analytical chemistry

pubs.acs.org/ac

Article

Adaptive Filtering Framework to Remove Nonspecific and Low-Efficiency Reactions in Multiplex Digital PCR Based on Sigmoidal Trends

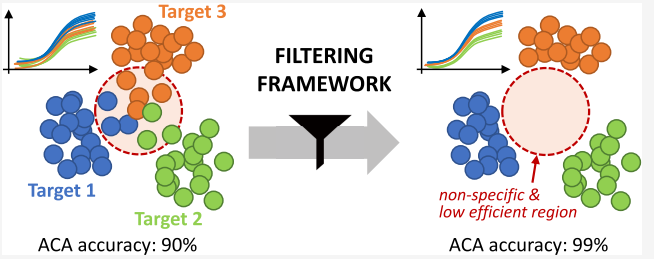
Luca Miglietta,[§] Ke Xu,[§] Priya Chhaya, Louis Kreitmann, Kerri Hill-Cawthorne, Frances Bolt, Alison Holmes, Pantelis Georgiou, and Jesus Rodriguez-Manzano*

Cite This: *Anal. Chem.* 2022, 94, 14159−14168

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Real-time digital polymerase chain reaction (qdPCR) coupled with machine learning (ML) methods has shown the potential to unlock scientific breakthroughs, particularly in the field of molecular diagnostics for infectious diseases. One promising application of this emerging field explores single fluorescent channel PCR multiplex by extracting target-specific kinetic and thermodynamic information contained in amplification curves, also known as data-driven multiplexing. However, accurate target classification is compromised by the presence of undesired amplification events and not ideal reaction conditions. Therefore, here, we proposed a novel framework to identify and filter out nonspecific and low-efficient reactions from qdPCR data using outlier detection algorithms purely based on sigmoidal trends of amplification curves. As a proof-of-concept, this framework is implemented to improve the classification performance of the recently reported data-driven multiplexing method called amplification curve analysis (ACA), using available published data where the ACA is demonstrated to screen carbapenemase-producing organisms in clinical isolates. Furthermore, we developed a novel strategy, named adaptive mapping filter (AMF), to adjust the percentage of outliers removed according to the number of positive counts in qdPCR. From an overall total of 152,000 amplification events, 116,222 positive amplification reactions were evaluated before and after filtering by comparing against melting peak distribution, proving that abnormal amplification curves (outliers) are linked to shifted melting distribution or decreased PCR efficiency. The ACA was applied to assess classification performance before and after AMF, showing an improved sensitivity of 1.2% when using inliers compared to a decrement of 19.6% when using outliers (p-value < 0.0001), removing 53.5% of all wrong melting curves based only on the amplification shape. This work explores the correlation between the kinetics of amplification curves and the thermodynamics of melting curves, and it demonstrates that filtering out nonspecific or low-efficient reactions can significantly improve the classification accuracy for cutting-edge multiplexing methodologies.



Downloaded via IMPERIAL COLLEGE LONDON on December 1, 2022 at 16:25:54 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

This manuscript demonstrates that undesired amplification reactions from real-time digital PCR (qdPCR) can be detected and filtered out by only evaluating the sigmoidal shape of an amplification curve. Here, we propose a novel methodology that can be used with multiplex PCR assays without the need for post-amplification analysis, increasing the result's accuracy and reliability.[1,2]

During the last decade, the gold-standard PCR technologies along with other nucleic acid amplification chemistries have resulted in key procedures for molecular diagnostics in both academic and clinical environments.[3−7] However, limitations such as sample availability, trained personnel, and overall laboratory costs can represent obstacles to the scalability and adoption of PCR-based approaches.[8,9] To overcome these barriers, multiplexing has been used to unlock the potential of conventional instruments, increasing the number of targets that can be detected in a single reaction.[10−12] Since the adoption of multiplexing techniques, researchers and industries have successfully applied them to different areas such as molecular diagnostics, RNA signature polymorphism, and quantitative analysis.[13−15] Moreover, in an effort to increase the overall multiplex PCR capabilities, several studies have recently been published on the use of machine learning (ML) to identify the biological nature of an amplification event, improving throughput, clinical and analytical reliability, and sample classification accuracy.[16,17] As described by Athamanolap et al. in 2014, ML methods were applied to high-resolution melt

Received: April 29, 2022
Accepted: September 22, 2022
Published: October 3, 2022



ACS Publications

© 2022 The Authors. Published by American Chemical Society

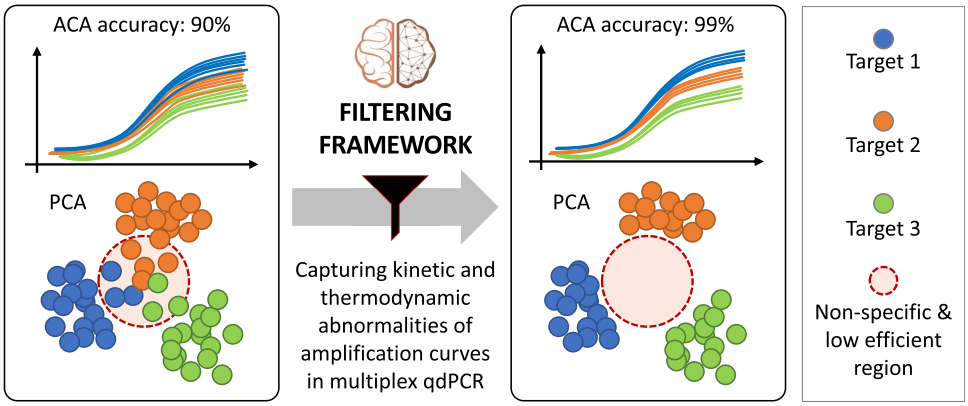


Figure 1. Concept figure. Left: raw amplification curves and their corresponding ACA clusters (represented by principal component analysis or PCA) include nonspecific and low-efficient reactions (confined in the red-circled region). The presence of outliers blurs the boundaries of the different clusters, negatively impacting ACA classification accuracy. By applying the proposed filtering framework, kinetic and thermodynamic abnormalities from amplification events can be captured. Right: outliers are removed from the original data, resulting in more separated clusters and clearer boundaries. Therefore, ACA classification accuracy is improved.

curves to increase both the tolerance of melting temperature (T_m) deviation among targets and the reliability of classification for genetic variants (such as polymorphic genetic loci).[18] In Jacky et al., ML techniques were used to enable high-level multiplexing using TaqMan probes by leveraging on single-feature classification (i.e., final fluorescence intensity or FFI) and PCR platforms with multiple fluorescent channels.[19] While data-driven methods have mostly been employed to improve the accuracy of target identification, with the aim to increased multiplexing capability, some groups have also explored such techniques for outlier removal, both in digital and in bulk PCR. For instance, Yao et al.[20] developed a process-based classification model to identify false-positive curves in dPCR (leading to a 64% improvement compared with classical techniques), and Burdukiewicza et al.[21] developed an algorithm to automatically detect hook effect-like curvatures, allowing for streamlined quality control in qPCR.

Recently, Moniri et al. in 2020 proposed a new approach called amplification curve analysis (ACA) for single-channel multiplexing without explicitly extracting features.[22] The ACA method comprises a supervised ML classifier to analyze kinetic information encoded in the entire amplification curve by looking into sigmoidal shapes across different targets.[22,23] Furthermore, using ACA along with melting curve analysis (MCA),[24] a new method called amplification and melting curve analysis (AMCA) was developed, enabling higher-level multiplexing in a single channel. While the melting curve is determined by the thermodynamic properties of the amplicon, mainly related to its nucleotide sequence, the features of the amplification curve are also influenced by the concentration of templates and amplicon, as well as PCR efficiency (and its cycle-to-cycle variation), thus also providing information on the kinetics of the amplification reaction. The AMCA couples both ACA and MCA coefficients from the classifier to improve classification accuracy. This has been demonstrated through the detection of nine mobilized colistin resistance genes and clinical isolates containing five common carbapenemase resistance genes.[25,26] Moreover, multiplex PCR (coupled with innovative approaches such as ACA or AMCA) is bringing about a change of paradigm in molecular diagnostics by enabling faster, more accurate, and higher-throughput detection of several biomarkers in one reaction. Its applications are wide ranging, including precision medicine in cancer, genetic testing, and syndromic testing in clinical microbiology and infectious diseases, where it enables precise multitarget identification of multiple pathogens and antimicrobial resistance genes.

