UNIVERSITY OF CALIFORNIA SAN DIEGO

Advancing Rapid Infectious Disease Screening Using a Combined Experimental/Computational Approach

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Nanoengineering

by

Lennart Langouche

Committee in charge:

Professor Stephanie I. Fraley, Chair
Professor Jesse Vincent Jokerst, Co-Chair
Professor Yi Chen
Professor Todd Prentice Coleman
Professor Martin Hoenigl
Professor Shelley M. Lawrence
Professor Donald James Sirbuly

2021

The dissertation of Lennart Langouche is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To Nicole, Flor, Vincent and Kiran for believing in me and encouraging me to pursue my dreams, even if it means living thousands of miles away from them.

EPIGRAPH

There's plenty of room at the bottom.

—Richard Feynman

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Very Special Thanks to..

Sadik Esener, Yu-Hwa Lo and Stephanie Fraley for accepting me into their labs and believing in me. To Todd Coleman for his unrivaled excitement for science and great mentorship. To the many colleagues I got to know along the way, of which some have become true friends and without whom I would have never been able to bring this journey to a successful end. To my family and friends back in Belgium for always believing in me and supporting my dreams, even if it means living thousands of miles away from them.

Co-Authors and Publishers

Chapter 1, in part, was published in Bioinformatics on 23 December 2020. Lennart Langouche, April Aralar, Mridu Sinha, Shelley M Lawrence, Stephanie I Fraley, Todd P Coleman. The dissertation author was the primary investigator and author of this material.

Chapter 2, in part, was published in Bioinformatics on 23 December 2020. Lennart Langouche, April Aralar, Mridu Sinha, Shelley M Lawrence, Stephanie I Fraley, Todd P Coleman. The dissertation author was the primary investigator and author of this material. Authors would like to extend thanks to thank Zachary Dwight for his assistance in obtaining synthetic melt curves using uMelt software.

Chapter 3, in part, has been submitted for publication as it may appear in Bioinformatics Advances 2021. Lennart Langouche, Augustine C. Obirieze, Mridu Sinha, Hannah Mack, William Leineweber, April Aralar, David T. Pride, Todd P. Coleman, Stephanie I. Fraley. The dissertation author was the primary investigator and author of this material.

VITA

2011	B. S. in Engineering, K.U. Leuven, Belgium
2013	M. S. in Nanoscience and Nanotechnology *magna cum laude*, K.U. Leuven, Belgium
2013	M. S. in Nanoscience and Nanotechnology *magna cum laude*, Chalmers University, Sweden
2021	Ph. D. in Nanoengineering, University of California San Diego

PUBLICATIONS

A Dankert, L Langouche, MV Kamalakar, SP Dash, “High-performance molybdenum disulfide field-effect transistors with spin tunnel contacts”, *ACS nano*, 8 (1), 476-482, 2014.

W Cai, E Wang, PW Chen, YH Tsai, L Langouche, YH Lo, “A microfluidic design for desalination and selective removal and addition of components in biosamples”, *Biomicrofluidics*, 13 (2), 024109, 2019

L Langouche, A Aralar, M Sinha, SM Lawrence, SI Fraley, TP Coleman, “Data-driven noise modeling of digital DNA melting analysis enables prediction of sequence discriminating power”, *Bioinformatics*, 2020

ABSTRACT OF THE DISSERTATION

Advancing Rapid Infectious Disease Screening Using a Combined Experimental/Computational Approach

by

Lennart Langouche

Doctor of Philosophy in Nanoengineering

University of California San Diego, 2021

Professor Stephanie I. Fraley, Chair
Professor Jesse Vincent Jokerst, Co-Chair

Outbreaks of infectious diseases are rising around the world. In addition to outbreaks, several known infectious diseases have a significant impact on mortality in the US and around the world. In this doctoral work we use a combined experimental/computational approach to study the performance and limitations of digital High Resolution Melt (dHRM), an infectious disease screening platform that was previously established in the Fraley lab. This allows comparisons with other microbiological technologies and helps shine some light on its potential use-cases. First, we developed a computational framework for estimating the resolving power of dHRM technology for defined sequence profiling tasks. By deriving noise models from experimentally generated dHRM datasets and applying these to in silico predicted melt curves, we enable the production of synthetic dHRM datasets that faithfully recapitulate real-world variations arising from sample and machine variables. Second, we present an advancement in universal microbial high resolution melting (HRM) analysis that is capable of accomplishing both known genotype identification and novel genotype detection. Specifically, this novel surveillance functionality is achieved through probabilistic modeling of sequence-defined HRM curves, which is uniquely enabled by the large-scale melt curve datasets generated using our high-throughput digital HRM platform. Our hope is that in the future, the dHRM platform can translate into a near-point of care, cost-effective tool for infectious disease screening.

Chapter 1

Introduction

1.1 Infectious Disease: Overview

Infectious disease outbreaks are rising around the world [1]. Outbreaks occur when the number of disease cases increases above what would statistically be expected in a defined geographical area, season or community. Figure 1, adapted from [1], shows that the number of outbreaks and variety of causal diseases exhibit a significant increase since 1980, as do the number of outbreaks and variety of causal diseases for each subcategory of pathogen taxonomy (viruses, bacteria, fungi, parasites or protozoa), pathogen transmission mode (vector transmitted or non-vector transmitted) and host type (human specific or zoonotic) [1].

Besides outbreaks, infectious diseases have a significant impact on mortality in the US and around the world. There are many different types of infections, including but not limited to the bloodstream, urinary tract, gastrointestinal tract, respiratory tract and central nervous system. One example we will briefly discuss is sepsis. Sepsis occurs when an infection is accompanied by organ dysfunction from a dysregulated host response to that infection [2]. The CDC reports that each year, at least 1.7 million adults in America develop sepsis, nearly 270,000 Americans die as a result of sepsis and 1 in 3 patients who dies in a hospital has sepsis [3, 4]. Sepsis survivors

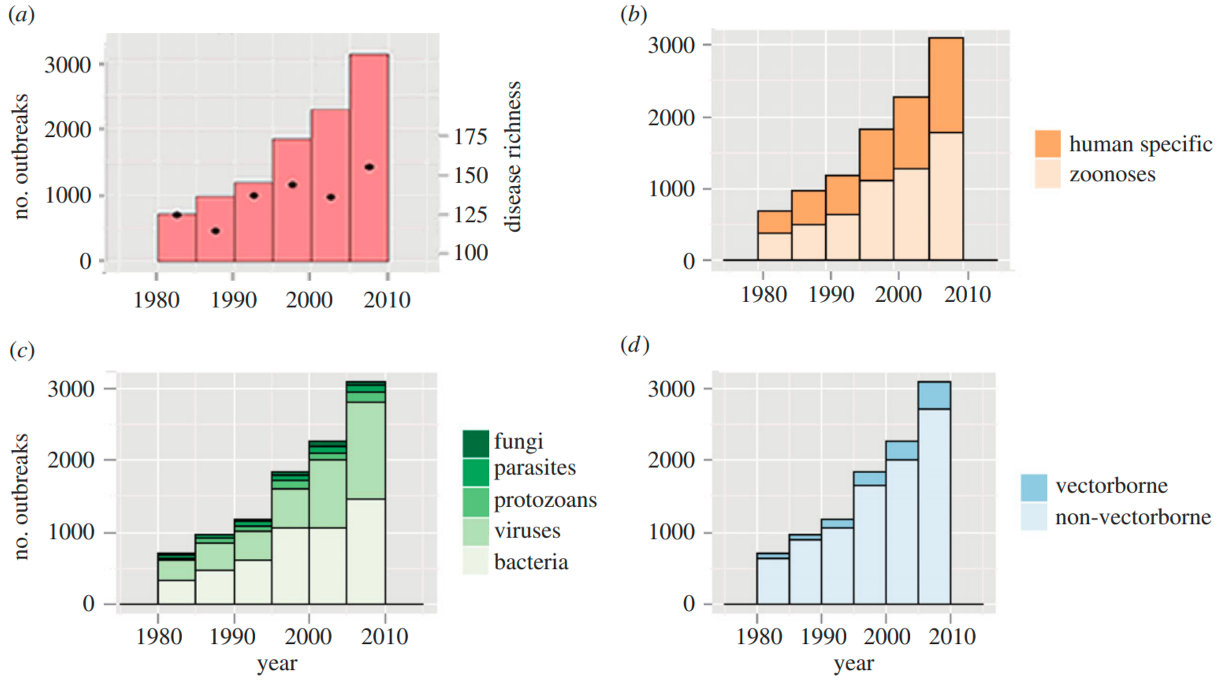


Figure 1.1: Global number of human infectious disease outbreaks and richness of causal diseases 1980 – 2010. Outbreak records are plotted with respect to (a) total global outbreaks (left axis, bars) and total number of diseases causing outbreaks in each year (right axis, dots), (b) host type, (c) pathogen taxonomy and (d) transmission mode. Figure and caption adapted from [1].

often suffer long-term physical, psychological, and cognitive disabilities [5]. Despite the high mortality rate, there is no approved diagnostic test, and the diagnosis requires clinical judgment based on evidence of both infection and organ dysfunction [5].

Generally, the microbiology laboratory plays a critical role in infectious disease detection, as the clinician seeks the aid of a clinical microbiology laboratory to help with diagnosis [6].

1.2 Infectious Disease Diagnostics

The microbiology laboratory tries to answer three basic questions, i.e., whether the patient is infected, if so, with which pathogen(s) and third, what will treat it [7]. Often, broad empirical antimicrobials are given to a patient during the window of time it takes for specimen collection, pathogen identification and antimicrobial susceptibility testing. Because of this, many patients are overtreated with antimicrobials that afterwards turn out to be inappropriate, which in addition has led to the phenomenon of antimicrobial resistance. Molecular diagnostic technologies have enabled the microbiology lab to provide answers to these questions more accurately and faster than ever before. Rapid diagnostic technologies are able to decrease the window of time needed, which in turn enables providing faster, effective and targeted antimicrobial treatment while also decreasing the use of unnecessary empirical treatment.

Microorganisms such as bacteria, viruses, fungi, and parasites can now promptly be identified using molecular technologies, to diagnose causative pathogens in infections of the bloodstream, respiratory tract, urinary tract, gastrointestinal tract, and central nervous system [8]. In parallel, new gene-based resistance detection methods are rapidly being developed to guide antimicrobial use. Novel rapid diagnostic technologies include nucleic acid-based diagnostics (e.g. single-target PCR testing and multiplex PCR panels) [9, 10, 11, 12, 13], microarray panels [14], peptide nucleic acid fluorescent in situ hybridization (FISH) technologies [15], magnetic resonance-based testing [16], matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) [17], and next-generation sequencing (NGS) [18]. These technologies span the range from single-target pathogen-specific or gene-specific antimicrobial resistance testing to syndromic panels [19] containing many common pathogens causing a disease process, to unbiased sequencing with the ability to detect unsuspected or novel pathogens [20].

1.3 Universal Digital High-Resolution Melt (U-dHRM)

HRM rapidly analyzes a DNA sequence by measuring how the bonds between double-stranded DNA break in response to heating. The readout of HRM analysis is based on the fluorescence of a generic DNA intercalating dye, which binds to double-stranded DNA and fluoresces, but loses fluorescence as the DNA unwinds to become single-stranded (Fig. 1.2A). This heating and unwinding process, which takes about five to ten minutes, produces a melt curve that can be plotted as a fluorescence versus temperature graph (Fig. 1.2B). Melt curves are sensitive to the content and order of nucleotides, as well as heating rate during the melt process [21, 22, 23]. This allows melt curves to serve as unique signatures for DNA sequences. As such, machine learning classification algorithms can be used to identify DNA sequences based on their melt curve signatures [24, 25, 26]. For example, we previously generated a database of melt curves for a variety of bacterial organisms using their hypervariable 16S rRNA gene sequence. Machine learning enabled us to identify the organisms by their melt curve with 99% accuracy [25]. We recently developed a high-throughput digital HRM (dHRM) analysis system [23, 22] with robust temperature control [22] that uniquely enables the reliable generation of thousands of melt curves from a DNA sequence in a single experiment (Fig. 1.2C). This advance has a 200-fold increase in throughput compared to traditional well-plate formats, enabling rich data-driven analyses. Using universal primers on a dHRM system (Universal or U-dHRM [27]) enables the detection of a large number of pathogens using a single primer set.

Given its simplicity and speed, dHRM is a promising technique for diagnostic applications. The main limitation of dHRM is the run-to-run and well-to-well variation in melt curves due to sample and machine variables [22].

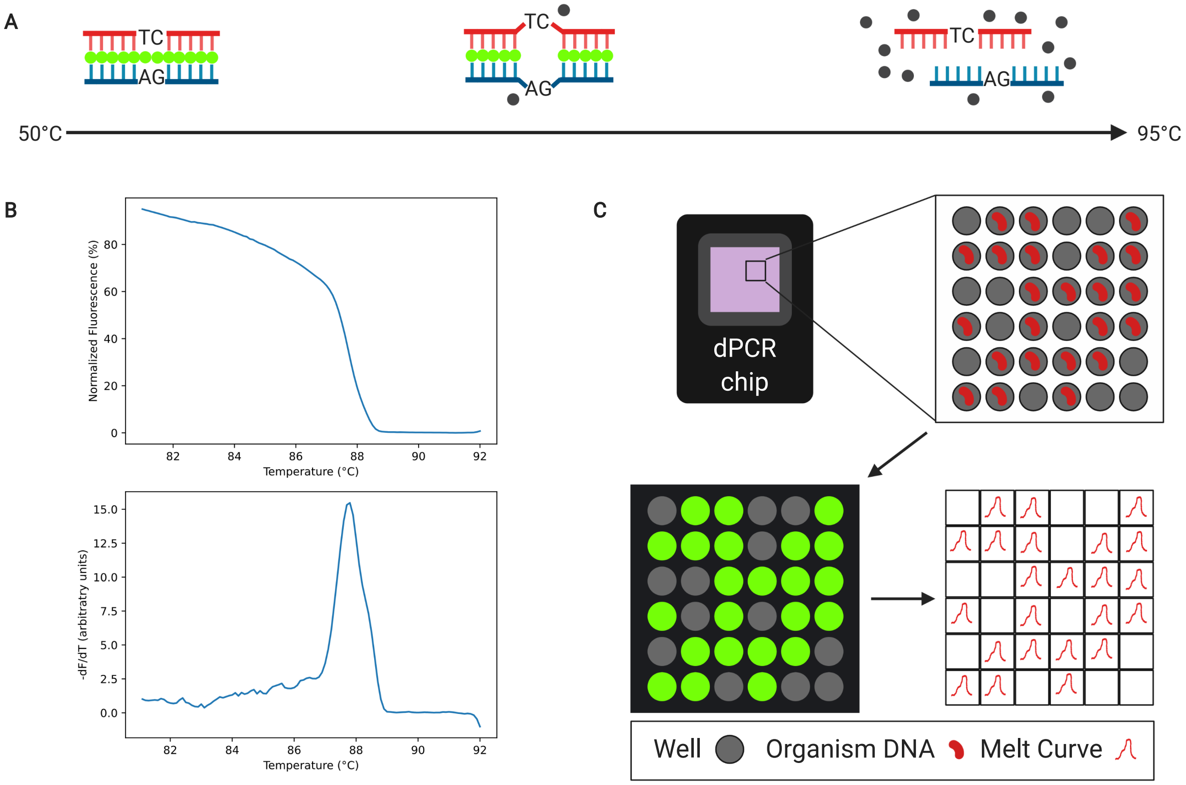


Figure 1.2: dHRM Overview. **A** Intercalating dye (green), which binds to double-stranded DNA (top and bottom strands) and fluoresces. The dye loses fluorescence as the temperature increases and DNA unwinds to become single-stranded. **B** Loss-of-fluorescence curve (top) and its negative derivative (bottom). **C** Diagram of high-throughput digital HRM workflow to generate thousands of melt curves simultaneously to enable data-driven analysis. Top left: digital PCR chip with 20,000 picoliter-sized wells.

