanda H *et al*. Loop-mediated isothermal amplification (LAMP) of gene sequences and simple visual detection of products. *Nat Protoc* 2008;**3**:877–82. https://doi.org/10.1038/nprot.2008.57
8. Han E-T, Watanabe R, Sattabongkot J *et al*. Detection of four Plasmodium species by genus- and species-specific loop-mediated isothermal amplification for clinical diagnosis. *J Clin Microbiol* 2007;**45**:2521–28. https://doi.org/10.1128/JCM.02117-06
9. Gelaw B, Shiferaw Y, Alemayehu M *et al*. Comparison of loop-mediated isothermal amplification assay and smear microscopy with culture for the diagnostic accuracy of tuberculosis. *BMC Infect Dis* 2017;**17**:79. https://doi.org/10.1186/s12879-016-2140-8
10. World Health Organization *The Use of Loop-mediated Isothermal Amplification (TB-LAMP) for the Diagnosis of Pulmonary Tuberculosis: Policy Guidance*. Genève, Switzerland: World Health Organization, 2016
11. LAMP primer designing software primerexplorer. https://primerexplorer.jp/e/ (21 February 2025, date last accessed).
12. Nagamine K, Hase T, Notomi T. Accelerated reaction by loop-mediated isothermal amplification using loop primers. *Mol Cell Probes* 2002;**16**:223–29. https://doi.org/10.1006/mcpr.2002.0415
13. Gandelman O, Jackson R, Kiddle G *et al*. Loop-mediated amplification accelerated by stem primers. *Int J Mol Sci* 2011;**12**:9108–24. https://doi.org/10.3390/ijms12129108
14. Jia B, Li X, Liu W *et al*. GLAPD: whole genome based lamp primer design for a set of target genomes. *Front Microbiol* 2019;**10**:2860. https://doi.org/10.3389/fmicb.2019.02860
15. Kronenberger JA, Wilcox TM, Mason DH *et al*. eDNAssay: a machine learning tool that accurately predicts qPCR cross-amplification. *Mol Ecol Resour* 2022;**22**:2994–3005. https://doi.org/10.1111/1755-0998.13681
16. Kayama K, Kanno M, Chisaki N *et al*. Prediction of PCR amplification from primer and template sequences using recurrent neural network. *Sci Rep* 2021;**11**:7493. https://doi.org/10.1038/s41598-021-86357-1
17. Kimura Y, de Hoon MJL, Aoki S *et al*. Optimization of turn-back primers in isothermal amplification. *Nucleic Acids Res* 2011;**39**:e59. https://doi.org/10.1093/nar/gkr041
18. Sanekata Y, Kayama K, Endoh T *et al*. Development of a LAMP simulation and selection pipeline to predict primer success. *Philipp J Vet Med* 2024;**61**:26–38.
19. Döring M, Kreer C, Lehnen N *et al*. Modeling the amplification of immunoglobulins through machine learning on sequence-specific features. *Sci Rep* 2019;**9**:10748. https://doi.org/10.1038/s41598-019-47173-w
20. Conard AM, DenAdel A, Crawford L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdiscip Rev Comput Stat* 2023;**15**:e1617. https://doi.org/10.1002/wics.1617
21. Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics* 2022;**12**:237. https://doi.org/10.3390/diagnostics12020237
22. Beckh K, Müller S, Jakobs M *et al*. Harnessing prior knowledge for explainable machine learning: an overview. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Raleigh, NC, USA. pp. 450–63. IEEE, 2023.
23. Radovanović S, Delibašić B, Jovanović M *et al*. Framework for integration of domain knowledge into logistic regression. In *Proceedings of the Eighth International Conference on Web Intelligence, Mining and Semantics*. New York, NY, USA: ACM, 2018.
24. Schneider VA, Graves-Lindsay T, Howe K *et al*. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res* 2017;**27**:849–64. https://doi.org/10.1101/gr.213611.116
25. Altschul SF, Gish W, Miller W *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2
26. Quyen TL, Ngo TA, Bang DD *et al*. Classification of multiple DNA dyes based on inhibition effects on real-time loop-mediated isothermal amplification (LAMP): prospect for point of care setting. *Front Microbiol* 2019;**10**:2234. https://doi.org/10.3389/fmicb.2019.02234
27. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 2003;**31**:3406–15. https://doi.org/10.1093/nar/gkg595
28. R: a language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria, 2024
29. Wu, Z Generalized linear models in family studies. *J Marriage Fam* 2005;**67**:1029–47. https://doi.org/10.1111/j.1741-3737.2005.00192.x
30. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol* 2005;**67**:301–20. https://doi.org/10.1111/j.1467-9868.2005.00503.x
31. Cessie SL, Van Houwelingen JC. Ridge estimators in logistic regression. *J R Stat Soc Ser C Appl Stat* 1992;**41**:191.
32. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol* 1996;**58**:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
33. King G, Zeng L. Logistic regression in rare events data. *Polit Anal* 2001;**9**:137–63. https://doi.org/10.1093/oxfordjournals.pan.a004868
34. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinf* 2006;**7**:91. https://doi.org/10.1186/1471-2105-7-91
35. Brodersen KH, Ong CS, Stephan KE *et al*. The balanced accuracy and its posterior distribution. In *2010 20th International Conference on Pattern Recognition*. Istanbul, Turkey. pp.3121–24 IEEE, 2010.
36. Seabold S, and Perktold J. Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the Ninth Python in Science Conference*. Austin, TX, USA. pp.92–96, 2010.
37. Akiba T, Sano S, Yanase T *et al*. (2019) Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, New York, NY, USA.
38. Minetti CASA, Remeta DP, Miller H *et al*. The thermodynamics of template-directed DNA synthesis: base insertion and extension enthalpies. *Proc Natl Acad Sci U S A* 2003;**100**:14719–24. https://doi.org/10.1073/pnas.2336142100

39. Breslauer KJ, Frank R, Blöcker H *et al*. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A* 1986;**83**:3746–50. https://doi.org/10.1073/pnas.83.11.3746
40. Dangerfield TL, Paik I, Bhadra S *et al*. Kinetics of elementary steps in loop-mediated isothermal amplification (LAMP) show that strand invasion during initiation is rate-limiting. *Nucleic Acids Res* 2023;**51**:488–99. https://doi.org/10.1093/nar/gkac1221
41. Endoh T, Sanekata Y, Kayama K *et al*. Development of machine learning algorithm for loop-mediated isothermal amplification including influence of temperature. *Sci Engg J* 2024;**17**:202–44.

Synthetic Biology, 2025, 10(1), ysaf007 , DOI: https://doi.org/10.1093/synbio/ysaf007, Advance Access Publication 31 March 2025, Research Article
Submitted: 26 March 2025; **Received (in revised form)**: 11 March 2025; **Accepted**: 26 March 2025