A barrier to wider adoption of the aforementioned techniques is that they may be limited by instrumentation specifications such as thermal profile performance, available optical channels/filters, and software setup. For example, MCA methodologies are particularly limited in point-of-care devices, as many do not have melting curve capabilities. Furthermore, in assays based on probe-based chemistries (such as TaqMan), where intercalating dyes are not present, the melting curve cannot be generated. In these circumstances, the ACA method still stands as a valid option for multiplexing and therefore it has been the methodology of choice for the work proposed in this manuscript. However, across all of these ML-based multiplexing strategies, the ACA approach can be negatively affected by the presence of abnormal amplification products due to primer dimerization, amplification of undesired targets, the miscalibration of the instrument, and intramolecular secondary structures. These abnormal behaviors tend to alter the kinetic information of the sigmoidal curves, causing low efficiency or delaying the amplification reaction.[27−29] As represented in Figure 1, when considering shapes of amplification curves from a multiplex assay, similarities among different targets can reduce the accuracy of the ACA classifier, as the presence of nonspecific or low-efficient reactions results in blurred boundaries among clusters. To overcome this problem, we developed an intelligent algorithm to filter out outliers from multiplex amplification events. Furthermore, to validate the correctness of outlier removal, amplification curves (inliers and outliers) are compared with labeled melting curves (correct and wrong).

In this work, we demonstrated that nonspecific and low-efficient PCR reactions affect the shape of the amplification curve and therefore they can be filtered out considering only the sigmoidal trend. Furthermore, we developed an outlier removal algorithm called adaptive mapping filter (AMF), which in combination with the ACA approach was used to improve the multitarget classification accuracy. This represents a step forward to incorporate ACA in clinical applications and ensure that by filtering in correct amplification curves, higher diagnosis reliability is delivered to the patient. These concepts were explored using data obtained from qdPCR experiments reported by Miglietta et al.[26] As a case study, three of the "The big 5"

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

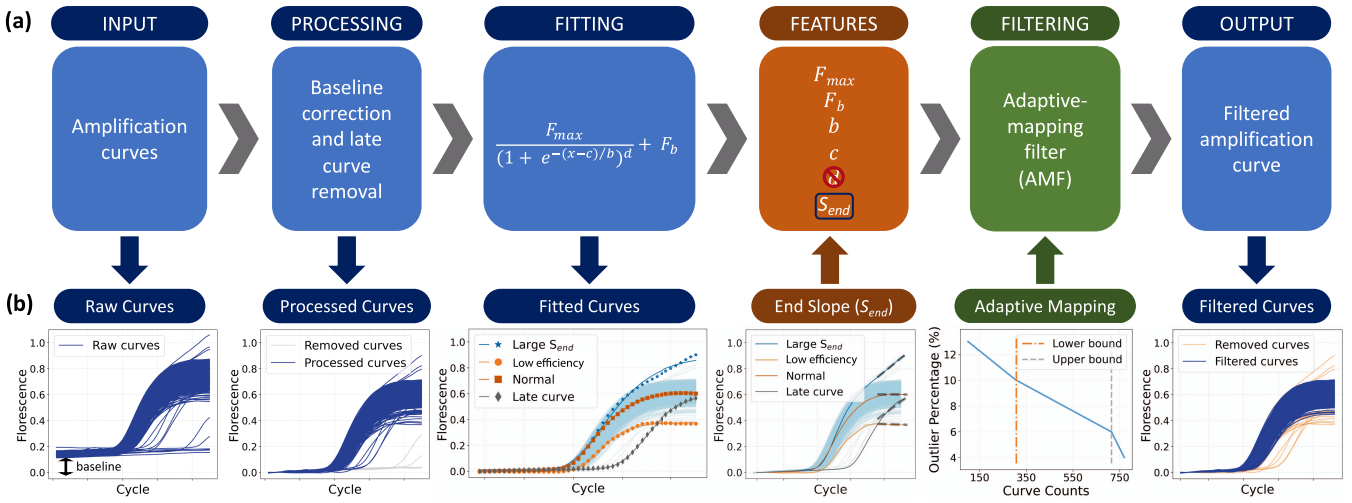


Figure 2. Proposed framework. (a) Framework steps: raw data input, processing, curve fitting, feature extraction, adaptive mapping filtering (AMF), and filtered curve output. (b) Input or output of each step. From left to right, the input of the framework was raw amplification curves, some of which are flat or late curves. By applying the processing step, the baselines were removed, and flat/late curves were discarded. Following this, the processed curves were fitted using a five-parameter sigmoid function, after which each curve was condensed into five features. A new feature S_{end} plus four of the parameters was used to form a set, which is the input of the filtering step. The d parameter was discarded from the feature set for filtering as it is unsuitable for the used algorithms. We further developed the AMF with a monotonic decreasing map between positive curve numbers within a panel and the outlier percentage. The outputs of the framework are the curves after filtering (inliers).

carbapenemase genes (bla_{NDM}, bla_{IMP}, and bla_{OXA-48}) were considered in this study.

Our vision is that by sharing this new approach we can significantly improve the quality of data from qdPCR instruments and enhance the sensitivity and accuracy of ML-based multiplexing methods relying only on amplification curves. Moreover, extending this framework to other amplification chemistries and real-time platforms will improve the multiplexing capabilities of existing diagnostic workflows and platforms.

EXPERIMENTAL SECTION

In this section, a new framework for outlier removal in qdPCR is proposed. As depicted in Figure 2, this framework took raw amplification curve data as the input and applied baseline and flat/late curve removal in the processing step. Then, each processed curve was fitted by a sigmoid function and the fitted parameters, as well as a newly developed feature referred to as S_{end}, were used as the input for a filtering algorithm, which identified outliers automatically. Finally, the framework outputs the amplification curves after filtering, marked as inliers.

Data Input. As a case study, data from Miglietta et al. were used in this work.[26] Data from synthetic DNA (gBlocks gene fragments, IDT) containing bla_{NDM} (N = 18,480), bla_{IMP} (N = 17,710), and bla_{OXA-48} (N = 17,710) gene sequences were used as the training data set. From the original study, a total of 198 clinical isolates labeled with these three targets were used as the testing samples to maintain a balanced data set and due to their high prevalence and clinical significance in U.K. hospitals. Each sample contained 770 raw curves for a total of 152,460 curves across all of the samples, within which 116,222 were positive after the processing step. It is expected that data from clinical isolates are much noisier and thus contain more outliers than those from gBlocks.