1.4 Thesis

In this doctoral work we use a combined experimental/computational approach to study dHRM's performance and potential limitations. This allows comparisons with other microbiological technologies and helps shine some light on its potential use-cases. In Chapter 2 we investigate the resolving power of dHRM and discuss the implications. In Chapter 3 we discuss novelty detection and investigate whether dHRM can be used to detect unsuspected or novel pathogens, similar to unbiased NGS.

1.5 Acknowledgments

Disclaimer: Figure 1 is adapted from [1]. It is not property of the dissertation author or his co-authors. Chapter 1, in part, was published in Bioinformatics on 23 December 2020. Lennart Langouche, April Aralar, Mridu Sinha, Shelley M Lawrence, Stephanie I Fraley, Todd P Coleman. The dissertation author was the primary investigator and author of this material. The publication can be accessed online at https://doi.org/10.1093/bioinformatics/btaa1053. Fig. 1.2A and Fig. 1.2C were created with BioRender.com.

Chapter 2

Data-driven Noise Modeling of Digital DNA Melting Analysis Enables Prediction of Sequence Discriminating Power

2.1 Abstract

Motivation: The need to rapidly screen complex samples for a wide range of nucleic acid targets, like infectious diseases, remains unmet. Digital High-Resolution Melt (dHRM) is an emerging technology with potential to meet this need by accomplishing broad-based, rapid nucleic acid sequence identification. Here, we set out to develop a computational framework for estimating the resolving power of dHRM technology for defined sequence profiling tasks. By deriving noise models from experimentally generated dHRM datasets and applying these to in silico predicted melt curves, we enable the production of synthetic dHRM datasets that faithfully recapitulate real-world variations arising from sample and machine variables. We then use these datasets to identify the most challenging melt curve classification tasks likely to arise for a given application and test the performance of benchmark classifiers.

Results: This toolbox enables the in silico design and testing of broad-based dHRM screening assays and the selection of optimal classifiers. For an example application of screening common human bacterial pathogens, we show that human pathogens having the most similar sequences and melt curves are still reliably identifiable in the presence of experimental noise. Further, we find that ensemble methods outperform whole series classifiers for this task and are in some cases able to resolve melt curves with single-nucleotide resolution.

Availability: Data and code available on https://github.com/lenlan/dHRM-noise-modeling

2.2 Introduction

HRM rapidly analyzes a DNA sequence by measuring how the bonds between double-stranded DNA break in response to heating. The readout of HRM analysis is based on the fluorescence of a generic DNA intercalating dye, which binds to double-stranded DNA and fluoresces, but loses fluorescence as the DNA unwinds to become single-stranded (Fig. 2.1A). This heating and unwinding process, which takes about five to ten minutes, produces a melt curve that can be plotted as a fluorescence versus temperature graph (Fig. 2.1B). Melt curves are sensitive to the content and order of nucleotides, as well as heating rate during the melt process [21, 22, 23]. This allows melt curves to serve as unique signatures for DNA sequences. As such, machine learning classification algorithms can be used to identify DNA sequences based on their melt curve signatures [24, 25, 26]. For example, we previously generated a database of melt curves for a variety of bacterial organisms using their hypervariable 16S rRNA gene sequence. Machine learning enabled us to identify the organisms by their melt curve with 99% accuracy [25]. We recently developed a high-throughput digital HRM (dHRM) analysis system [23, 22] with robust temperature control [22] that uniquely enables the reliable generation of thousands of melt curves from a DNA sequence in a single experiment (Fig. 2.1C). This advance has a 200-fold increase in throughput compared to traditional well-plate formats, enabling rich data-driven analyses. Using universal primers on a dHRM system (Universal or U-dHRM [27]) enables the detection of a large number of pathogens using a single primer set.

Given its simplicity and speed, dHRM is a promising technique for diagnostic applications. The main limitation of dHRM is the run-to-run and well-to-well variation in melt curves due to sample and machine variables [22]. This has led us to ask: what is the resolving power of dHRM? This will be affected by two main factors: the noise inherent to dHRM and the classification approach used to discern the melt curves. This work takes a combined experimental/computational approach to predict the resolving power. We hope to provide insight into further scalability and enable comparisons with other technologies that are emerging for use in infectious disease diagnostics, such as next-generation sequencing (NGS).

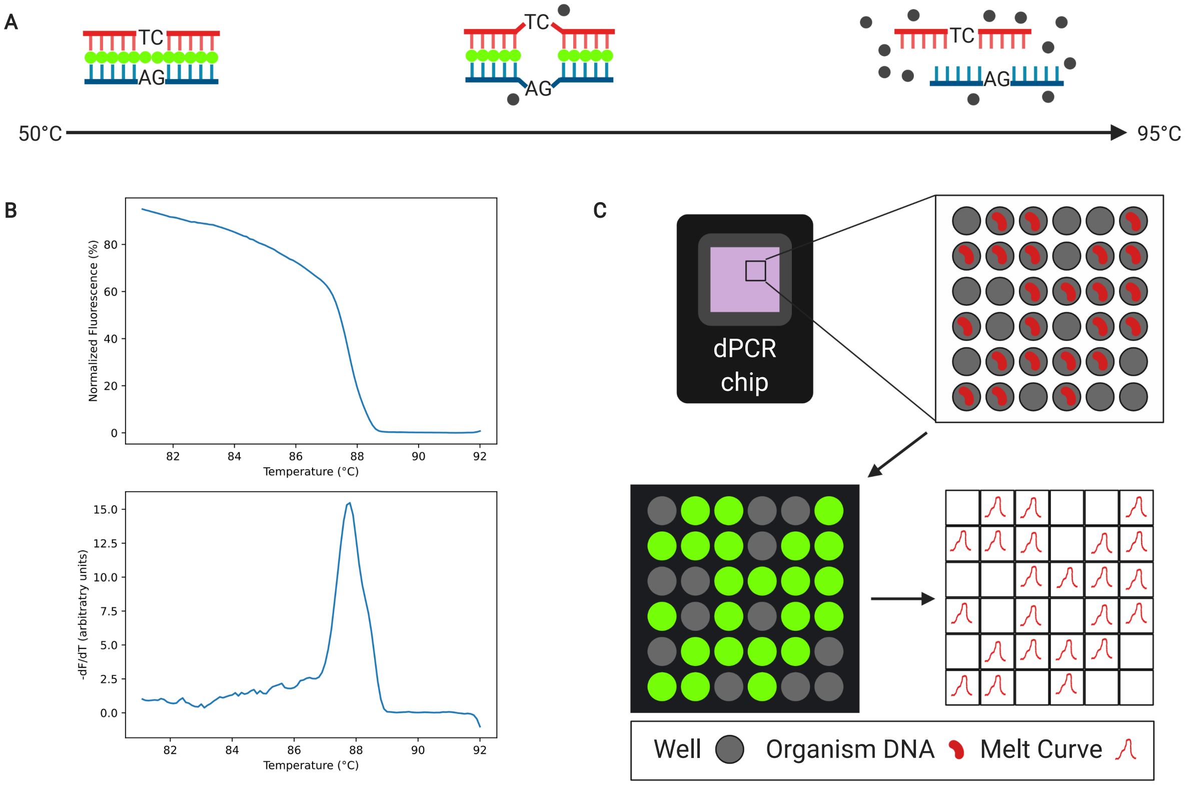


Figure 2.1: dHRM Overview. **A** Intercalating dye (green), which binds to double-stranded DNA (top and bottom strands) and fluoresces. The dye loses fluorescence as the temperature increases and DNA unwinds to become single-stranded. **B** Loss-of-fluorescence curve (top) and its negative derivative (bottom). **C** Diagram of high-throughput digital HRM workflow to generate thousands of melt curves simultaneously to enable data-driven analysis. Top left: digital PCR chip with 20,000 picoliter-sized wells.

2.3 Materials and Methods

2.3.1 Melt Curve Generation

Digital High Resolution Melt data was collected as detailed previously [23]. It consists of melt curves from the 16S rRNA gene (regions V1 to V6) of ten different bacterial organisms, which are listed along with their amplicon length in Table 2.1. The melt curves have been cropped to 160 datapoints, a datapoint every 0.1°C with a temperature range between 75.6°C and 91.5°C (Fig. 2.2). The melt curves have been smoothed (using the Matlab function imgaussfilt with σ = 3) and their derivative is taken with respect to temperature to obtain -dF/dT. An average of 1828 melt curves were generated for each organism (the lowest is 843). Fig. 2.2 provides a snapshot of the data collected from the ten species described in Table 2.1, where each subplot shows all melt curves originating from a single chip and single organism with the mean melt curve superimposed. Fig. 2.3 shows the residuals from the mean and their variance for the same dataset.

Table 2.1: Overview of experimental data set.

Nr.	Species	Amplicon Length	Melt Curves (Nr.)	Mean Peak (°C)	Std. Peak (°C)	Median DTW
1	*Citrobacter koseri*	969	2052	89.06	0.42	2.31
2	*Enterococcus faecium*	978	1137	88.53	0.10	1.68
3	*Escherichia coli*	981	843	88.49	0.13	2.60
4	*Haemophilus influenzae*	967	2265	88.29	0.10	2.14
5	*Listeria monocytogenes*	994	2244	87.79	0.08	2.23
6	*Staphylococcus aureus (MSSA)*	981	2028	87.43	0.19	1.51
7	*Streptococcus gallolyticus*	978	2006	88.43	0.78	3.83
8	*Streptococcus 'group B' (GBS)*	978	2249	88.90	0.10	2.37
9	*Streptococcus pneumoniae*	978	1255	88.71	0.15	1.62
10	*Streptococcus sanguinis*	972	2202	88.45	0.09	3.00
	Average	977	1828	88.41	0.21	2.33

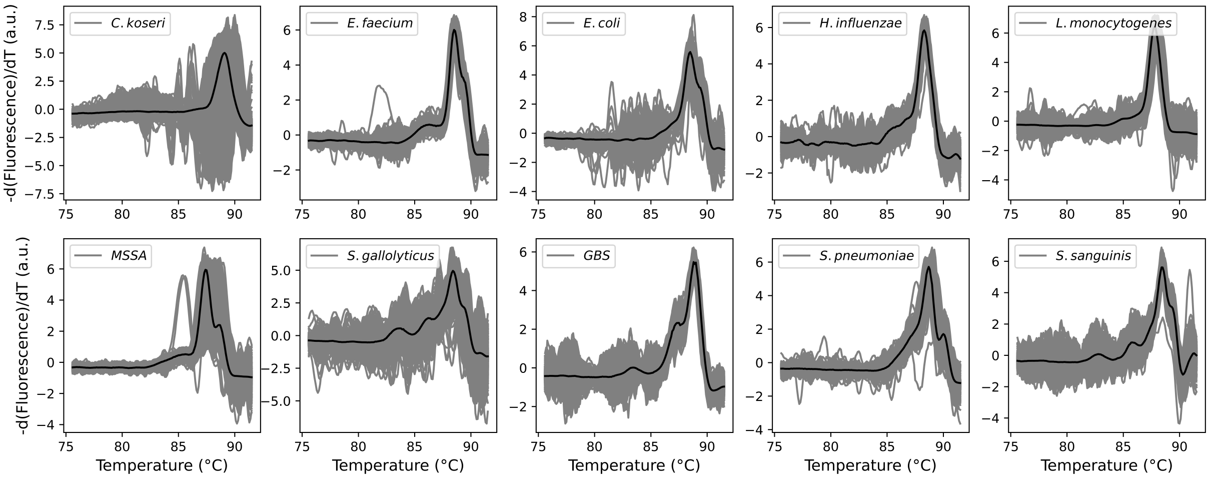


Figure 2.2: dHRM Melt curves with superimposed mean (black) for ten bacterial pathogenic organisms.

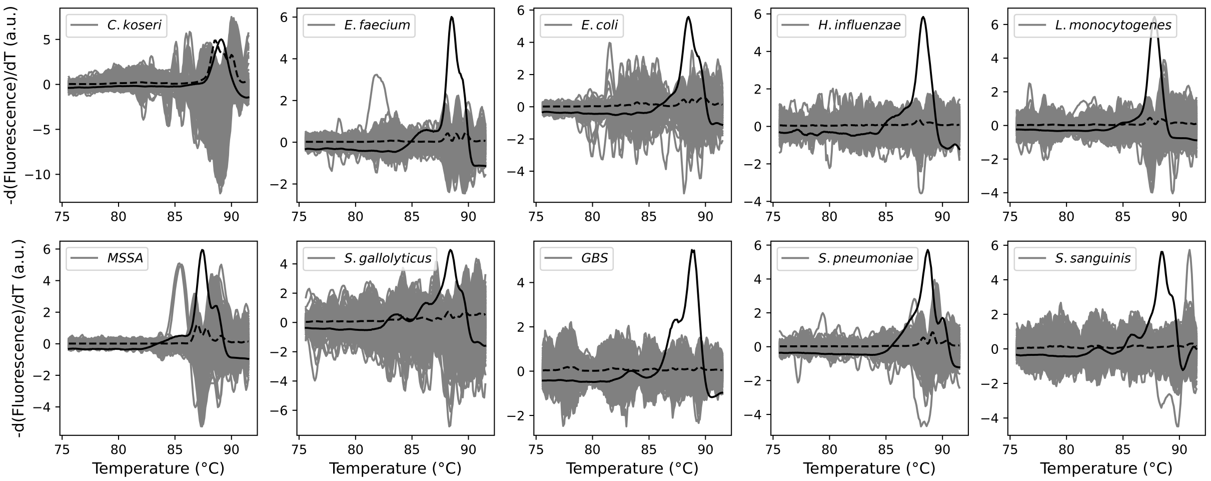


Figure 2.3: dHRM Melt curve residuals with superimposed mean (black). The dotted line at any temperature T is the variance of the melt curve at that T, calculated over all the wells associated with that species.

2.3.2 Characterization of Melt Curve Noise

Due to well-to-well and run-to-run variation [22], each chip has a distinct noise envelope as shown in Fig. 2.2. A shift or distortion along the temperature axis can be seen in some chips, here most clearly shown for MSSA (Fig. 2.2). This is an indicator that Dynamic Time Warping (DTW) might be successful at classifying these curves. DTW is an elastic distance measure that was introduced initially to deal with temporal distortions in the context of speech recognition [28], and has been used at least once for classifying HRM curves, which can exhibit a similar temperature distortion [29]. We used DTW as an investigative tool to obtain more insight into well-to-well variation in melt curve shapes. Figure 2.4. shows an example of how the DTW distance between melt curves is calculated. Table 2.1 shows the median DTW distance per chip (this is the median pairwise DTW distance between all curves from that chip). This can be interpreted as the amount of well-to-well variance in shape. We employ DTW without a window constraint here, which means the curves can be freely warped along the temperature axis. Focusing on MSSA again as an example, it can be observed that despite a seemingly broad noise envelope (Fig. 2.2), it has the second lowest median DTW or well-to-well variance in shape.