Data Processing. The first step of the framework is processing the raw curves using a baseline correction and a flat/late curve removal to exclude the negative curves of the unprocessed data from the qdPCR output. The baseline of the real-time PCR reaction during the initial cycles presents little change in the fluorescent signal. The low-level signal of the baseline equates with the background or noise of the reaction. Therefore, we processed the baseline of each raw curve by averaging the fluorescent value of the first five cycles and subtracting it from the time series. Following this, flat/late curves were removed by applying an upper and lower fluorescence threshold at the 40th cycle, as suggested by the manufacturer.[30,31]

Fitting and Feature Extraction. Following the processing step, a curve fitting step was introduced to represent the processed amplification curves with sigmoid parameters, which were later downselected and used as input features for outlier removal and classification algorithms. A 5-parameter sigmoid model,[29] which is shown below, was used to fit the amplification curves

$$F(t) = F_b + \frac{F_{max}}{(1 + e^{-(t-c)/b})^d}$$

where t is the PCR cycle number, $F(t)$ is the fluorescence at the tth cycle, F_b is the background fluorescence, F_{max} is the maximum fluorescence, b relates to the slope of the curve, c is the fractional cycle of the inflection point, and d is the asymmetric parameter. To solve the nonlinear least-squares optimization problem for the curve fitting, the trust region reflective (TRF) algorithm with specific bounds was used.[32] Here, we set $[10, 0.3, 10, 50, 100]$ and $[0, -0.1, -10, -50, -10]$ as the upper and lower bounds for the search of the 5-parameter set $p = [F_{max}, F_b, b, c, d]$, respectively. The initial parameter set p_0 was optimized through pivot fitting on 5% of the training data. After fitting, each amplification curve was given as five parameters, which are condensed representations of curve information. The fitting quality was assessed using mean-squared error (MSE) and is reported in Figure S1. All parameters except for d were considered input features for outlier removal algorithms because parameter values of outliers may have significant differences

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

from those of normal curves. The d parameter shows a bimodal distribution with two distant peaks, which is unsuitable for the outlier removal step because many of the outlier algorithms require a unimodal distribution of features. Therefore, the d parameter was discarded from the feature set for filtering.

In addition, we further introduced a new feature called the end slope (S_{end}), with the aim to provide further information about the amplification curve shape. This was calculated by taking the average of the first derivatives at the last five cycles of the amplification curve

$$S_{end} = \frac{1}{5}[D(N-4) \quad D(N-3)] \quad \cdots \quad D(N)\boldsymbol{e_5}^T$$

where

$$D(x) = \frac{\mathrm{d}\,F(t)}{\mathrm{d}t}$$

$$\boldsymbol{e}_5 = [1 \quad 1 \quad 1 \quad 1 \quad 1]$$

and N is the total cycle number.

Using the S_{end} feature, the information in the tail of amplification curves was extracted, which contributes to distinguishing inliers and outliers. For example, as illustrated in the "fitting curves" step of Figure 2b, curves that do not reach the plateau may have larger end slopes. These curves cannot be precisely represented by the fitted parameters since the fitting equation is not capable of capturing this nonplateaued trend. Therefore, S_{end} would benefit from the result of the outlier removal by providing additional information to the feature set. Including S_{end} and discarding d, the final feature set for outlier removal algorithms is $\boldsymbol{x}_f = [F_{max}, F_b, b, c, S_{end}]$.

Outlier Removal Algorithms. In this research, seven outlier removal algorithms were considered, which can be split into the following categories according to their principal ideas of filtering: proximity-based, linear, outlier ensembles, and angle-based algorithms. (i) Proximity-based outlier detection algorithms rely on using a distance metric (e.g., Euclidean or Manhattan) to identify outliers. We applied two proximity-based algorithms, which are local outlier factor (LOF) and density-based spatial clustering of applications with noise (DBSCAN).[33,34] The LOF algorithm considers the k-nearest neighbors (KNNs) to every point in the data set and computes a local outlier factor for each of them. DBSCAN classifies the points into the core, border, and noise of clusters based on the number of points (min points) within the radius (epsilon) of the considered point. (ii) The linear outlier detection methods used were one-class support vector machine (OC-SVM) and elliptical envelope.[35,36] OC-SVM applies the concept of finding a hyperplane that separates the inlier points from the origin, such that the hyperplane is closest to the inlier points as possible. The elliptical envelope aims to fit the smallest ellipse possible to the core cluster of data points, with any point outside being considered outliers. (iii) Outlier ensemble-based detection methods considered were isolation forest and feature bagging.[37,38] Isolation forest uses random forests to recursively randomly partition data, after which data points with fewer partitions to isolate are marked as outliers. Feature bagging considers multiple outlier algorithms and randomly selects a group of features. From those features, the resulting outlier scores from each algorithm are merged to find the strongest outliers. (iv) Angle-based outlier detection considers the angles made by a point with all other pairs of points in the data set.[39] For each point, the variance is calculated from all of the angles obtained, where for a potential outlier the variance is small since the point is distant from the main cluster of data.

Adaptive Mapping Filter (AMF). Most of the outlier detectors explained in the previous section require a hyperparameter called "contamination ratio" or "outlier percentage", which represents the percentage of outliers to be removed from the original data. To adaptively set up this hyperparameter, we developed a mapping strategy that maps the number of positive reactions per panel in the qdPCR chip (processed curves) to the contamination ratio used in the outlier removal algorithm.

In digital PCR, as the number of positive curves increases, the probability of having more than one molecule in a single well increases, resulting in a shift of reaction state from digital to bulk. Moreover, as the reaction goes toward the bulk region, a higher number of positive curves will be present in a panel, which can result in a lower probability of observing a nonspecific or low-efficient reaction (outlier) in a well.[22,40] Let us suppose that for each well the probability of observing an outlier is $p(M_i)$, where M_i is the number of processed curves for the ith sample. Since $p(M_i)$ are independent and identically distributed (IID) for all of the wells, the total number of outliers X_i observed in the ith sample follows the distribution of $X_i \sim B(M_i, p(M_i))$. Therefore, the expected percentage of outliers in the ith sample should be

$$\text{outlier percentage} = \frac{E(X_i)}{M_i} = \frac{M_i p(M_i)}{M_i} = p(M_i)$$

which means that the expected outlier percentage is a monotonic decreasing function to the number of positive curves. In this research, we applied a piecewise linear function with empirical turning points, as illustrated in the filtering step of Figure 2b.

Coupling the adaptive mapping with an outlier removal algorithm, we developed a novel method called adaptive mapping filter (AMF), which takes as input the feature set and output the inliers.

Melting Labeling. An algorithm was developed to automatically label the melting curves as specific (which we called "correct") or nonspecific (referred to as "wrong") ones. By using this methodology, the percentage of wrong melting curves within all of the curves of a sample (wrong melting percentage or WMP) was calculated, and this WMP further served as a metric for performance evaluation.

To apply melting labeling, the reference melting peak for each target needs to be determined. For a target tg$\in$ [bla_{NDM},bla_{IMP},$bla_{OXA\text{-}48}$], a reference melting peak temperature T_m^{tg} was given by calculating the median value of all of the melting peak temperatures of the gBlock curves with target tg. After that, the steps below were followed to label every single melting curve of the clinical data set:

(1) Find the global maximum melting peak's temperature T_m^g of the current melting curve.

(2) If $T_m^g \notin \left[T_m^{tg} - \frac{W}{2}, T_m^{tg} + \frac{W}{2}\right]$, where W is the tolerance width of the T_m^g distribution, the current curve is labeled directly as a wrong melting curve. Here, considering our instrument resolution for melting curve analysis, our W is equal to $\pm$0.5 °C.

(3) Otherwise, find the local maximum melting peaks' temperatures on the left and right sides of T_m^g on the current curve and mark them as T_m^l and T_m^r, respectively. Note that either T_m^l or T_m^r may not exist. If neither exists, the current curve will be labeled as a correct melting curve.

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

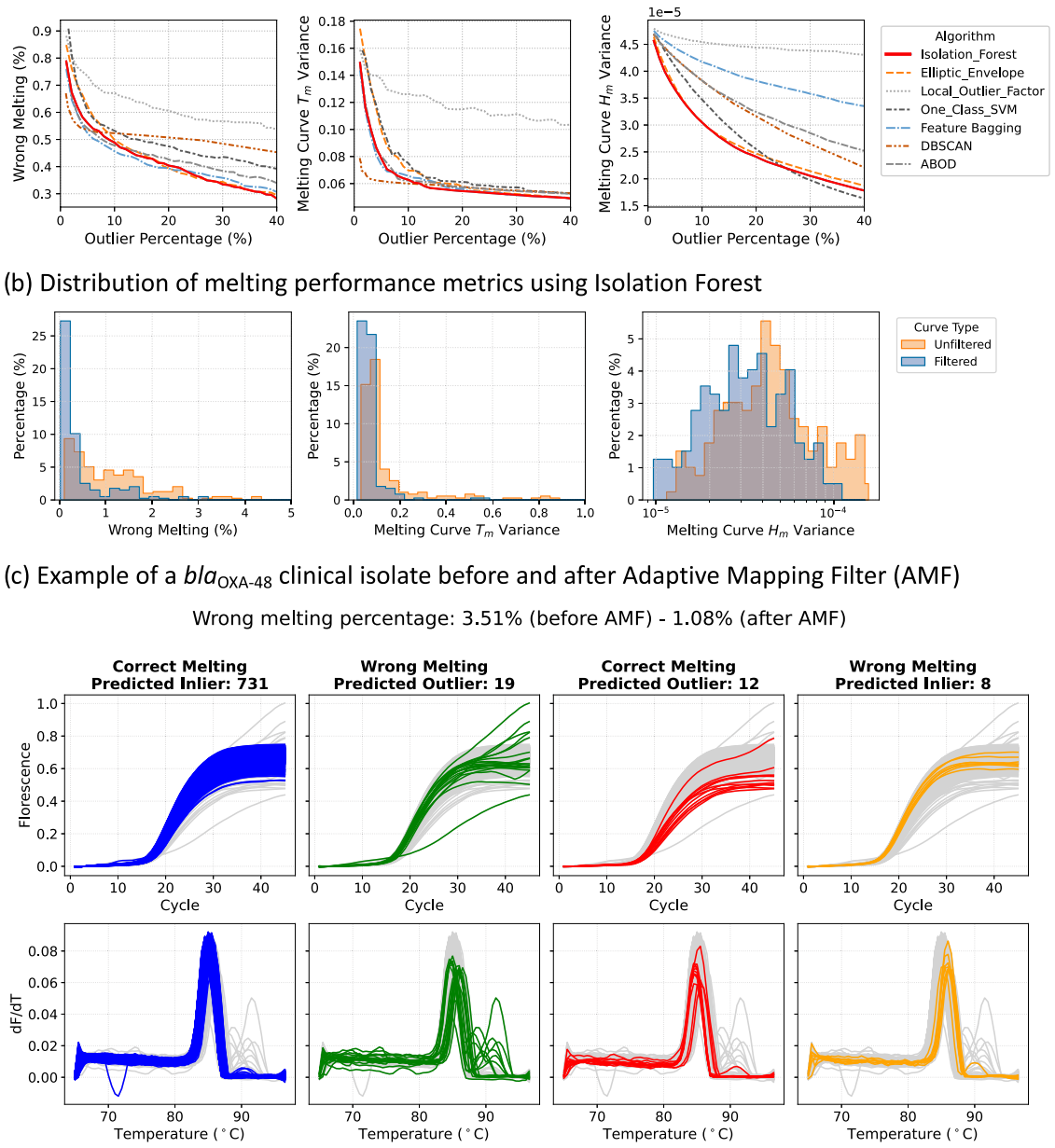


Figure 3. Melting curve analysis on filtering results. (a) Melting performance shown with wrong melting percentage (WMP), T_m, and H_m variances vs fixed outlier percentage. As the outlier percentage increases, all of the metrics show decreasing trends, which tend to plateau after a certain percentage. As illustrated by the firm red line, isolation forest performs the best overall for the three metrics. (b) The distribution of melting performance metrics shows that, after filtering, the WMP becomes significantly smaller, and T_m and H_m have a narrower distribution. (c) An example of bla_{OXA-48} clinical isolate. Each column shows the amplification curve and corresponding melting curve of the correct melting and predicted inliers (N = 731), wrong melting and predicted outliers (N = 19), correct melting and predicted outliers (N = 12), and wrong melting and predicted inliers (N = 8).

(4) If at least one of T_m^l and T_m^r exists, a set of this (these) local melting peak(s) will be constructed. For each element T_m^e in this set, check whether

$$H_e \in [\, H_{mean} - 4H_{std}, H_{mean} + 4H_{std} \,]$$

where H_e is the height of the current melting curve at temperature T_m^e, H_{mean} and H_{std} are the mean and standard deviation of $[H_{1,T_m^e} H_{2,T_m^e} \cdots H_{M,T_m^e}]$, respectively, in which H_{n,T_m^e} means the height of the nth melting curve of the sample at temperature T_m^e, and M is the total curve number in the sample. If at least one of the above tests fails, the current curve will be labeled as a wrong melting curve. Otherwise, it will be marked as the correct one.

With the above steps, it is ensured that both curves with large deviations of T_m^g from reference melting peaks and curves with large nonspecific local melting peaks can be labeled as wrong. In this way, all of the curves had been marked as either "correct" or "wrong" and further used to calculate the wrong melting percentage (WMP)

$$\mathrm{WMP} = \frac{N_{wrong}}{N_{total}} \times 100$$

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

where N_{wrong} is the number of wrong melting curves within the sample, and N_{total} is the total number of curves in the sample.

It is worth mentioning that the proposed algorithm of automatic melting labeling is not a part of the filtering framework. The labeling was used to calculate the WMP, which functioned as a metric for filtering evaluation, where a lower WMP indicates better filtering performance.

Data Visualization. Visualization is a vital step for understanding the distribution of a given data set. In this article, principal component analysis (PCA) with two components was used to visualize the feature sets of the curves before and after applying the outlier removal algorithm to scatter plots. Visual inspection was performed to illustrate how separated the clusters of different targets were. Following this, several metrics for measuring the density and degree of separation among those clusters were used to quantitatively evaluate how well they were divided.

Specifically, after the PCA of the feature set $\boldsymbol{x}_f = [F_{max}, F_b, b, c, S_{end}]$ from the amplification curves of each target, the Silhouette coefficient for each feature set was calculated.[41] The mean value of these coefficients, known as the mean Silhouette score, was then used to indicate how well the curves of the same targets are clustered. A higher Silhouette score implies denser and better-separated clusters observed. Two additional metrics, the Calinski−Harabasz score and the Davies−Bouldin score, were also implemented for clustering evaluation, where a higher Calinski−Harabasz score or a lower Davies−Bouldin score relates to larger intercluster distances among targets.[42,43]

Classification of Amplification Curves: Data-Driven Multiplexing. The ACA method uses kinetic information encoded in the amplification curve to classify different nucleic acid molecules from a PCR test. The performance of the ACA was assessed using different curve representations (Table S1), and the five fitted parameters were used in this study. To illustrate the influence of the AMF on the ACA, a random forest classifier with 100 trees was applied to the feature set $\boldsymbol{x}_c = [F_b, F_{max}, b, c, d]$, which differs from the $\boldsymbol{x}_f$ used for outlier removal algorithms. Here, parameter d was reintroduced because more curve-related information is needed, provided that the proposed classifier is relatively less sensitive to the feature distributions. S_{end} was discarded for classification because, after outlier removal, abnormal curves with large end slopes were not present in the data set. For the remaining curves, S_{end} was extremely close to zero; thus, it was not necessary for S_{end} to be included again. All of the other features were normalized with the mean and the variance of the training data before being input into the classifier.

In this research, after applying data processing and feature extraction on both training and testing sets, the extracted features of the training set were used to train a random forest classifier. This trained classifier was then evaluated on the testing set with or without adaptive mapping filtering (the progress of AMF is totally unsupervised, so it can be applied to the testing data set without the true labels). For the testing set, we utilized both the inliers and the outliers marked by the aforementioned AMF algorithm and tested them. As a comparison, two randomly downselected data sets with the same number of curves as the inliers and the outliers were also constructed and tested.