2.3.3 U-dHRM Classification

We have chosen to compare four classification methods on our data in this work. This selection is informed by a recent review comparing methods for time series classification (TSC) on the ‘The UCR Time Series Classification Archive’ [30, 31]. The authors recommend 1-nearest neighbor with Euclidean distance (1-NN ED) as a starting point on any new dataset, as this is generally a low benchmark that is easily beaten by other benchmark classifiers. Rotation Forest (RotF), Random Forest (RandF) and 1-nearest neighbor with Dynamic Time Warping (1-NN DTW) with a warping window set through cross validation [32] make up the review’s top three benchmark classifiers whereas 1-NN ED comes in sixth place [30].

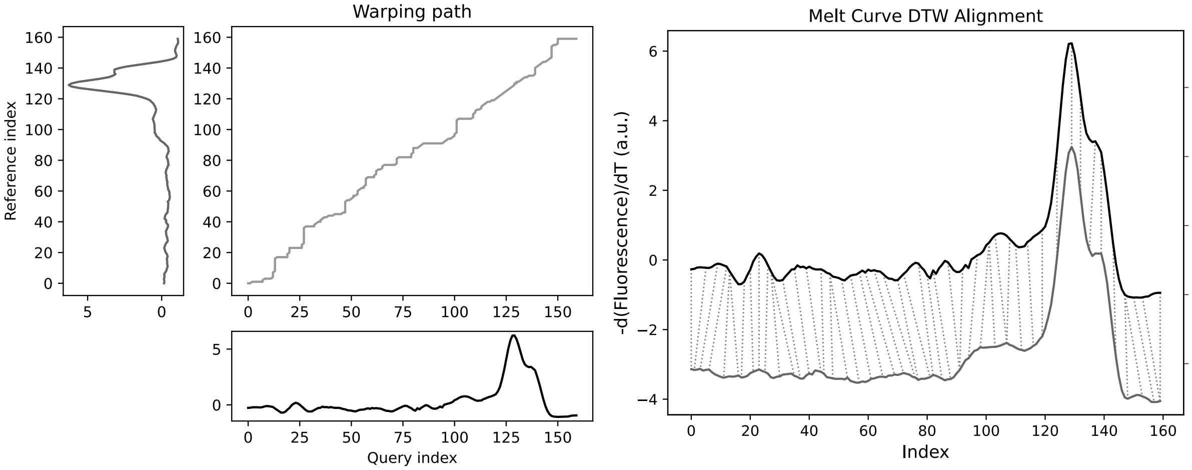


Figure 2.4: Example of dHRM melt curve DTW alignment. Left: warping path to align two melt curves. Right: dotted lines show which points get aligned under DTW (no window constraint). DTW forces end points to align, which can be seen in the right figure. ψ-DTW can be used to relax that constraint.

1-NN ED and 1-NN DTW are straightforward ‘whole series classifier’ algorithms, which try to find the distance between two time series using either inelastic (Euclidean distance) or elastic (DTW) measures. The other two methods are both ensemble classifiers, which have proved popular in recent TSC research and are highly competitive on general classification problems [30]. Random forest, introduced by Breiman in 2001 [33], consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest produces a class prediction and the class with the most votes becomes the model’s prediction. Node splitting in a random forest model is based on a random subset of features for each tree.

Rotation Forest, introduced by Rodríguez et al. in 2006 [34], is another ensemble classifier based on feature extraction. Here, Principal Component Analysis (PCA) is applied to each subset of features. All principal components are retained in order to preserve the variability information in the data (which is equivalent to a rotation of each subset of futures). The idea of the rotation approach is to encourage simultaneously individual accuracy and diversity within the ensemble [34].

We used Python’s scikit-learn [35] for RandF (50 trees) and the nearest neighbor classifiers, with the DTW distance measure from dtaidistance [36] which includes psi-relaxation [37]. We used cross validation to determine the optimal window (w), relaxation parameter psi (ψ) and number of neighbors (k). For RotF (50 trees), we used the implementation from [38].

2.3.4 U-dHRM Resolving Power

It has been shown that HRM is able to distinguish single-nucleotide polymorphisms (SNPs) under certain conditions [39, 40]. Two factors play a role here: amplicon size and type of mutation. Generally, SNPs are easier to differentiate in short amplicons as the melting temperature differences among genotypes increase as the amplicon size decreases [39]. The second factor that determines whether a SNP can be differentiated is the type of mutation, e.g. C/T, C/A, G/A or G/T substitutions are generally easier to differentiate than C/G or T/A substitutions [39], because %GC-content has a strong effect on melt temperature.

In order to differentiate pathogenic species, we select a specific barcoding region and desired amplicon length. Within the context of using clinical samples, we target longer amplicons (around 1000bp) to overcome the challenges of high background and environmental contamination relative to pathogen level (3). In this work, we have used the primers V1F: 5'-GYGGCGNACGGGTGAGTAA-3' and V6R: 5'-AGCTGACGACANCCATGCA-3' corresponding to amplicons including barcoding regions V1 to V6. The usage of such long amplicons typically results in having multiple sequence variations differentiating two species instead of a single SNP. We selected 58 clinically relevant bacterial pathogens, including category A and B biothreat agents and their surrogates from [41]. We adapted code from the primerTree package [42] to automate primer-BLAST [43] searching and return all matching amplicons given a specified primer pair and a list of species. We then used uMelt software [44] which allows prediction of high-resolution melting curves and dynamic melting profiles of PCR products.

uMelt

uMelt uses existing models of DNA melting that utilize nearest neighbor thermodynamics and recursive calculations using statistical mechanics [45, 46, 47, 48] to do fluorescent melting analysis of PCR products in a web application [44]. Instead of treating a DNA double helix as a string of separate interactions between base pairs, the nearest-neighbor model treats a DNA helix as a string of interactions between ’neighboring’ base pairs [49]. It means that the interaction between bases on different strands depend to some extent on the neighboring bases. So, for example, have a look at the DNA shown below.

5’ G-C-T-T-C-A 3’
3’ C-G-A-A-G-T 5’

The free energy of forming this DNA from the individual strands, ΔG, is represented (at 37 °C) as

$$\Delta G^{\circ}37(predicted) = \Delta G^{\circ}37(G/C\ initiation) + \Delta G^{\circ}37(GC/CG) + \Delta G^{\circ}37(CT/GA)$$
$$+ \Delta G^{\circ}37(TT/AA) + \Delta G^{\circ}37(TC/AG) + \Delta G^{\circ}37(CA/GT) + \Delta G^{\circ}37(A/T\ initiation).$$

First you have the G/C initiation term, representing the free energy of the first base pair, GC, in the absence of a nearest neighbor. The second term includes both the free energy of formation of the second base pair, CG, and stacking interaction between this base pair and the previous GC base pair. The remaining terms are constructed in a similar manner. In general, the free energy of forming a nucleic acid duplex is the sum of the initiation terms and the nearest neighbor interaction terms. Each ΔG° term has enthalpic, ΔH°, and entropic, ΔS°, parameters. Values of ΔH° and ΔS° have been determined for the ten possible pairs of interactions and can be found in thermodynamic tables.

We tried both low- and high-resolution settings of 1.0°C and 0.1°C to assess which melt curves would be most comparable to the experimentally obtained curves. Our experimentally obtained data is collected at a resolution of 0.1°C, so this was our first choice, but as shown in Figure 2.5, these curves have much narrower peaks than our experimentally obtained data. We decided to start with the 1.0°C resolution, and then smooth (Savitzky–Golay filter, window = 25, polynomial order = 2) and interpolate these curves to a 0.1°C resolution. This way, we obtained synthetic melt curves with shapes similar to our experimental data (Fig. 2.5). Since uMelt has its limitations, e.g. the algorithm does not account for the thermodynamics of dye binding and current libraries do not account for the rapid melting rates [44], the obtained melt curves are not exactly the same (neither in shape nor position) as our experimentally obtained data. However, for the purposes of this model, this is not necessary, as we are only interested in obtaining melt curves with realistic shapes and will not be comparing synthetic data with experimentally obtained data.

We calculated the Euclidean distance matrix for these 58 synthetic melts and selected the top 15 pairs with the smallest Euclidean distance between them (Table 2.2), as these will be the most difficult to differentiate. To create a fair and meaningful classification challenge, we applied the noise from our experimental dHRM data to these synthetic uMelts. We did this by calculating the residuals to the mean for each of the ten chips (Fig. 2.3) and applying those residuals to the synthetic uMelts. The residuals are shifted so that the noise at the peak location of the experimental data aligns with the peak of the uMelts. The residuals are also scaled by the ratio of the peak heights from the experimental data and the uMelts.

Fig. 2.6 outlines the complete workflow. The collection of residuals from one organism is randomly split in half and each half is applied to one uMelt from a pair. The resulting three sets are randomly split into test/training sets with a 2/3-1/3 split. This random splitting in half and random train/test splitting is both done three times to ensure consistency. The result is nine

Table 2.2: Generated uMelt pairs and similarity measures.

Pair	Organism 1	Organism 2	Euclidean distance	Sequence similarity (%)	Nucleotide mismatches
1	*Yersinia pestis*	*Yersinia pseudotuberculosis*	1.10	99.79	2
2	*Bacillus anthracis*	*Bacillus cereus*	1.11	99.69	3
3	*Proteus vulgaris*	*Pseudomonas aeruginosa*	1.55	87.62	120
4	*Streptococcus sanguinis*	*Yersinia enterocolitica*	1.66	84.74	148
5	*Proteus mirabilis*	*Proteus vulgaris*	1.76	98.56	14
6	*Bordetella parapertussis*	*Bordetella pertussis*	1.86	99.90	1
7	*Staphylococcus epidermidis*	*Staphylococcus lugdunensis*	1.93	98.27	17
8	*Staphylococcus lugdunensis*	*Staphylococcus saprophyticus*	2.00	98.37	16
9	*Mycobacterium gordonae*	*Mycobacterium kansasii*	2.03	98.34	16
10	*Yersinia enterocolitica*	*Yersinia pestis*	2.16	96.70	32
11	*Micrococcus luteus*	*Mycobacterium fortuitum*	2.25	90.87	87
12	*Staphylococcus aureus*	*Staphylococcus epidermidis*	2.32	98.57	14
13	*Yersinia enterocolitica*	*Yersinia pseudotuberculosis*	2.52	96.49	34
14	*Proteus mirabilis*	*Pseudomonas aeruginosa*	2.71	87.82	118
15	*Acinetobacter calcoaceticus*	*Aerococcus viridans*	2.82	81.31	183

train/test splits per pair and noise model. This is implemented for all 15 pairs and 10 noise models. Fig. 2.7 gives some examples of pairs and noise models. The classification task at hand consists of distinguishing the light from the dark gray curves.

We chose to apply one noise model at a time (all residuals coming from the same chip), as we found that mixing noise models can enable the classifiers to learn the noise model, rather than the actual underlying melt curves (data not shown). Finally, we also investigated to which extent the choice of noise model affects the classification result.

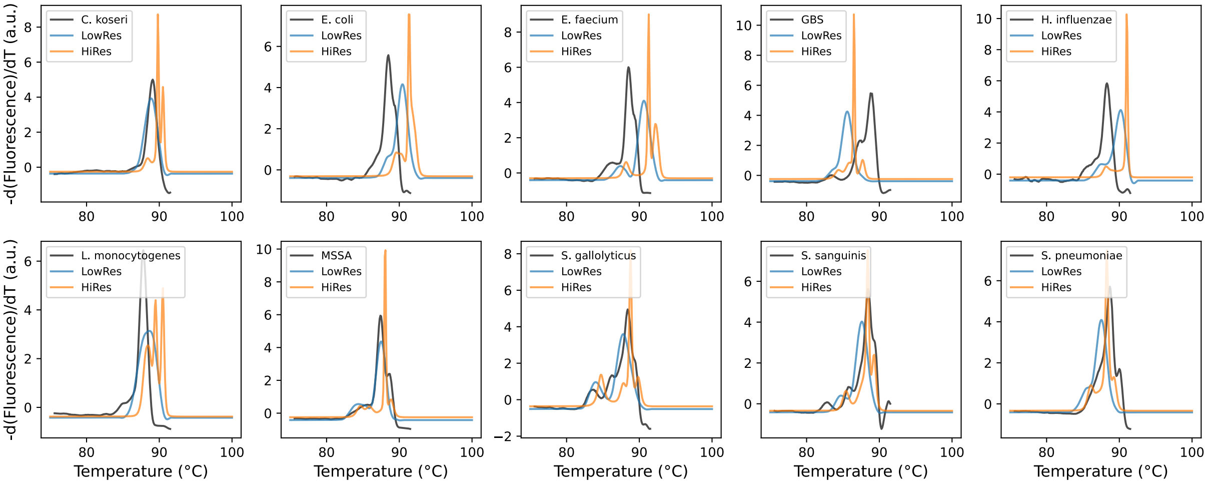


Figure 2.5: Comparison of experimental dHRM mean melt curves with synthetic uMelts (high resolution and smoothed low resolution). The smoothed low-resolution curves are more similar in shape to the experimental data than the high-resolution ones.

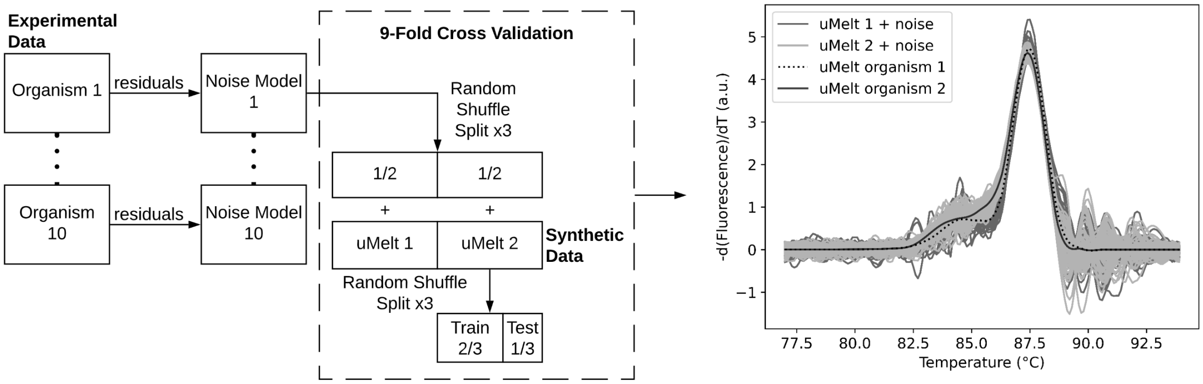


Figure 2.6: Workflow to obtain classification challenge to study dHRM resolving power.

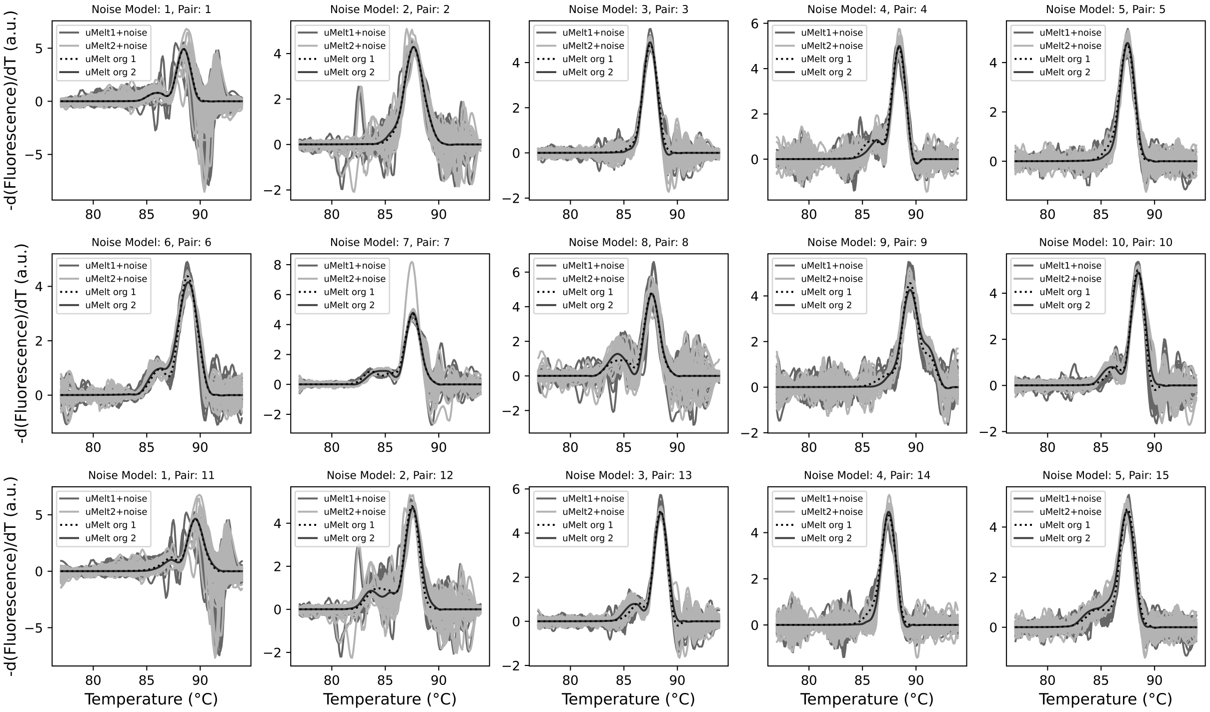


Figure 2.7: Examples of dHRM noise applied to pairs of synthetic uMelts. The classification challenge consists of differentiating the light from the dark gray curves.

2.4 Results

2.4.1 U-dHRM Classification

A comparison of the four classification methods applied to the melt curves from the top-10 sepsis causing pathogens (which account for at least 63% of cases) [50] listed in Table 2.1 is shown in Fig. 2.8 (left). All four classifications methods perform extremely well (accuracy > 99.5%) on this dataset and variation was limited between the five train-test splits. For the nearest neighbor classifiers, parameters were chosen after cross validation (Fig. 2.8 right and figures 2.9 and 2.10).

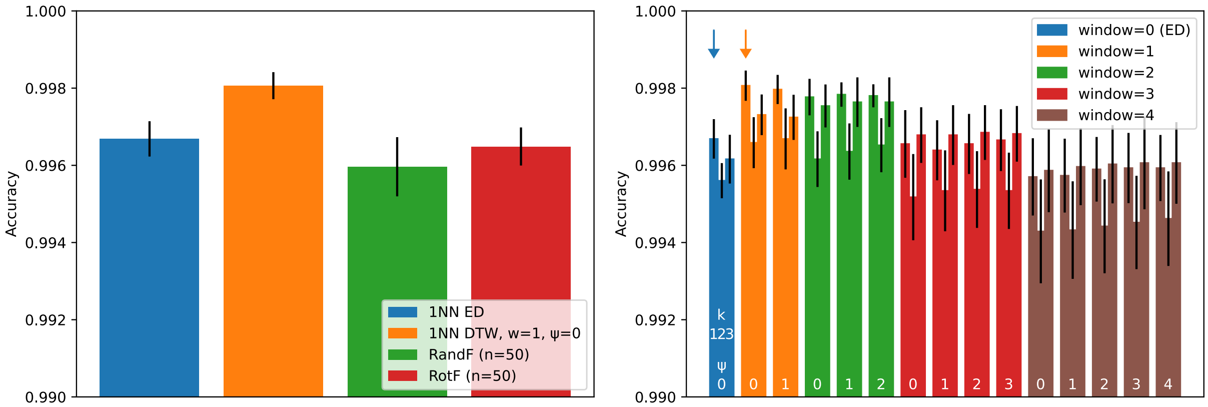


Figure 2.8: Left: Classification results of four classifiers on experimental dHRM data (5 train/test splits). Right: Cross validation to determine optimal window size (w), number of neighbors (k) and relaxation parameter psi (ψ). For this dataset, increasing the relaxation parameter psi or the number of neighbors does not improve performance. Best performing ED and DTW parameters are indicated by arrows.

2.4.2 U-dHRM Resolving Power

A comparison of the four classification methods tested on the 15 pairs of synthetic melts is shown in Fig. 2.11. First of all, it is important to notice how the whole series classifiers (1-NN ED and 1-NN DTW) struggle with this classification problem, only scoring >90% in 6/15 (1-NN ED) and 3/15 (1-NN DTW) cases (Fig. 2.12). This confirms we have not created a trivial model. Next, the ensemble methods (RotF and RandF) outperform the optimized whole series classifiers (50-NN ED and 50-NN DTW) across the board (Fig. 2.11 left and Fig. 2.13). The only pair which cannot be resolved with a desirable accuracy (>90%) by any of the classifiers is the first and most challenging pair. The sequences of pair 1 differ by two insertions, an extra G and C in Yersinia pestis compared to Yersinia pseudotuberculosis. To our surprise, pair 6, which only has a single SNP, a C/T substitution, can be resolved with an average accuracy >99% (across ten noise models) by both RotF and RandF. This shows that, according to this model, in some cases resolving melt curves with single nucleotide resolution can be achieved in dHRM.

To compare the performance of the noise models, we performed a Z-test using the mean error rates of each classifier for each noise model and adjusting for the number of curves in the

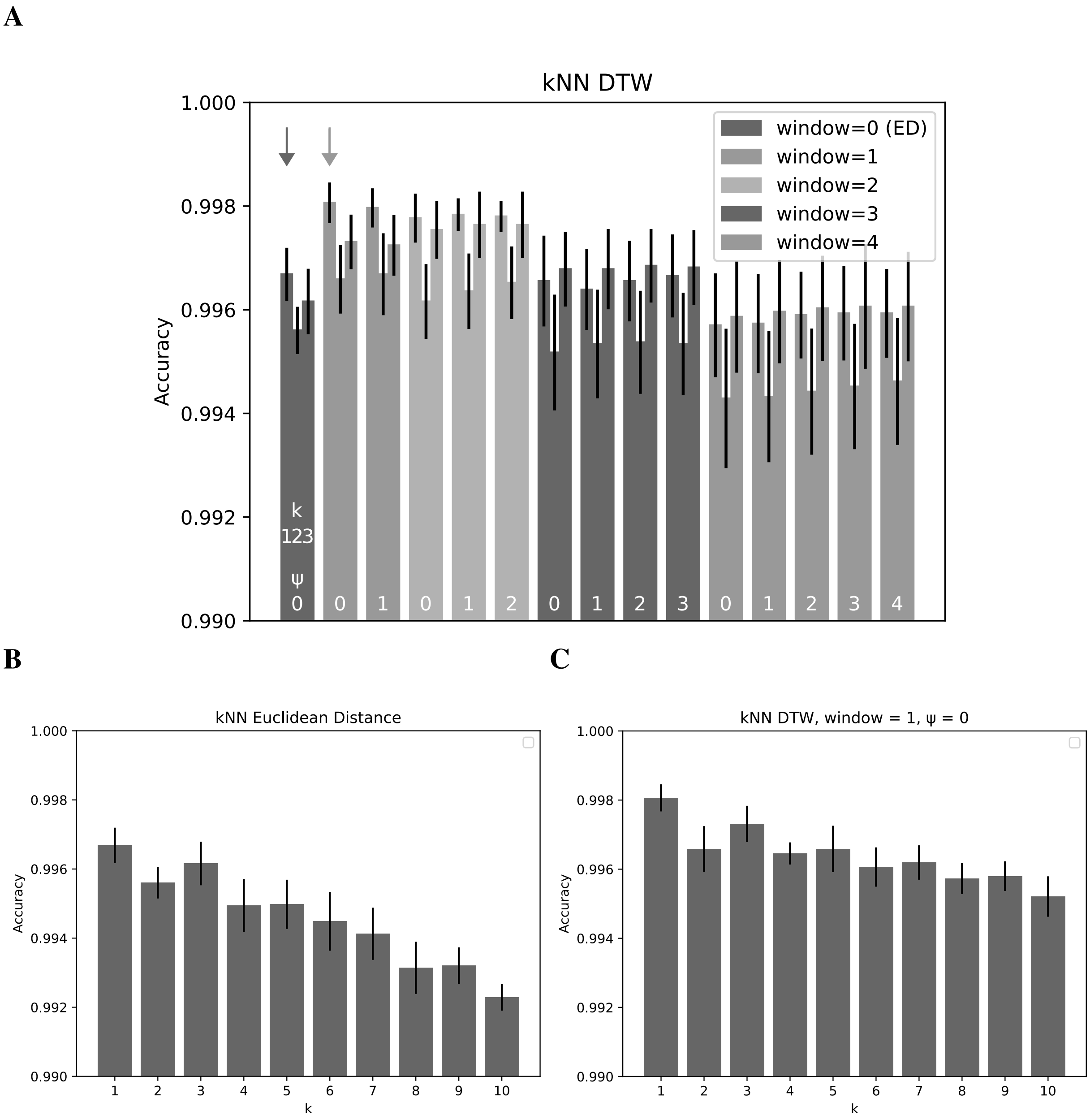


Figure 2.9: Generalization of 1NN ED and 1NN DTW to kNN. **A** Average classification accuracy for experimental dHRM data (5 train/test splits). Increasing the relaxation parameter psi (ψ) does not increase performance, increasing the number of neighbors (k) does not either. Best performing ED and DTW settings indicated by arrows. **B** Increasing k beyond 3 does not improve performance for kNN ED. **C** Increasing k beyond 3 does not improve performance for kNN DTW either.

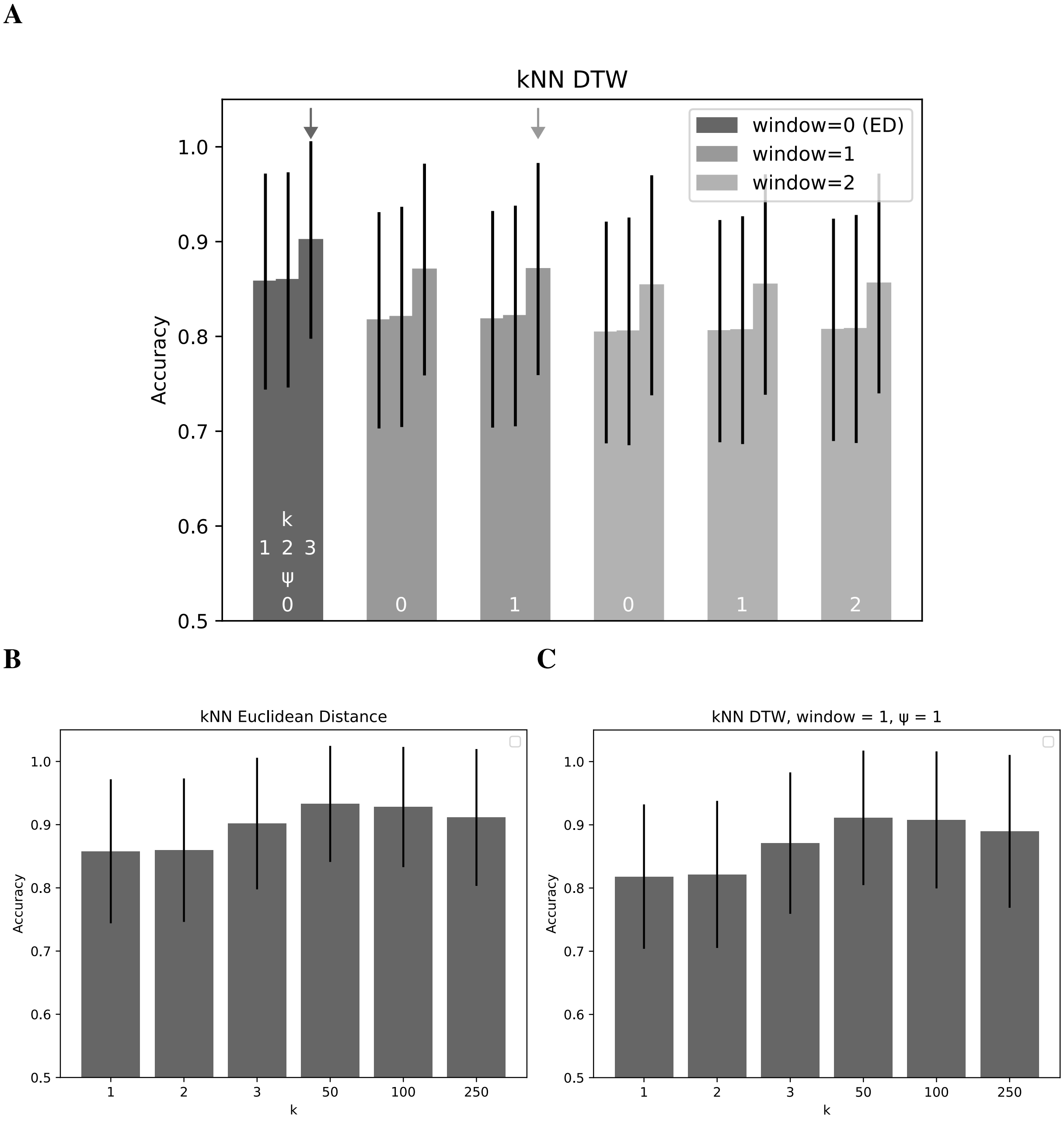


Figure 2.10: Generalization of 1NN ED and 1NN DTW to kNN. **A** Average classification accuracy across all noise models and synthetic melt pairs (9 train/test splits per noise model and pair). Increasing the relaxation parameter psi (ψ) does not significantly increase performance, increasing the number of neighbors (k) does. Best performing ED and DTW settings indicated by arrows. **B** Increasing k beyond 3 improves performance even further. An optimum can be found close to k=50. **C** Increasing k beyond 3 improves performance even further. An optimum can be found close to k=50.