Statistical Analysis. Two-sided Wilcoxon signed-rank tests were used to determine the statistical significance of the changes in WMP and melting peak distributions (distributions of the melting peak temperature, T_m, and height, H_m) before and after outlier removal. Two-sided Mann−Whitney U rank tests were used to compare the distributions of C_t, FFI, and maximum slopes between inlier and outlier amplification curves. Those three metrics were chosen for their relationship with the amplification curve efficiency. Many studies suggest that sigmoidal modeling of the entire amplification curve can be used to define the rate of PCR efficiency. Therefore, low-efficient PCR reactions are related to low fluorescent values and low maximum slope.[44,45]

Moreover, the significance of the comparison between inliers and outliers in clustering Silhouette coefficients was determined by a two-sided Wilcoxon signed-rank test. This test was also used in the evaluation of the classification performance. A p-value of 0.001 with Bonferroni correction was used as the threshold for statistical significance.

RESULTS AND DISCUSSION

In this study, a new framework is presented to detect outliers from amplification reactions in qdPCR. The outlier identification relies on the AMF, which is comprised of an outlier detection algorithm and a mapping strategy to adapt the contamination ratio hyperparameter to the positive amplification reaction counts (or positive wells) of the qdPCR chip.

Evaluation of Outlier Detection Algorithms. As shown in Figure 3a, we evaluated the detection performance of seven outlier removal algorithms on filtering amplification curves against outlier percentages by using three metrics: (i) wrong melting percentage (WMP), (ii) melting curve T_m variance, and (iii) melting curve H_m variance. The changing values of metrics for different algorithms with fixed outlier percentages from 0.1 to 40% are shown in Figure 3a. After the filtering is applied, the WMP shows a significant reduction from 1.1% (from the unfiltered data set) to a maximum of 0.9% after filtering across all of the algorithms. The graph depicts that outlier percentage and WMP are inversely proportional, but the trend can vary among methods. Proximity-based outlier detectors perform worse overall compared to the rest, so they are unable to achieve a dramatic decrease in WMP, even with very large contamination ratios. On the other hand, ensemble-based detectors such as feature bagging and isolation forest have better performance with the lowest WMP among all of the outlier percentages. As shown in the center and right end graphs, the variances of T_m and H_m have a decreasing trend that can be observed as the outlier percentage increases, indicating that both of their distributions are narrowed down. In the T_m variance plot, it is noticed that DBSCAN achieves better performance at lower outlier percentages, but this trend reaches a plateau as the outlier percentage further increases. Once again, ensemble-based methods have similar behavior for the T_m variance as for the WMP. For instance, isolation forest outperforms all other detectors after the outlier percentage reaches 12%. Moreover, isolation forest and elliptic envelop show the best performance for H_m variance up to 26% contamination ratio.

In this analysis, WMP was used to show the change of wrong melting proportion after applying outlier detection algorithms, indicating the direct effect of the filtering on removing wrong melting curves. It is important to consider that we do not relate wrong meltings with wrong target sequences as the true nature of the amplicons resulting from the PCR reaction can only be established by sequencing, which is impractical in digital PCR. The WMP is used to evaluate the shift of melting peak or the presence of multiple low-intensity peaks, which result from nonspecific or low-efficiency amplification reactions. This can

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

Table 1. Comparison of C_t, FFI, and Maximum Slope between Predicted Inliers and Outliers with Correct Melting Peaks[a]

	C_t(mean ± std)		FFI (mean ± std)		max slope (mean ± std)	
target	inliers	outliers	inliers	outliers	inliers	outliers
bla_{NDM}	21.45 ± 3.28	26.40 ± 6.11	0.67 ± 0.06	0.60 ± 0.12	0.07 ± 0.01	0.06 ± 0.01
bla_{IMP}	30.33 ± 2.23	31.25 ± 3.43	0.44 ± 0.06	0.41 ± 0.07	0.0276 ± 0.003	0.0271 ± 0.01
bla_{OXA-48}	18.82 ± 3.08	21.03 ± 4.34	0.65 ± 0.08	0.51 ± 0.16	0.05 ± 0.01	0.04 ± 0.02

[a]For all of the targets, inliers have significantly smaller C_t and larger FFI and max slope, with all p-values < 0.0001.

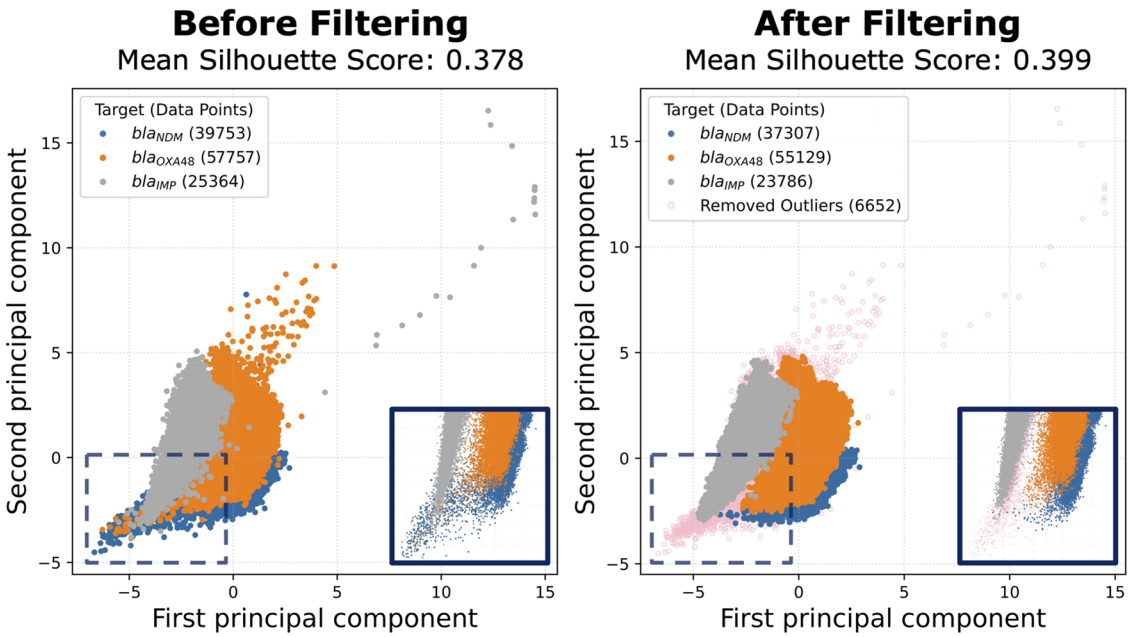


Figure 4. Data visualized using two-dimensional (2D) principle component analysis before and after filtering. The processed data plot shows that most outliers have been removed from the original unfiltered data, and the clusters are more separated with clearer boundaries and fewer overlaps. The segmented squares on the bottom side of both figures show the areas where cluster overlapping is more evident; thus, they are zoomed. The mean Silhouette score increases from 0.378 to 0.399 after filtering.

largely affect the ACA classification depending on the presence of the abovementioned phenomena; therefore, filtering such events can result in improved target identification. Moreover, a smaller T_m variance indicates a narrower T_m distribution, which in combination with the WMP methods shows that curves with large deviations from the reference T_m^{tg} are removed by the filtering algorithm. In molecular biology, those curves may be generated after nonspecific events such as undesired target interaction or primer dimerization.[46] In addition, melting curves presenting low $-df/dt$ (or H_m) are associated with low-efficient amplification reactions. Therefore, the narrowed distribution of H_m indicates that low-efficient curves, which are present at the tail of the distributions, are removed.[47] All of the algorithms provide better performance compared to the original benchmark calculated on the unfiltered data. However, it is noticed that isolation forest is always among one of the best methods for all of the metrics and does not show any defects, which is common for other algorithms (outlier distribution reported in Figure S2). In the following sections, we use isolation forest to further demonstrate the proposed framework.

Filtering Performance Analysis of the AMF. In the following step, AMF was applied to the unfiltered data, and the distributions of inner-sample WMP, T_m, and H_m variances are illustrated in Figure 3b. Across these three metrics, significant shifts of distributions to smaller values are shown after filtering, supported by all of the p-values < 0.0001. This indicates that the proposed AMF can significantly remove both nonspecific and low-efficiency curves only by looking at amplification curves. This proves our hypothesis that amplification curves contain not only kinetic but also thermodynamic information as numbers of outliers correspond to wrong melting curves.

An example of the AMF visual performance on a clinical isolate containing the carbapenemase gene bla_{OXA-48} is illustrated in Figure 3c. Columns represent both amplification and melting curves of (i) correct melting and predicted inliers (N = 731, 94.9%), (ii) wrong melting and predicted outliers (N = 19, 2.5%), (iii) correct melting and predicted outliers (N = 12, 1.6%), and (iv) wrong melting and predicted inliers (N = 8, 1%). The first column shows the correctly identified inliers representing specific products of PCR tests. In the second column, nonspecific reactions are correctly identified and labeled as outliers, which emphasizes the effectiveness of the filtering. We noticed that a small number of specific curves were predicted as outliers, as shown in the third column of Figure 3c. This phenomenon does not deny the efficacy of the filter, as these "incorrectly" removed curves have (i) significantly larger C_t values, (ii) significantly smaller FFI, (iii) and smaller values of maximum slope compared to the inliers. Across the entire clinical isolate data set (N = 116,222), compared to melting curve analysis, 115,535 were correctly predicted inliers and 791 were correctly predicted outliers. Furthermore, 5,861 were wrongly classified as outliers, whereas 687 were wrongly classified as inliers. Further statistical analyses on the entire data set also endorse these significant differences between inliers

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

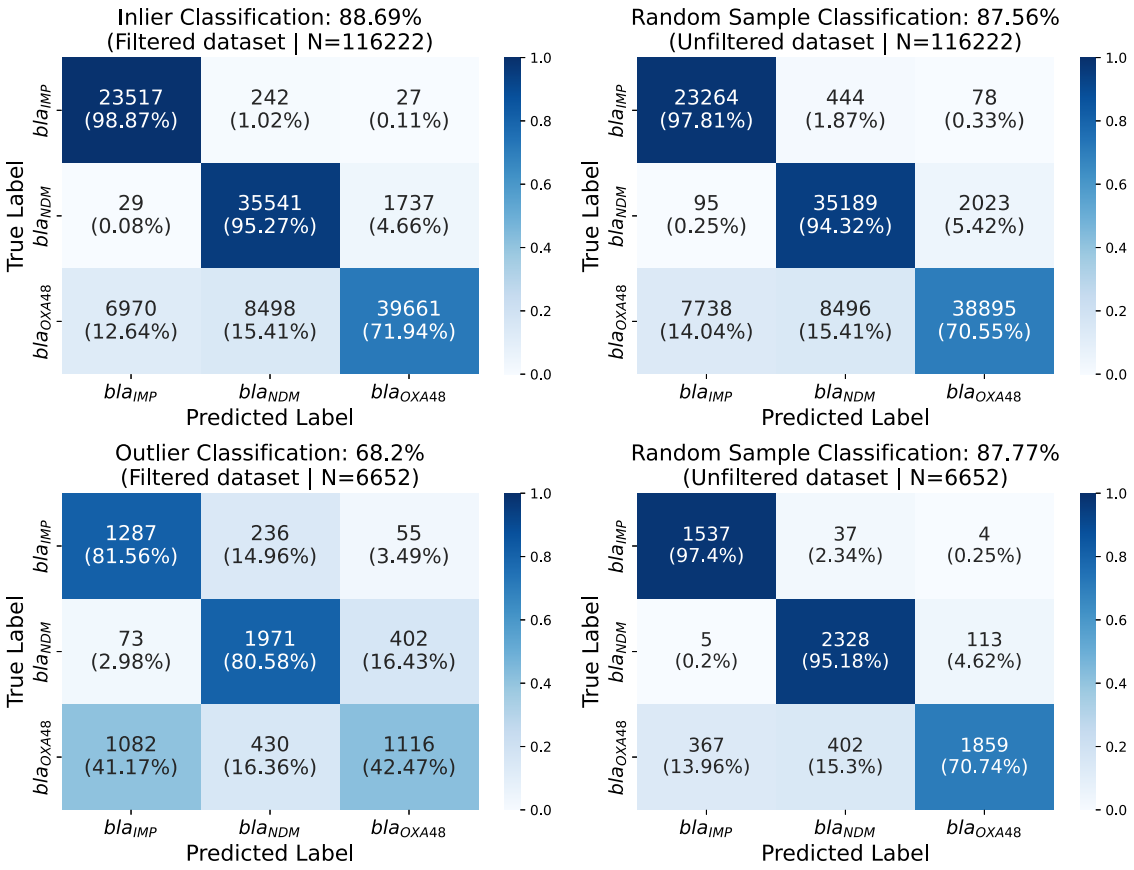


Figure 5. Confusion matrices for inlier and outlier classifications. The four confusion matrices are shown for (a) inliers, (b) randomly downselected data with the same curve numbers as inliers, (c) outliers, and (d) randomly downselected data with the same curve numbers as outliers. The title of each matrix reports the sensitivity of the model. Moreover, each square of the matrix has the number of predicted curves for the corresponding true label and the respective sensitivity of the square.

and outliers for C_t, FFI, and maximum slope values, as illustrated in Table 1. This indicates that AMF removes certain curves because they are of low amplification efficiencies even though they have "correct" melting peaks. A few curves labeled as "wrong" melting may be predicted as inliers, as shown in the fourth column of Figure 3c. This can be explained by the relatively low-temperature resolution of the equipment, which results in mislabeled wrong melting curves due to the large quantization noise of T_m^g during temperature measurement. In fact, by visually inspecting the last column of Figure 3c, it can be seen that amplification curves are of very similar shapes to correctly predicted inliers (shown in the first column of Figure 3c). The WMP of the illustrated sample has dropped from 3.51 to 1.08%. Overall speaking, in this demonstrated data set, 1.2% of wrong meltings were reported before filtering, and after applying AMF, we reduced the WMP by half to 0.59%.

Feature Set Visualization. To visualize the effect of the AMF, PCA-based feature visualization before and after filtering is depicted in Figure 4. On the left of the figure, the unfiltered data shows larger overlapping within clusters of different targets and a higher number of outliers compared to the data after filtering. The segmented squares are used to emphasize the differences in cluster overlapping before and after the AMF, where clearer boundaries between bla_{IMP} and both bla_{OXA-48} and bla_{NDM} can be seen. These differences highlight that (i) outliers can be effectively removed by the AMF and (ii) removing outliers enhances the separation and reduces the overlap among different target clusters, which will ease the classification of the ACA method.

To numerically evaluate the degree of separation across target clusters, the mean Silhouette score of all of the data points was calculated before and after filtering, showing an increment from 0.378 to 0.399 (p-value < 0.0001). In addition, the Calinski–Harabasz score increased from 101,002.729 to 130,134.802, and the Davies–Bouldin score dropped from 0.886 to 0.839. All of those results indicate that AMF makes target clusters denser and better separated.