Table 2.3: Standard deviation in classification accuracy for each pair and classifcation method (averaged over 10 noise models)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Average
1NN ED	3.50	9.82	7.51	7.24	6.49	2.75	7.02	6.59	5.91	6.72	6.45	3.99	7.89	4.99	4.01	6.06
1NN DTW, w=1, ψ=1	3.64	7.46	6.61	7.52	7.63	4.47	6.97	6.24	5.96	7.23	6.57	6.41	7.61	6.54	6.91	6.52
50NN ED	5.38	8.27	5.85	5.44	4.20	2.68	3.99	3.78	3.70	3.74	6.03	3.45	4.59	3.07	2.56	4.45
50NN DTW, w=1, ψ=1	5.83	7.71	6.99	5.79	5.79	3.32	4.61	4.03	4.09	4.50	6.21	4.32	5.30	4.04	3.67	5.08
RandF (n=50)	5.88	4.78	2.95	1.86	1.19	0.84	1.01	0.99	1.09	1.36	2.31	0.80	1.66	0.86	0.69	1.88
RotF (n=50)	7.06	4.40	3.03	1.53	0.94	0.78	0.95	0.87	0.99	1.05	1.83	0.65	1.43	0.81	0.60	1.79

training sets of each of the noise models. Overall the results are very consistent, there is just one noise model that performs significantly different from all others across all classifiers: model 7 ($P < 0.05$ for RotF and $P < 0.01$ for all others, see Fig. 2.11 right). Model 7 corresponds to residuals from Streptococcus gallolyticus, which has the highest well-to-well variance across all 10 species (median DTW in Table 2.1). We will consider disregarding this noise model for future use.

Table 2.3 shows the standard deviation of the 9-fold cross validated classification results (3 initial splits of the residuals times 3 train/test splits in each of these), averaged across the 10 noise models. The ensemble methods show smaller variation, but in general, classification results are consistent across classifiers and noise models.

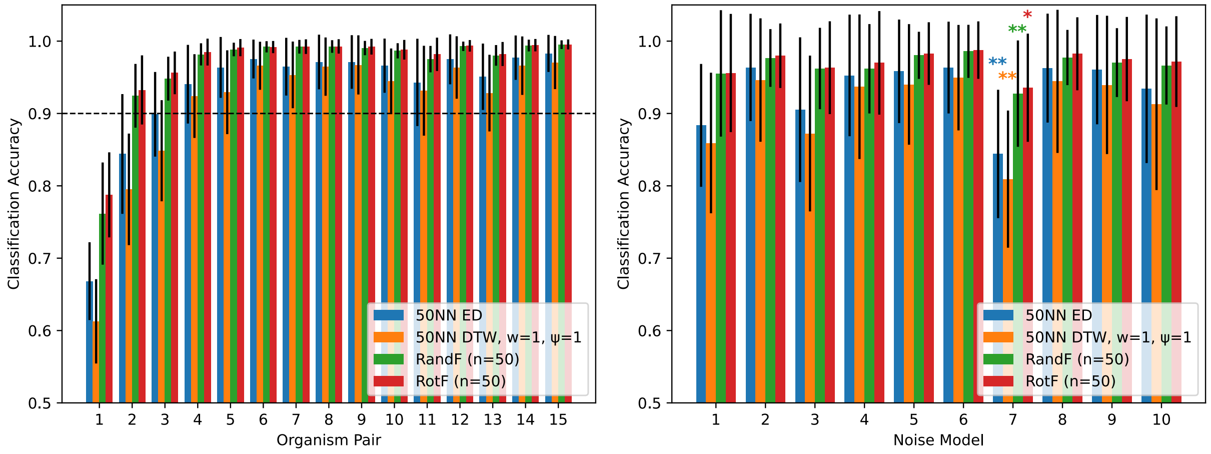


Figure 2.11: Left: Classification results of the four classifiers on the fifteen pairs of noise-augmented synthetic uMelts. The ensemble methods (RotF and RandF) perform with an accuracy >90% for all but the most challenging pair (pair 1). They outperform optimized whole series classifiers (50NN ED and 50NN DTW) across the board. Right: Comparison of noise model performance. Outlier models for each classifier are marked with asterisks (*P $<$ 0.05 and **P $<$ 0.01). Noise model 7, which corresponds to residuals from Streptococcus gallolyticus, performs significantly different from all other noise models across all four classifiers, which can be explained by its high DTW variability (Table 2.1).

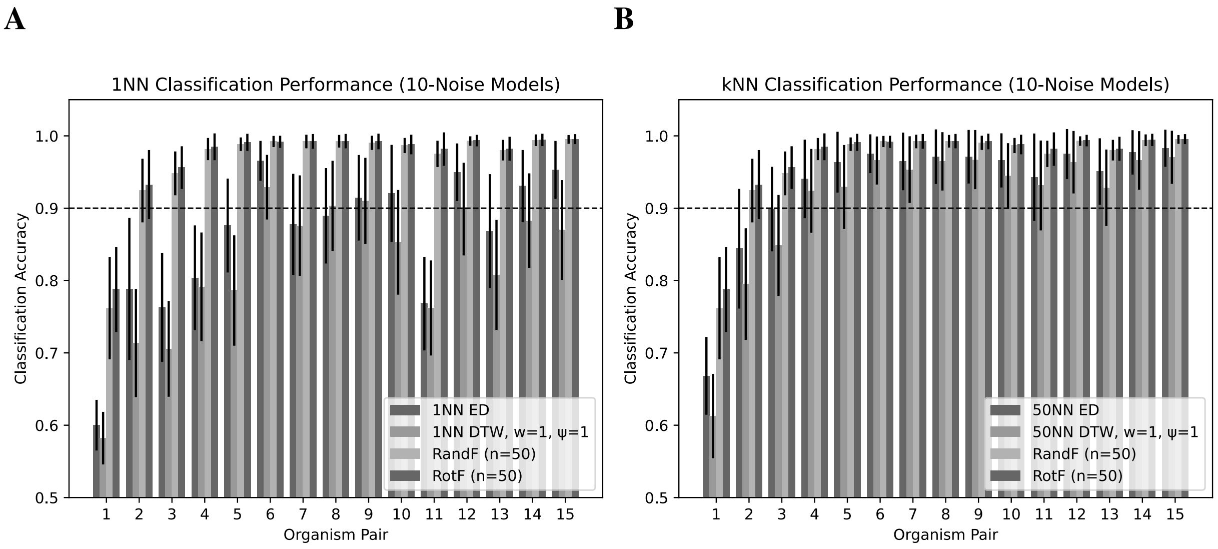


Figure 2.12: Generalization of 1NN ED and 1NN DTW to kNN. **A** 1NN Classification performance. **B** A significant improvement can be seen for both ED and DTW when increasing the number of neighbors to k = 50

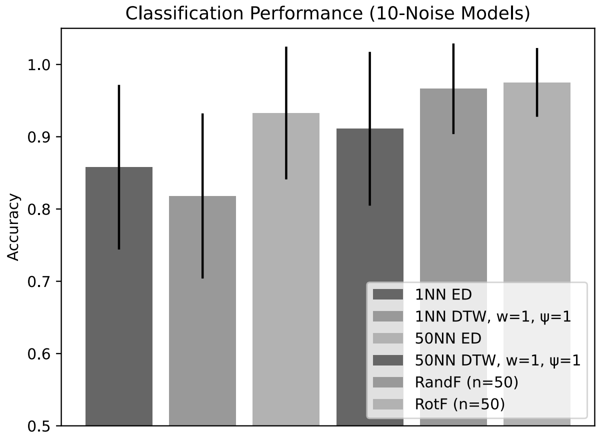


Figure 2.13: Average classification accuracy across all noise models and synthetic melt pairs (9 train/test splits per noise model and pair). The ensemble methods (RotF and RandF) outperform the whole series classifiers, even after generalizing them to kNN and selecting the optimum number of neighbors k = 50.

2.5 Discussion

The experimental classification results confirm the potential that U-dHRM has as a universal infectious disease diagnostic tool. The modeling results show great promise for using dHRM as a cheaper, faster and less complex solution to any application that involves classifying genetic sequences (infectious disease diagnostics, forensics, DNA data storage, etc.). Comparing dHRM with other emerging screening technologies, there are two remaining challenges to be overcome. First, until more accurate melt curve prediction models materialize, dHRM is only able to recognize sequences it already has available in its database. Second, due to its inherent variation, dHRM might be most useful in applications where single-nucleotide resolution is not required, although our model suggests that even there, it could play a role. Limitations of this work include: (1) We have not been able to find an underlying distribution that captures the variation in noise across different chips and it is therefore possible that future data will have different noise that could be more difficult to classify. Knowing the exact distribution of the noise might not be as important though, as long as there are machine classifiers that are able to capture those nuances.

(2) We have chosen to compare simple and readily available benchmark classification methods. Other methods such as COTE [51], which uses a collective of transformation ensembles, have been shown to significantly outperform the benchmark classifiers [30].

Reducing Noice

Having some level of noise is inherent to most technologies and this holds true for dHRM as well. We'll discuss some of the sources and potential mitigation strategies here:

- dPCR efficiency. A more efficient dPCR reaction will result in a higher number of amplicons inside the well, which corresponds to a higher fluorescence base level (at room T). Experience has shown that a high base level (above a certain threshold) is needed for the melt peak to be observable as separate from the background noise.
- Differences in amplicon sequences. Caused by polymerase base substitution errors. Mitigated by using high-fidelity polymerase enzymes.
- Low resolution per well. A single digital well is captured by around a dozen pixels in our dHRM setup. A higher pixel density could reduce variability.
- Different types of image noise such as Gaussian noise, Salt-and-pepper noise and shot noise. These types of noise are difficult to reduce on the hardware side but can be mitigated by using post-processing image filters at the chip level.
- Temperature variations between center and edge of chip can impact PCR efficiency and introduce variations in melting temperature.
- Development of bubbles at higher temperatures, mainly caused by degassing of dissolved gas within the reaction chamber. No easy solution is currently known but this can be addressed by image processing, either at the chip- or individual well-level.

Clinical Context

Mutations

Mutations in conserved regions do occur [52, 53], and if they coincide with primer regions they can compromise any assay based on amplification through PCR. A technology like whole-genome-NGS (WG-NGS) mitigates that risk, by providing non-targeted identification of microbes using deep sequencing of biological samples without a priori [54]. WG-NGS does have its own challenges with a time to results of 2–3 days and a high cost which makes it difficult to repeat the assay as frequently as culture. Interpretation of the results by physicians is another challenge for WG-NGS. It requires appropriate training and is certainly easier in normally sterile samples, like blood, than in polymicrobial samples (respiratory samples, feces, etc.) [54].

Polymicrobial Samples

One can imagine a sample in which a benign microbe, perhaps part of the natural flora, is present in excess of the pathogen causing symptoms. No matter how the sample is processed or diluted, it might be impossible to avoid getting excess microbe in the same dPCR well as the pathogen. The universal primers could anneal to the benign microbe and outcompete the amplification reaction targeting the pathogen. In the best case we would observe a mixture of melt curves and maybe with enhanced processing could try to subtract the melt curve attributed to the benign microbe. In the worst, and perhaps more likely case, we wouldn't observe that any mixture happened at all and attribute the melt curve to the benign microbe. Again, non-targeted WG-NGS could present a possible alternative [54], with the caveat that getting nucleic acid read-outs comprising the entire sample biome will need additional specialist/physician interpretation.

Sepsis

Applications such as sepsis, where time-to-result is perhaps the most important requirement for a diagnostic test [22], is where dHRM could potentially outshine other technologies. Blood culture analysis remains the gold standard for diagnosing sepsis despite advances in molecular diagnostic technologies [22]. However, because blood culture is slow and inconvenient it can't significantly influence the initial treatment of patients. The rapid initiation of effective antibiotic therapies depends on the capacity of a sepsis diagnostic test to capture clinically relevant organisms along with antimicrobial resistance within 1 to 3 h [22]. Such a test would require high sensitivity along with a high negative predictive value in order to be able to deliver the appropriate, narrow-spectrum antibiotics. It should also utilize small sample volumes and be able to detect polymicrobial infections and contaminants [22]. In this work we have only discussed the accuracy of correctly identifying pathogens that are already present in the dHRM database. Accuracy relates to sensitivity and specificity in the following way:

Sensitivity = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)
Specificity = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)
Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)

Although not explicitly calculated in this work, accuracies over 99.5% in this work will correspond to a very high sensitivity and specificity as well. We have not validated the requirement of a high negative predictive value in this work, nor have we investigated the performance of machine learning algorithms with novel or unseen pathogens. This last topic however will be discussed in the next chapter.

Future work will include improving the hardware to decrease variation between chips, improving the software to enable use of increasingly sophisticated algorithms, and experimentally validating the predictions we have made in this work. One key aspect will be estimating the optimal amplicon length for use in U-dHRM, as this is expected to be a trade-off between resolving power (shorter length equals higher resolving power) and overcoming background/environmental contamination levels (longer length desired). Another interesting avenue that has yet to be investigated is determining the extent to which amplicon %GC content plays a role.

2.6 Acknowledgements

Chapter 2, in part, was published in Bioinformatics on 23 December 2020. Lennart Langouche, April Aralar, Mridu Sinha, Shelley M Lawrence, Stephanie I Fraley, Todd P Coleman. The dissertation author was the primary investigator and author of this material. The publication can be accessed online at https://doi.org/10.1093/bioinformatics/btaa1053. Authors would like to extend thanks to thank Zachary Dwight for his assistance in obtaining synthetic melt curves using uMelt software. Fig. 2.1A and Fig. 2.1C were created with BioRender.com.

Support for this study was provided by a grant from the NIH NIAID to S.I.F., S.M.L., and T.P.C., award number R01AI134982 and a Burroughs Wellcome Fund Career Award at the Scientific Interface to S.I.F., award number 1012027.

Chapter 3

A Probabilistic Approach to Melt Curve-Based DNA Profiling Enables Novel Genotype Detection

3.1 Abstract

Surveillance for genetic variation of microbial pathogens, both within and among species, plays an important role in informing research, diagnostic, prevention, and treatment activities for disease control. However, large-scale systematic screening for novel genotypes remains challenging in part due to technological limitations. Towards addressing this challenge, we present an advancement in universal microbial high resolution melting (HRM) analysis that is capable of accomplishing both known genotype identification and novel genotype detection. Specifically, this novel surveillance functionality is achieved through probabilistic modeling of sequence-defined HRM curves, which is uniquely enabled by the large-scale melt curve datasets generated using our high-throughput digital HRM platform. Taking the detection of bacterial genotypes as a model application, we demonstrate that our algorithms accomplish

an overall classification accuracy over 99.5% and perform novelty detection with a sensitivity of 0.94, specificity of 0.96 and Youden index of 0.9. Since HRM-based DNA profiling is an inexpensive and rapid technique, our results add support for the feasibility of its use in surveillance applications.

3.2 Introduction

High Resolution Melting (HRM) analysis is a rapid, inexpensive, and powerful post-amplification nucleic acid characterization technique that is increasingly being used to profile DNA sequences for research and clinical diagnostic application [55, 39, 56, 57, 58, 59]. To accomplish HRM analysis, a DNA-binding dye is added to a sample of DNA, where it fluoresces upon intercalating into the double-stranded structure. Then the sample is heated. As the temperature increases, the double-stranded DNA denatures into single strands, releasing the intercalating dye and losing fluorescence. This loss in fluorescence with heating is recorded as a function of temperature, generating a melt curve.

Advances in HRM technology have taken this analysis from a simple check on amplification product homogeneity to a tool for heterozygote detection [55]. Subsequently, heat transfer and reaction engineering have improved HRM such that a melt curve could reliably be used as a sequence-specific signature [60]. Efforts to harness the potential of HRM as a broad-based sequence profiling tool have necessitated the use of machine learning for HRM curve analysis and classification. Machine learning approaches that have been employed include Naïve Bayes (NB) [26], Support Vector Machines (SVM) [24, 25], k-Nearest Neighbors using Dynamic Time Warping [29], Random Forest (RandF) [61, 62], and Neural Networks [63].