ACA Classification. After demonstrating that removing outliers improves the overall distance among clusters, we further explored its impact on the ACA classification for both inliers and outliers against randomly downselected data sets with the same numbers of curves. In Figure 5a, the confusion matrix shows that the sensitivity for the inliers is 88.96%, which is an increase of 1.13% compared to the randomly downselected ones (Figure 5b). For all of the targets, a significant sensitivity improvement can be observed of 1.06, 0.95, and 1.39% for bla_{IMP}, bla_{NDM}, and bla_{OXA-48}, respectively. Moreover, the overall classification accuracy was 84.94% for inliers and 83.76% for randomly downselected curves, showing a 1.18% improvement (p-value < 0.0001), which is in line with the overall WMP before filtering (WMP = 1.2%). Applying the filter will help increase the overall performance and specificity of the data set. This supports our hypothesis that melting information or thermodynamics are contained in the amplification curve.

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168

To show that the removed outliers are less informative for target recognition and harmful for the overall classification, in Figure 5c,d, we show the confusion matrices of the classification using both removed outliers and a randomly downselected data set with the same size. As expected, the performance for outliers is significantly worse than the randomly downselected ones, with only 68.2 and 54.78% sensitivity and accuracy, respectively, for outliers (p-values < 0.0001). This dramatic sensitivity decrement of 19.57% strongly suggests that outliers have less useful information for the classification of the selected targets.

Furthermore, the statistical analysis of the two randomly downselected data sets shows no significant differences in in-sample accuracy with a p-value of 0.448, which is in line with the central limit theorem as they originate from the same distribution. This is further proof that the efficacy of the proposed framework is not related to the size of the data.

CONCLUSIONS

In this paper, we presented a novel framework to adaptively remove abnormal curves from PCR amplification reactions. The method takes the raw input from a qdPCR run and processes it in three steps: background subtraction, late curve removal, and sigmoidal fitting. Moreover, a new feature called end slope (S_{end}) is developed in this study, which, along with sigmoidal parameters, is used in the adaptive mapping filter (AMF). The AMF is capable of removing nonspecific and low-efficient amplification curves, which are labeled as outliers. Melting curves of the outliers, previously removed, were compared with those of inliers using both wrong melting percentage (WMP) and melting peak distributions. Results show that nonspecific and low-efficient curves can be removed from amplification reaction by purely considering the sigmoidal trend. Further validation of the framework performance was conducted by assessing the classification accuracy and sensitivity of the ACA classifier on both inliers and outliers. This reinforces our hypothesis that removing abnormalities of amplification reaction in real-time PCR instruments would benefit data-driven multiplexing by removing undesired information.

In this research, we used data from qdPCR published in our previous work to demonstrate the effectiveness of the proposed framework, but its generality has not been tested in other settings. Future work will focus on evaluating this methodology on real-time data originating from various qPCR instruments, from different chemistries (such as isothermal amplification), and from point-of-care devices. Digital PCR allows us to generate amplification curves at low concentrations of samples, enabling the use of the developed framework. However, future work will focus on the application of this novel method in bulk reactions. Moreover, in the event of secondary amplification, the curve may show a second increasing phase with a large FFI and deviated shape from the sigmoid. However, as shown in Figure 2b, fitting step, the approximate shape of the distorted curve can still be depicted by the 5-parameter model, with still relatively small fitting error. After fitting, certain parameter values of the secondary amplification events will be different and distant from normal reactions and these events can be identified easily by the outlier detector. Regarding the presence of multiple targets in a single well, we expect to have a normal sigmoidal trend; therefore, the fitting error (MSE) will be low without affecting the AMF progress. However, the ACA classification of such an event may be challenging. In our previous work, we demonstrated that the presence of double targets can be resolved using the AMCA approach,[25] and other solutions such as FFI modulation by changing the probe concentration in TaqMan assay[19] may also help tackle this issue. Finally, the upcoming work will focus on introducing advanced machine learning techniques to enhance the classification efficacy of the ACA classifier and then on making this approach more reliable for use in clinical diagnostics.

In conclusion, this study reveals the interconnection between the kinetics of the amplification curve and the thermodynamics of the melting curves. For the first time, a framework is introduced, which is capable of removing abnormalities in kinetic and thermodynamic information by purely screening amplification curves.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.2c01883.

Curve-level performance comparison between 45-D and 5-D fitting parameter data; mean-squared error distribution of each class; and outlier distribution in clinical samples (PDF)

AUTHOR INFORMATION

Corresponding Author

Jesus Rodriguez-Manzano − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.*; orcid.org/0000-0002-2583-8366; Email: j.rodriguez-manzano@imperial.ac.uk

Authors

Luca Miglietta − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.; Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, U.K.*

Ke Xu − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.; Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, U.K.*; orcid.org/0000-0001-7376-2805

Priya Chhaya − *Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, U.K.*

Louis Kreitmann − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.*

Kerri Hill-Cawthorne − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.*

Frances Bolt − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.*

Alison Holmes − *Department of Infectious Disease, Faculty of Medicine, Imperial College London, London W12 0NN, U.K.*

Pantelis Georgiou − *Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.2c01883

Author Contributions

[§]L.M. and K.X. contributed equally to this work. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Imperial COVID-19 Research Fund (WDAI.G28059), the Department of Health and Social Care-funded Centre for Antimicrobial Optimisation (CAMO) at Imperial College London, and the Imperial College President's PhD Scholarships 2021 (K.X.). L.K. is also affiliated with the Research & Development of bioMérieux, S.A., and is supported by a scholarship from bioMérieux, S.A. The authors K.H.C., F.B., A.H., P.G., and J.R.M. are affiliated with the NIHR Health Protection Research Unit (HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at Imperial College London in partnership with the U.K. Health Security Agency (previously PHE) in collaboration with Imperial Healthcare Partners, the University of Cambridge, and the University of Warwick. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health and Social Care, or the U.K. Health Security Agency. A.H. is a National Institute for Health and Care Research (NIHR) Senior Investigator.