As digital PCR has gained popularity in the previous decade, so has the transition from regular HRM to digital HRM (dHRM) [64, 65]. Initially developed by our lab [23, 22, 66], our dHRM platform uses a custom heat transfer and imaging system to reliably melt thousands of digital PCR (dPCR) reactions simultaneously. Universal primers are used to target con-served sequences flanking hypervariable regions of the bacterial 16S rDNA gene in our dHRM system. This enables broad-based amplification of bacteria, while relying on melt and ML to specify organism identity based on the signatures of the hypervariable sequences. This approach also offers the ability to detect individual organisms via the ‘digital’ design, wherein each reaction contains only zero or one genome as template DNA. This development is accompanied by a large increase in available melt curve data. Where a regular HRM experiment might be performed on a 96-well plate, our digital HRM experiments make use of a commercial dPCR chip with 20,000 partitions. This means that one experiment can generate 20,000 HRM curves, as opposed to the standard 96-well format, a 200-fold increase. This increase in data availability opens many possible routes for further exploring this data using ML.

In the past, HRM analysis was focused primarily on SVM [24, 25], because of its ability to perform well in high dimensional spaces and on small datasets. SVM entails finding the best n-1 hyperplane in an n-dimensional space that maximizes the margin between the classes in the data. Even with small amounts of representative melt curve data to ’learn’ from, SVM algorithms show excellent performance [24, 25]. However, with the formulation of the one-versus-one SVM (OVOSVM) applied to HRM analysis, a problem emerges when the melt curve to be predicted has no representation in the database. For example, with a database containing melt curves from known pathogens, the currently used OVOSVM algorithm will erroneously classify test melt curves from an emerging pathogen (that is not represented in the database) into a known pathogen class. For HRM to reach its potential as a broad-based surveillance tool, there is the need for machine learning classification algorithms that enable the identification of such ’novel’, emerging DNA melt curves that are not yet represented in the database.

Probabilistic classification methods are suitable for such tasks [67]. By outputting classification probabilities in addition to class predictions, these methods allow for the assessment of the degree of uncertainty of the classification, which can then be used for novelty detection. To our knowledge, there has only been one attempt at implementing novelty detection for HRM [26]. The suggested approach uses Naïve Bayes with a custom distance metric based on the Hilbert Transformation. Useful information is lost in this approach as the melt peaks are aligned to a single temperature. We have chosen to pursue a more broadly applicable strategy. Our approach in the present study is to leverage the large-scale datasets generated using our dHRM platform to evaluate the utility of several classification algorithms for novelty detection, as well as the performance of these models in classifying bacteria-derived dHRM melt curves. This approach requires a larger amount of melt curve data than previously reported.

In addition to experimentally generated melt curves, we also included melt curves derived in silico using the uMelt tool [44]. We focus on one application as a test case: the detection of known and novel bacterial pathogens relevant to the diagnosis of bacteremia in neonates. Although a small number of bacterial organism species are implicated in the majority of neonatal bacteremia cases, opportunistic infections and emerging pathogens can occur [68]. We demonstrate the ability to automatically detect rare pathogens that are not represented in our library of organism melt curves through their 'anomalous' melt curve signatures.

3.3 Materials and Methods

3.3.1 Bacterial Strains

Table 3.1 lists the bacterial species included in the present study and the number of melt curves for each species. Figure 3.1 shows examples of their melt curves. These organisms make up the majority of the causative pathogens implicated in neonatal sepsis [68, 69]. The organisms were isolated and shared by Dr. David Pride (University of California San Diego School of Medicine) or purchased from the American Tissue Culture Collection (ATCC, Old Town Manassas, VA). Bacteria were cultured in Lurie-Bertani (LB) broth or Tryptic Soy broth (TSB), as required, and incubated overnight at 37°C.

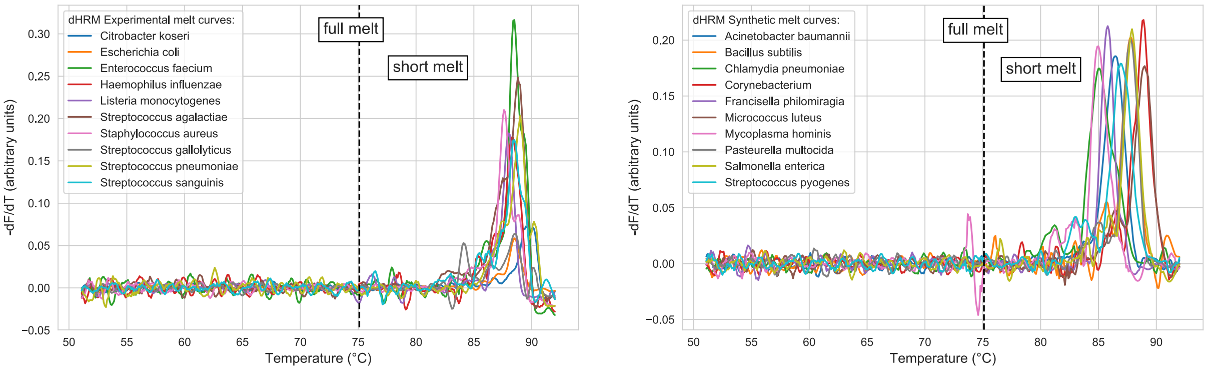


Figure 3.1: Overview of dHRM datasets. Left: Experimentally obtained dHRM melt curves. Right: Ten examples of synthesized melt curves using a combination of uMelts and real dHRM melt curve noise. Full and short melt refer to using the entire length of the melt curve or a shorter window around the melt peak location.

3.3.2 Bacterial Genomic DNA Extraction and PCR

Following overnight culture, bacterial genomic DNA was extracted using Wizard Genome DNA Purification kit (Promega Corporation, Madison, WI). Spectrophotometric absorbance measurements were used to assess the quality and concentration of the extracted DNA, and sequencing was conducted to further confirm the identity of the species. Genomic DNA dilutions were prepared for use with dPCR. Commercially available QuantStudio 3D Digital PCR 20K chip v2 (Applied Biosystems, Foster City, CA) were used for amplification following the manufacturer's recommended process, except for reagents. As described previously, the dPCR master mix was optimized for the recommended loading volume of 14.5 μL per reaction, and contained 1X Phusion HF Buffer con-taining 1.5 mM MgCl2 (Thermo Fisher Scientific, Waltham, MA), 0.15 μM forward primer 5'-GYGGCGNACGGGTGAGTAA-3' (Integrated DNA Technologies, Coralville, IA), 0.15 μM reverse primer 5'-AGCTGACGACANCCATGCA-3' (Integrated DNA Technologies, Coralville, IA), 0.2 mM dNTPs (Invitrogen, Carls-bad, CA), 2.5X EvaGreen (Biotium, Freemont, CA), 2X ROX (Thermo Fisher Scientific, Waltham, MA), 0.02 U/μL of Phusion HotStart Polymerase (Thermo Fisher Scientific, Waltham, MA), 1 μL of sample, and ultrapure PCR water (Quality Biological Inc., Gaithersburg, MD) to bring the total volume to 14.5 μL. To

Table 3.1: Overview of species.

Experimentally obtained melt curves from:
Citrobacter koseri (taxid:545)
Enterococcus faecium (taxid:1352)
Escherichia coli (taxid:562)
Haemophilus influenzae (taxid:727)
Listeria monocytogenes (taxid:1639)
Staphylococcus aureus (taxid:1280)
Streptococcus agalactiae (taxid:1311)
Streptococcus gallolyticus (taxid:315405)
Streptococcus pneumoniae (taxid:1313)
Streptococcus sanguinis (taxid:1305)

Synthesized melt curves from:
Acinetobacter calcoaceticus (taxid:471)
Acinetobacter baumannii (taxid:470)
Aerococcus viridans (taxid:1377)
Bacillus anthracis (taxid:1392)
Bacillus cereus (taxid:1396)
Bacillus subtilis (taxid:1423)
Bacteroides fragilis (taxid:817)
Bordetella parapertussis (taxid:519)
Bordetella pertussis (taxid:520)
Campylobacter jejuni (taxid:197)
Chlamydia pneumoniae (taxid:83558)
Chlamydia trachomatis (taxid:813)
Citrobacter freundii (taxid:546)
Clostridium difficile (taxid:1496)
Clostridium perfringens (taxid:1502)
Corynebacterium (taxid:1716)
Coxiella burnetii (taxid:777)
Enterobacter aerogenes (taxid:548)
Enterococcus faecalis (taxid:1351)
Enterococcus gallinarum (taxid:1353)
Francisella philomiragia (taxid:28110)
Francisella tularensis (taxid:263)
Helicobacter pylori (taxid:210)
Klebsiella pneumoniae (taxid:573)
Legionella pneumophila (taxid:446)
Micrococcus luteus (taxid:1270)

Synthesized melt curves from: (continued)
Moraxella catarrhalis (taxid:480)
Mycobacterium fortuitum (taxid:1766)
Mycobacterium gordonae (taxid:1778)
Mycobacterium kansasii (taxid:1768)
Mycoplasma hominis (taxid:2098)
Mycoplasma pneumoniae (taxid:2104)
Neisseria gonorrhoeae (taxid:485)
Neisseria meningitidis (taxid:487)
Oligella urethralis (taxid:90245)
Pasteurella multocida (taxid:747)
Propionibacterium acnes (taxid:1747)
Proteus mirabilis (taxid:584)
Proteus vulgaris (taxid:585)
Pseudomonas aeruginosa (taxid:287)
Salmonella enterica (taxid:28901)
Serratia marcescens (taxid:615)
Staphylococcus epidermidis (taxid:1282)
Staphylococcus lugdunensis (taxid:28035)
Staphylococcus saprophyticus (taxid:29385)
Streptococcus pyogenes (taxid:1314)
Treponema pallidum (taxid:160)
Yersinia enterocolitica (taxid:630)
Yersinia pestis (taxid:632)
Yersinia pseudotuberculosis (taxid:633)

load the chip, a master mix reaction volume of 14.5μL was spread across following manufacturer's recommendation. After loading, the dPCR chip was cycled on a flatbed thermocycler with the fol-lowing cycle settings: an initial enzyme activation (98 °C, 30 s), followed by 70 cycles (95 °C, 30 s, 59 °C, 30 s, 72 °C, 60 s).

3.3.3 DNA Melt Curve Generation and Preprocessing

The U-dHRM device developed by our group has been previously described [22, 23]. Briefly, it comprises a copper plate on which the microfluidic dPCR chip is placed, thermoelectric heater/cooler (TE Technology Inc., Traverse City, MI), proportional-integral-derivative (PID) controller (Meerstetter Engineering GmbH, Rubigen, Switzerland), Class 1/3B resistance temperature detector (RTD) (Heraeus, Hanau, Germany) embedded in the copper block, K-type thermocouple (OMEGA Engineering, Stamford, CT), and heat sink. A thin layer of thermal grease added between the dPCR chip and copper block ensures efficient heat transfer. A custom-made adapter secures the device on-stage for optimal fluorescent imaging. With heat ramping, simultaneous fluorescent images from the DNA-intercalating dye, EvaGreen (Ex/Em: 488 nm/561 nm) and the control dye, ROX (Ex/Em: 405 nm/488 nm) are acquired with a Nikon Eclipse Ti microscope (Nikon, Tokyo, Japan). Melt curves are generated by implementing an automated imaging processing algorithm in MATLAB. Background subtraction was performed similar to the linear method described in [70]. This was done in order to align the tails of the melt curves horizontally with the x-axis, so they are most similar to the theoretically predicted uMelt curves, as can be seen in Figure 3.1 and Figures 3.2,3.3 and 3.4.

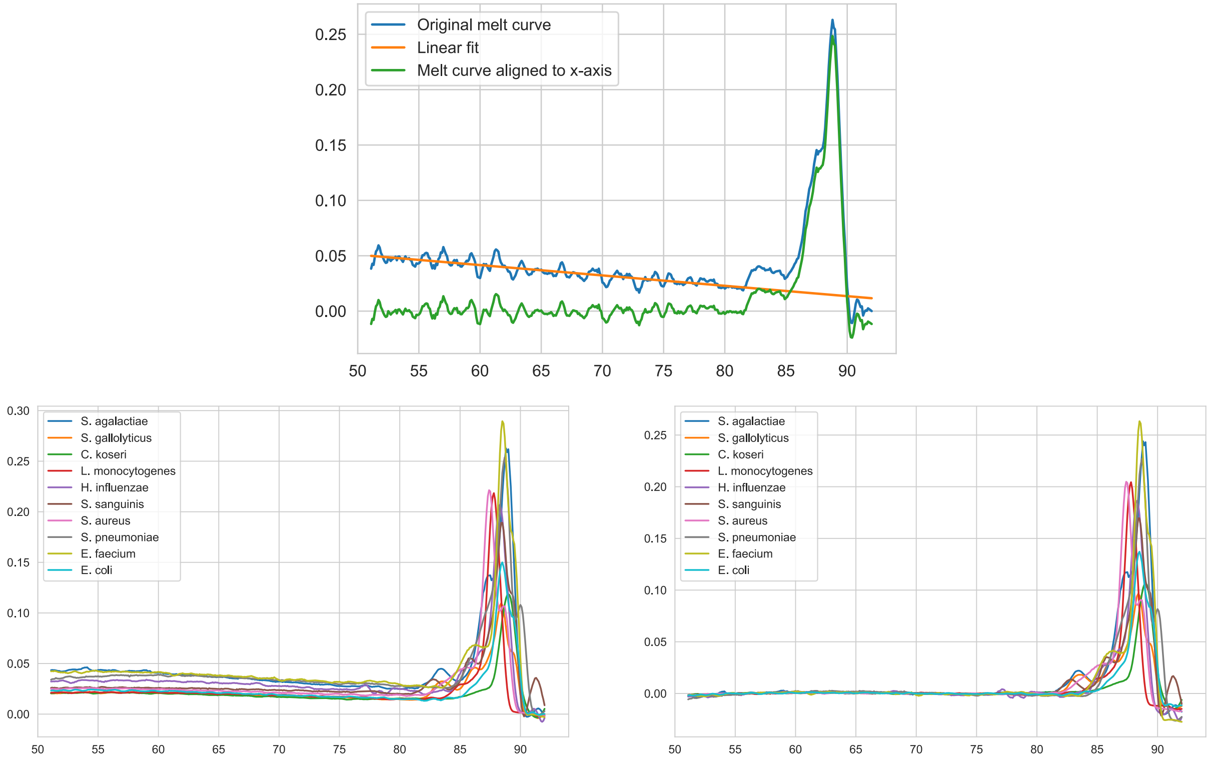


Figure 3.2: Data preparation. Linear background is subtracted to make sure experimental melt curves resemble synthetic ones. Top: mechanism. Left: experimental mean curves. Right: experimental mean curves after background subtraction.