■ REFERENCES

(1) Whale, A. S.; Huggett, J. F.; Tzonev, S. *Biomol. Detect. and Quantif.* **2016**, *10*, 15−23.
(2) McDermott, G. P.; Do, D.; Litterst, C. M.; Maar, D.; Hindson, C. M.; Steenblock, E. R.; Legler, T. C.; Jouvenot, Y.; Marrs, S. H.; Bemis, A.; Shah, P.; Wong, J.; Wang, S.; Sally, D.; Javier, L.; Dinio, T.; Han, C.; Brackbill, T. P.; Hodges, S. P.; Ling, Y.; Klitgord, N.; Carman, G. J.; Berman, J. R.; Koehler, R. T.; Hiddessen, A. L.; Walse, P.; Bousse, L.; Tzonev, S.; Hefner, E.; Hindson, B. J.; Cauly, T. H.; Hamby, K.; Patel, V. P.; Regan, J. F.; Wyatt, P. W.; Karlin-Neumann, G. A.; Stumbo, D. P.; Lowe, A. J. *Anal. Chem.* **2013**, *85*, 11619−11627.
(3) Valasek, M. A.; Repa, J. J. *Adv. Physiol. Educ.* **2005**, *29*, 151−159.
(4) Higuchi, R.; Fockler, C.; Dollinger, G.; Watson, R. *Nat. Biotechnol.* **1993**, *11*, 1026−1030.
(5) Bustin, S. A.; Benes, V.; Garson, J. A.; Hellemans, J.; Huggett, J.; Kubista, M.; Mueller, R.; Nolan, T.; Pfaffl, M. W.; Shipley, G. L.; Vandesompele, J.; Wittwer, C. T. *Clin. Chem.* **2009**, *55*, 611−622.
(6) Gingeras, T. R.; Higuchi, R.; Kricka, L. J.; Lo, Y. D.; Wittwer, C. T. *Clin. Chem.* **2005**, *51*, 661−671.
(7) Bustin, S. A.; Benes, V.; Nolan, T.; Pfaffl, M. W. *J. Mol. Endocrinol.* **2005**, *34*, 597−601.
(8) Kaltenboeck, B.; Wang, C. *Adv. Clin. Chem.* **2005**, *40*, 219−259.
(9) Mackay, I. M.; Arden, K. E.; Nitsche, A. *Nucleic Acids Res.* **2002**, *30*, 1292−1305.
(10) Rodriguez-Manzano, J.; Moniri, A.; Malpartida-Cardenas, K.; Dronavalli, J.; Davies, F.; Holmes, A.; Georgiou, P. *Anal. Chem.* **2019**, *91*, 2013−2020.
(11) Elnifro, E. M.; Ashshi, A. M.; Cooper, R. J.; Klapper, P. E. *Clin. Microbiol. Rev.* **2000**, *13*, 559−570.
(12) Markoulatos, P.; Siafakas, N.; Moncany, M. *J. Clin. Lab. Anal.* **2002**, *16*, 47−51.
(13) Arya, M.; Shergill, I. S.; Williamson, M.; Gommersall, L.; Arya, N.; Patel, H. R. *Expert Rev. Mol. Diagn.* **2005**, *5*, 209−219.
(14) Souissi, A.; Ben Said, M.; Ben Ayed, I.; Elloumi, I.; Bouzid, A.; Mosrati, M. A.; Hasnaoui, M.; Belcadhi, M.; Idriss, N.; Kamoun, H.; Gharbi, N.; Gibriel, A. A.; Tlili, A.; Masmoudi, S. *J. Adv. Res.* **2021**, *31*, 13−24.
(15) Gibriel, A. A. Y. *Briefings Funct. Genomics* **2012**, *11*, 311−318.
(16) Moniri, A.; Rodriguez-Manzano, J.; Malpartida-Cardenas, K.; Yu, L.-S.; Didelot, X.; Holmes, A.; Georgiou, P. *Anal. Chem.* **2019**, *91*, 7426−7434.
(17) Otoo, J. A.; Schlappi, T. S. *Biosensors* **2022**, *12*, No. 124.
(18) Athamanolap, P.; Parekh, V.; Fraley, S. I.; Agarwal, V.; Shin, D. J.; Jacobs, M. A.; Wang, T.-H.; Yang, S. *PLoS One* **2014**, *9*, No. e109094.
(19) Jacky, L.; Yurk, D.; Alvarado, J.; Belitz, P.; Fathe, K.; MacDonald, C.; Fraser, S.; Rajagopal, A. *Anal. Chem.* **2021**, *93*, 4208−4216.
(20) Yao, J.; Luo, Y.; Zhang, Z.; Li, J.; Li, C.; Li, C.; Guo, Z.; Wang, L.; Zhang, W.; Zhao, H.; Zhou, L. *Biosens. Bioelectron.* **2022**, *199*, No. 113873.
(21) Burdukiewicz, M.; Spiess, A.-N.; Blagodatskikh, K. A.; Lehmann, W.; Schierack, P.; Rödiger, S. *Biomol. Detect. Quantif.* **2018**, *16*, 1−4.
(22) Moniri, A.; Miglietta, L.; Malpartida-Cardenas, K.; Pennisi, I.; Cacho-Soblechero, M.; Moser, N.; Holmes, A.; Georgiou, P.; Rodriguez-Manzano, J. *Anal. Chem.* **2020**, *92*, 13134−13143.
(23) Miglietta, L.; Chen, Y.; Luo, Z.; Xu, K.; Ding, N.; Peng, T.; Moniri, A.; Kreitmann, L.; Cacho-Soblechero, M.; Holmes, A.; Georgiou, P.; Rodriguez-Manzano, J.*Smart-Plexer: A Breakthrough Workflow for Hybrid Development of Multiplex PCR Assay*, 2022.
(24) Montgomery, J. L.; Sanford, L. N.; Wittwer, C. T. *Expert Rev. Mol. Diagn.* **2010**, *10*, 219−240.
(25) Moniri, A.; Miglietta, L.; Holmes, A.; Georgiou, P.; Rodriguez-Manzano, J. *Anal. Chem.* **2020**, *92*, 14181−14188.
(26) Miglietta, L.; Moniri, A.; Pennisi, I.; Malpartida-Cardenas, K.; Abbas, H.; Hill-Cawthorne, K.; Bolt, F.; Jauneikaite, E.; Davies, F.; Holmes, A.; Georgiou, P.; Rodriguez-Manzano, J. *Front. Mol. Biosci.* **2021**, *8*, No. 775299.
(27) Shen, H.; Rogelj, S.; Kieft, T. L. *Mol. Cell. Probes* **2006**, *20*, 147−153.
(28) Markey, A. L.; Mohr, S.; Day, P. J. R. *Methods* **2010**, *50*, 277−281.
(29) Spiess, A.-N.; Feig, C.; Ritz, C. *BMC Bioinf.* **2008**, *9*, 221.
(30) Sidstedt, M.; Romsos, E. L.; Hedell, R.; Ansell, R.; Steffen, C. R.; Vallone, P. M.; Rådström, P.; Hedman, J. *Anal. Chem.* **2017**, *89*, 1642−1649.
(31) Fluidigm. *User Guide Digital PCR Analysis (PN 68000100)*.
(32) Coleman, T. F.; Li, Y. *SIAM J. Optim.* **1996**, *6*, 418−445.
(33) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. In *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; KDD'96; AAAI Press: Portland, Oregon, 1996; pp 226−231.
(34) Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; Sander, J. *ACM SIGMOD Rec.* **2000**, *29*, 93−104.
(35) Schölkopf, B.; Platt, J.; Shawe-Taylor, J.; Smola, A.; Williamson, R. *Neural Comput.* **2001**, *13*, 1443−1471.
(36) Rousseeuw, P. J.; Driessen, K. *Technometrics* **1999**, *41*, 212−223.
(37) Liu, F. T.; Ting, K. M.; Zhou, Z.-H. In *Isolation Forest*, 2008 Eighth IEEE International Conference on Data Mining; IEEE: Pisa, Italy, 2008; pp 413−422.
(38) Lazarevic, A.; Kumar, V. In *Feature Bagging for Outlier Detection*, Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05; Association for Computing Machinery: New York, NY, USA, 2005; pp 157−166.
(39) Kriegel, H.-P.; Schubert, M.; Zimek, A. In *Angle-Based Outlier Detection in High-Dimensional Data*, Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining; KDD '08; Association for Computing Machinery: New York, NY, USA, 2008; pp 444−452.
(40) Sun, B.; Rodriguez-Manzano, J.; Selck, D. A.; Khorosheva, E.; Karymov, M. A.; Ismagilov, R. F. *Angew. Chem., Int. Ed.* **2014**, *53*, 8088−8092.
(41) Rousseeuw, P. J. *J. Comput. App. Math.* **1987**, *20*, 53−65.
(42) Davies, D. L.; Bouldin, D. W. *IEEE Trans. Pattern Anal. Mac. Intell.* **1979**, *PAMI-1*, 224−227.
(43) Caliński, T.; Harabasz, J. *Commun. Stat.* **1974**, *3*, 1−27.
(44) Rutledge, R. G. *Nucleic Acids Res.* **2004**, *32*, No. e178.
(45) Rutledge, R. G.; Stewart, D. *BMC Biotechnol.* **2008**, *8*, No. 47.
(46) Ruiz-Villalba, A.; van Pelt-Verkuil, E.; Gunst, Q. D.; Ruijter, J. M.; van den Hoff, M. J. *Biomol. Detect. Quantif.* **2017**, *14*, 7−18.
(47) Guescini, M.; Sisti, D.; Rocchi, M. B.; Stocchi, L.; Stocchi, V. *BMC Bioinf.* **2008**, *9*, No. 326.

https://doi.org/10.1021/acs.analchem.2c01883
Anal. Chem. 2022, 94, 14159−14168