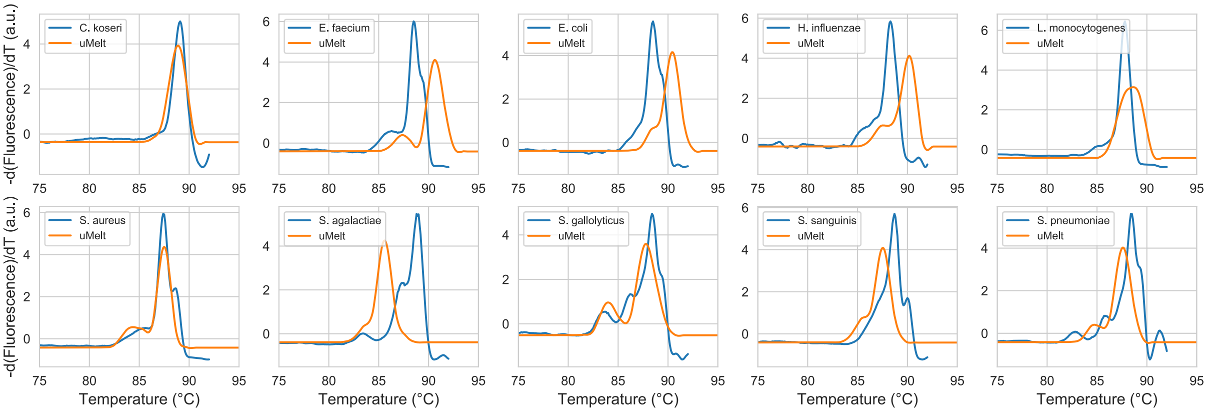


Figure 3.3: Comparison of experimentally obtained melt means and synthesized uMelt curves.

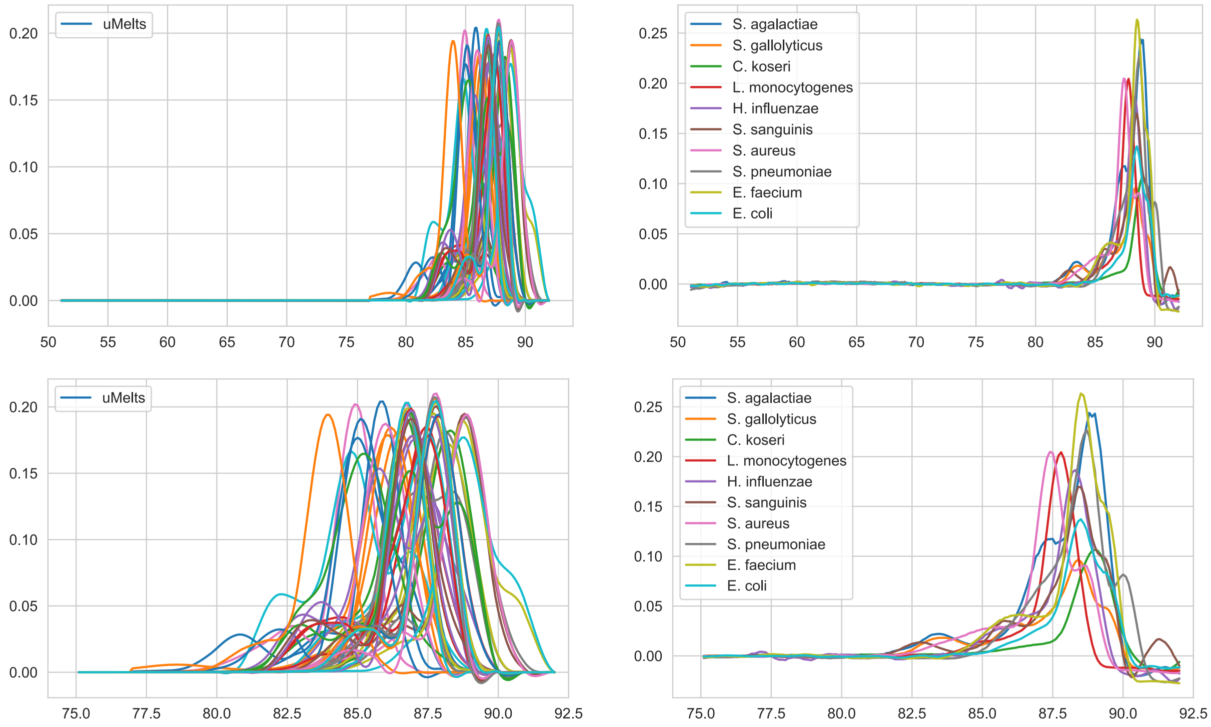


Figure 3.4: Overview of experimentally obtained and synthesized melt curves. Top: full melts. Bottom: short versions. For the experimental curves (right), the mean curve is shown. For the uMelts the original curve is shown (without applying dHRM noise). We wanted to assure these looked similar as the noise residuals to the mean (from the experimental data) are added to the uMelts.

3.3.4 Machine Learning Models

Logistic Regression

Logistic Regression is a supervised machine learning method used to assign discrete labels to observations. It returns a probability value by transforming its output using the ’logistic’ or ’sigmoid’ function [71]. Maximum Likelihood Estimation (MLE) is used in order to find the parameters of the Logistic Regression that result in an accurate model with minimum error. [72]. MLE consists of using gradient descent in order to find the parameters that maximize the log-likelihood function. [71, 72].

Naïve Bayes

Like other supervised machine learning methods, Naive Bayes uses features to assign labels to observations [73]. What is special about this method is that it assumes features are uncorrelated or independent, which is why it is called naive [73]. The algorithm is based on Bayes’ theorem which is defined by $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$, with $P(c|x)$ the posterior probability of class c given feature x, $P(c)$ the prior probability of class c, $P(x|c)$ the likelihood or the probability of feature x given class c, and $P(x)$ the prior probability of feature x [74].

Support Vector Machines

Support Vector Machines (SVMs) are a class of supervised machine learning methods that create lines (hyperplanes in higher dimensions) to separate data into classes [75]. The SVM algorithm first finds support vectors, which are the points from both classes that are closest to the line or hyperplane [75]. Then, the distance between the line and the support vectors, known as the margin, is computed. The goal of SVM is to maximize the margin and the optimal hyperplane is the one with the maximum margin [75]. Sometimes data is not linearly separable which means a straight line cannot be used to classify the data. In that case, the data can be converted to data that is linearly separable in a higher dimension using kernel functions. Now that the data is linearly separable, the data can be classified. [75].

Neural Networks (Multi-layer Perceptron)

Multi-layer Perceptron (MLP) is a supervised machine learning method that can learn complex (non-linear) functions by training on a dataset [76]. The algorithm consists of at least three layers: an input, an output, and one or more hidden layers. Each layer consists of neurons, which transform the value from the previous layer with a weighted summation, followed by a non-linear activation function like the hyperbolic tan function [76]. Finally, the values from the last hidden layer are transformed into output values or labels by the output layer. The MLP used in this chapter trains using gradient descent and the gradients are calculated using backpropagation [76].

Random Forest

Random Forest is a supervised machine learning method that uses decision trees as building blocks [77]. Decision trees split data at each node according to a feature ensuring that the resulting roots are as different from each other as possible and the members of each root are as similar to each other as possible [77]. A large number of decision trees that operate as an ensemble make up a Random Forest. Each tree can predict a class outcome and the class is chosen by majority vote. The random forest model performs well because a large number of uncorrelated trees operating as an ensemble will outperform any of the individual trees [77]. Random sampling of training data (Bagging) in combination with feature randomness (each tree can only pick from a random subset of features to split nodes), are used to make sure the trees in the Random Forest are uncorrelated. [77].

3.3.5 Machine Learning Model Selection

Some ML methods are better suited than others for probabilistic classifications. SVMs for example are inherently not probabilistic but have been adapted to provide probabilistic outputs [78]. ML methods vary widely in how they assign probability values to classifications [79]. Naïve Bayes for example, which makes unrealistic independence assumptions, pushes probabilities towards 0 and 1. Other models, such as neural networks, do not have these biases and predict well calibrated probabilities. Efforts have been made to correct these distorted probabilities to true posterior probabilities using Platt Scaling and Isotonic Regression [79]. We set out to compare five ML methods: Logistic Regression, Naïve Bayes (NB), Support Vector Machines (SVM), Neural Networks and Random Forest (RandF). At the end of this work we briefly discuss how calibrating the probabilities affects the results. We built and implemented all algorithms using the scikit-learn package with-in Python programming language [35]. All data and code are available on https://github.com/lenlan/dHRM-novelty-detection.

3.3.6 Leave-one-group-out (LOGO) Experiments

To assess the ability of a machine learning method to detect novel melt curves, i.e., melt curves belonging to species not previously represented in our database, we carried out leave-one-group-out experiments with a species being equivalent to a group. Using melt curves from all ten bacterial species, we held out melt curves for each of the ten species in turn. Melt curves of the remaining nine species were then randomly split into training and test sets (at a ratio of 80:20). The left-out species was then added to the test set to make up the new test set. The machine learning model was fit on the training set and evaluated on its ability to correctly identify the melt curves belonging to the left-out species as novel from the mixture comprising the 20% split of the remaining nine species and the withheld species. This was repeated ten times, for the ten species. Figure 3.5 presents a schematic of the novelty detection experimental approach. We assessed the

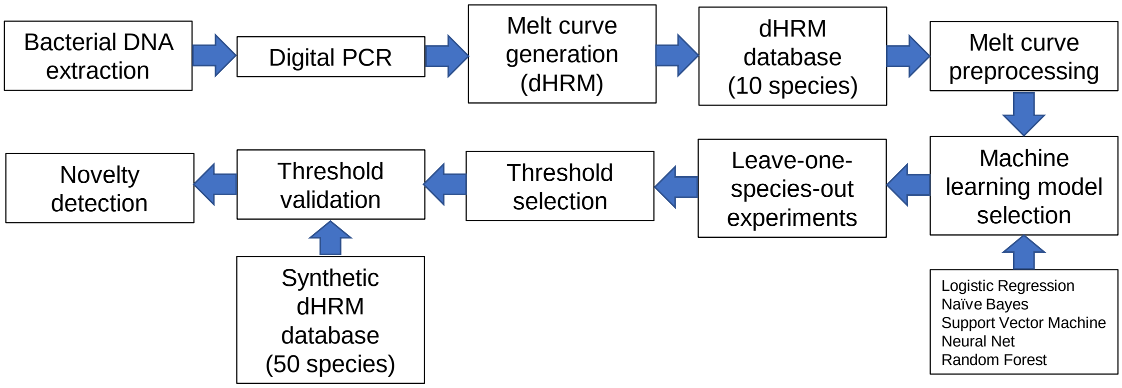


Figure 3.5: Workflow for probabilistic novelty detection.

performance of the novelty detection model using Youden's index, which is a single statistic that captures the performance of a dichotomous diagnostic test. It is the point on an ROC curve at which sensitivity + specificity - 1 is maximized. A prerequisite for using any ML method for novelty detection is that it can classify the non-novel data in our database with high accuracy. We briefly discuss classification performance in the results section.

3.3.7 Threshold Selection

Figure 3.6(left) shows the results of one LOGO experiment for one ML method (RandF, n=100). The optimal probability threshold can be found by plotting the ROC curve (Figure 3.6 (right)) and selecting its Youden index.

The next step is to find a threshold that works for all ten LOGO experiments combined. After all, we are interested in finding a practical threshold that can be used to differentiate the ten species in our dHRM database from any unknown or novel species. To find this 'practical' threshold, we accumulate the LOGO experiments and create their ROC curve to find the optimal threshold for the accumulated LOGO experiments (Figure 3.7 (left)). We investigated weighting each experiment (e.g. by number of left-out curves per LOGO experiment) but didn't observe a significant change in the selected threshold (data not shown). Having found the 'practical' thresh-

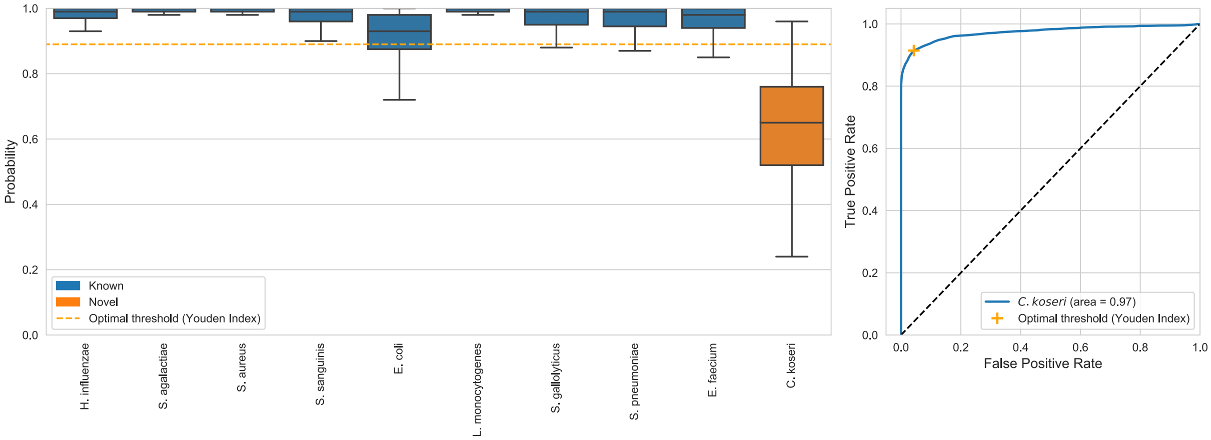


Figure 3.6: Leave-one-species-out cross validation to determine probability threshold. A. Boxplot of the classification probabilities of each of the ten species. In this experiment C. koseri is the left-out species, which means it is left out of the training set and added to the test set. This experiment is repeated for each of the ten species, and an optimal threshold can be found for each of them. This experiment is repeated for all ML methods (method shown here is RandF (n=100)). B. ROC curve that is used to find the optimal threshold. Youden's Index is chosen as the optimal threshold, it is the point on the ROC curve where sensitivity + specificity – 1 is maximized.

old, we can apply it to each of the ten LOGO experiments separately to see what performance would be like if we would have picked this threshold as our initial choice. This means that each LOGO experiment separately will be performing at a different (sub-optimal) operating point on their ROC curve, which is shown in Figure 3.7 (right).

3.3.8 Threshold Validation with Synthetic Melt Curves

We selected 50 clinically relevant bacterial pathogens, including category A and B biothreat agents and their surrogates from [41] (Table 3.1). We obtained their melt curves as outlined in [62] and used a similar model to [62] to add real dHRM noise to these synthetic melt curves (Figure 3.7). We chose to create 100 melt curves per species, with a different noise residual applied to each individual melt curve, making sure no residual is used twice in this creation process. Figure 3.1 (right) shows ten examples of the synthetic melt curves with added dHRM noise. Figure 3.8 shows an overview of the synthetically created melt curves for all 50 pathogens.

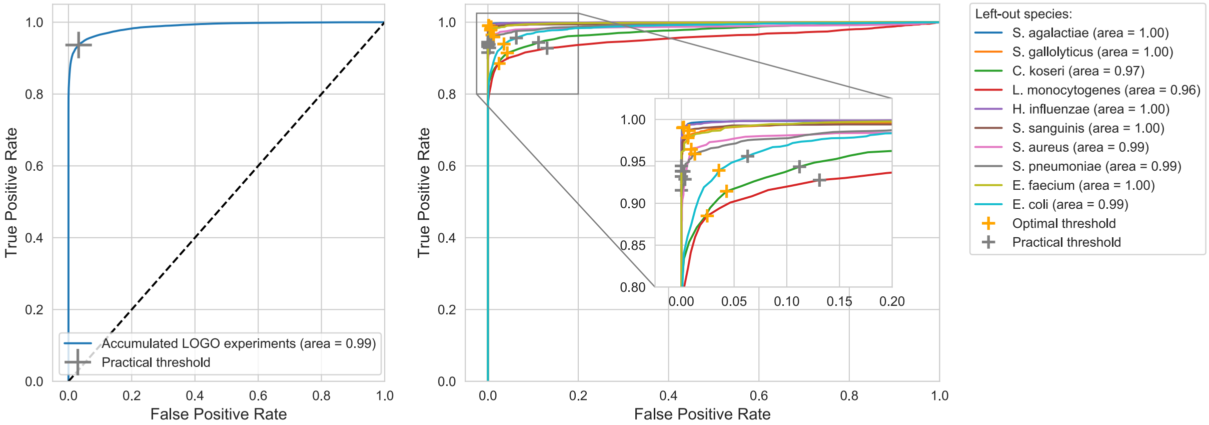
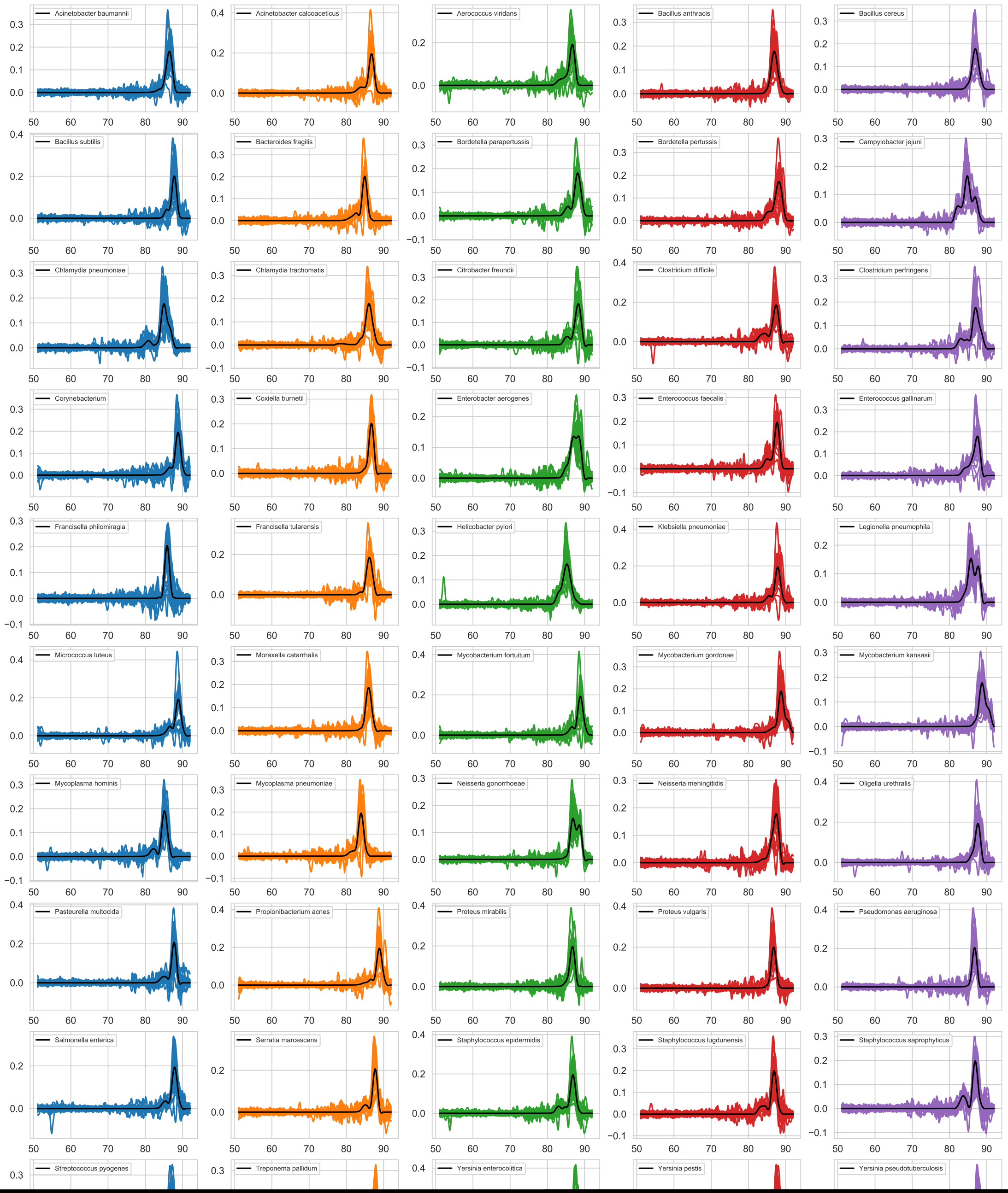


Figure 3.7: Accumulating the leave-one-group-out (LOGO) experiments results in a ’practical’ threshold. A. ROC curve for all ten LOGO experiments accumulated. The optimal threshold is again found by Youden’s Index. We have named it the ‘practical’ threshold as one threshold has to be chosen (as opposed to a separate threshold for each LOGO experiment) when further validating the model on unseen ’novel’ melt curves. It is the optimal threshold for all ten LOGO experiments combined. B. Choosing a practical threshold implies that each LOGO experiment individually will be performing at a suboptimal thresh-old, which translates to a suboptimal operating point on the ROC curve.



[redacted]

Figure 3.8: Overview of all synthesized melts. Residuals to the mean from each experimental melt curve are applied to the uMelts. The noise is scaled to the ratio of the peak heights (real peak/uMelt peak) and shifted (position real peak – position uMelt peak) so the noise should be in a similar position relative to the peak. No residual is used twice which means all the curves shown are unique. Each uMelt uses ten residuals from each of the experimental melts, which results in 100 curves per uMelt species.

3.4 Results

3.4.1 Classification Performance

The ML methods show very similar classification accuracy, which is summarized in Table 3.2. No significant difference can be observed between the ’full melt’ and ’short melt’ classification results, which implies there might not be any additional information in the tail of the melt curve. These results do not provide any hint as to which method would be most suited for novelty detection. The model fit time is provided to give the reader a sense of how these methods compare to each other performance-wise and can be interpreted as a possible tiebreaker in cases of similar classification or novelty detection performance.

3.4.2 Leave-one-group-out (LOGO) Experiments and Threshold Selection

Figure 3.9 summarizes the results of the LOGO experiments. The performance, as measured by Youden’s index is shown as a function of classification method. Each bar shows the average performance across ten LOGO experiments. ’Optimal thresholds’ means selecting the best threshold for each LOGO experiment individual-ly. ’Practical threshold’ means selecting the optimal threshold for the accumulated LOGO experiments and applying it to all LOGO experiments separately. Random Forest outperforms the other methods, but Neural Nets and SVMs still perform relatively well. This is not surprising, as random forests are known to be an efficient way of performing outlier detection in high-dimensional datasets[80].

Table 3.2: ML methods overview and classification results.

	Full melt		Short melt	
	Time (s)	Accuracy	Time (s)	Accuracy
Logistic Regression	21.55	0.995	13.18	0.995
Gaussian Naïve Bayes	0.17	0.977	0.07	0.974
SVM (rbf)	125.68	0.996	31.34	0.996
SVM (linear)	29.34	0.995	6.75	0.995
Neural net (identity)	22.2	0.992	25.24	0.992
Neural net (logistic)	23.35	0.996	25.17	0.997
Neural net (tanh)	17.77	0.996	19.99	0.997
Neural net (relu)	14.74	0.995	14.83	0.997
RandF (n=100)	1.29	0.995	0.89	0.995
RandF (n=500)	5.3	0.996	3.53	0.996

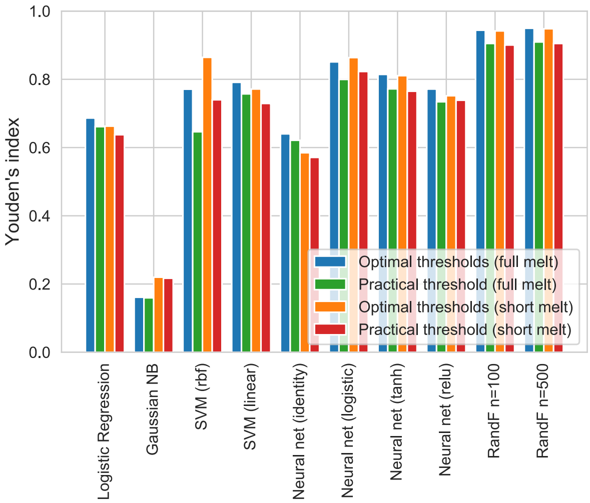


Figure 3.9: Summary of LOGO novelty detection results. Average novelty detection performance across ten species measured by Youden's index as a function of classification method. Optimal means selecting the best threshold for each leave-one-species-out experiment. Practical means selecting one threshold and applying it to all leave-one-species-out experiments.

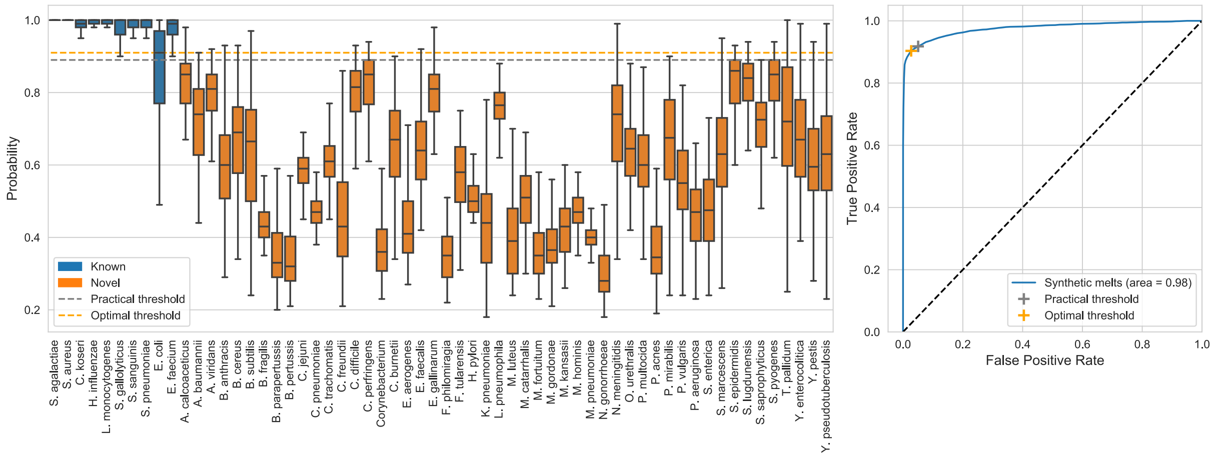


Figure 3.10: Validation of practical threshold on synthesized set of melt curves. A. The practical threshold, selected through the ten LOGO experiments, was validated on a new dataset consisting of 50 species, each with 100 synthetic melt curves with real dHRM noise. The performance of E. coli seemed to be an outlier, and this was confirmed using one-tailed t-tests. B. ROC curve. When the practical threshold is close to the optimal threshold, it serves as a validation for the threshold selection process. The ML method shown here is RandF (n=100).

3.4.3 Threshold Validation and Novelty Detection

Figure 3.10 shows an example of a boxplot (Figure 3.10(left)) and ROC curve (Figure 3.10 (right)) used to validate the previously obtained 'practical' threshold. The method shown in Figure 3.10 is RandF (n = 100), which was one of the best performing methods, figures 3.13-3.22 in section 3.6 show all other methods for comparison. When the practical threshold is close to the optimal threshold, the practical operating point on the ROC curve (Figure 3.10 (right)) will be close to the optimal one (Youden's index). When this is the case, it confirms that our proposed method for obtaining a practical threshold for novelty detection, is indeed valid. Figure 3.11 (left) shows an overview of the results for all ML methods. The average difference between the optimal and practically attained Youden index across the ten ML methods is just 0.015 with a standard deviation of 0.013. This serves as a confirmation of our threshold selection process using the accumulated LOGO experiments. Random Forest and SVM (rbf) perform the best, with the Neural Nets a close third.

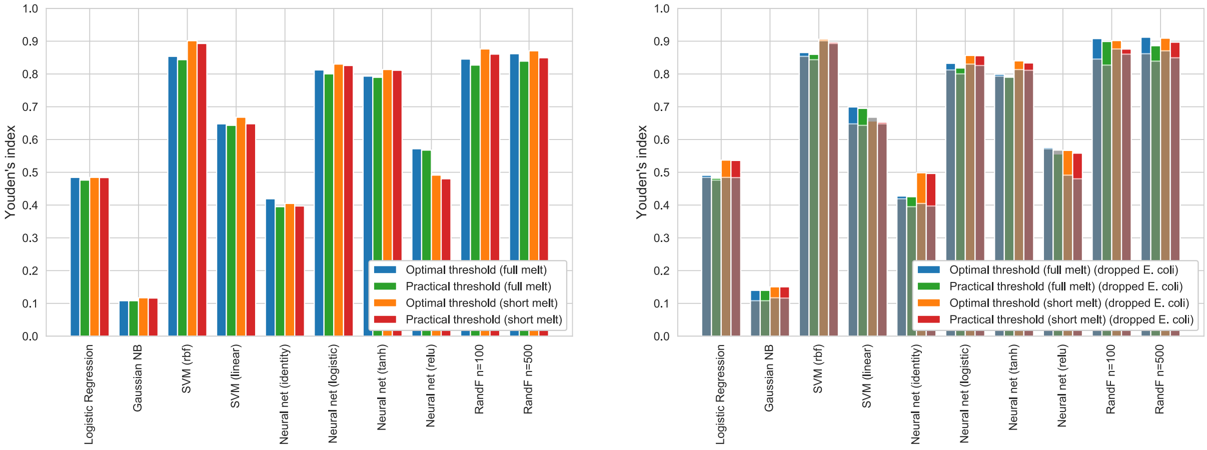


Figure 3.11: Summary of practical threshold validation results. A. Average novelty detection performance on 50 unseen species measured by Youden's index as a function of classification method. B. Dropping E. coli, an outlier group, results in im-proved performance for almost all methods (results including E. coli are overlayed in gray). SVM (rbf) is more robust against this outlier behavior and sees less improvement.

3.4.4 Further Improvements

We observed that the classification probabilities for the E. coli group of melt curves are more spread out (Figure 3.10 (left)) and might be an outlier group compared to the other species. This was apparent for multiple ML methods and was confirmed with one-tailed t-tests (e.g. for the short melt curves: $P < 0.01$ for Logistic Regression, Gaussian NB, SVM (rbf and linear), Neural Net (identity and relu) and RandF (n=100 and n=500)). As a result of this, we ran all steps again, but this time leaving out E. coli, to see if we could further optimize our novelty detection method. Figure 3.11 (right) shows that results do indeed improve when leaving E. coli out. Random Forest and SVM (rbf) are the top performers, and their results are further summarized in Table 3.3. The best performance achieved is a Youden index of 0.90, corresponding to a specificity of 0.96 and sensitivity of 0.94.

We also investigated whether calibrating the probabilities using scikit-learn's 'Calibrated-ClassifierCV' function would improve the outcome. We tested both the 'sigmoid' method, which corresponds to Platt's method (i.e., a logistic regression model) or the 'isotonic' method, which is

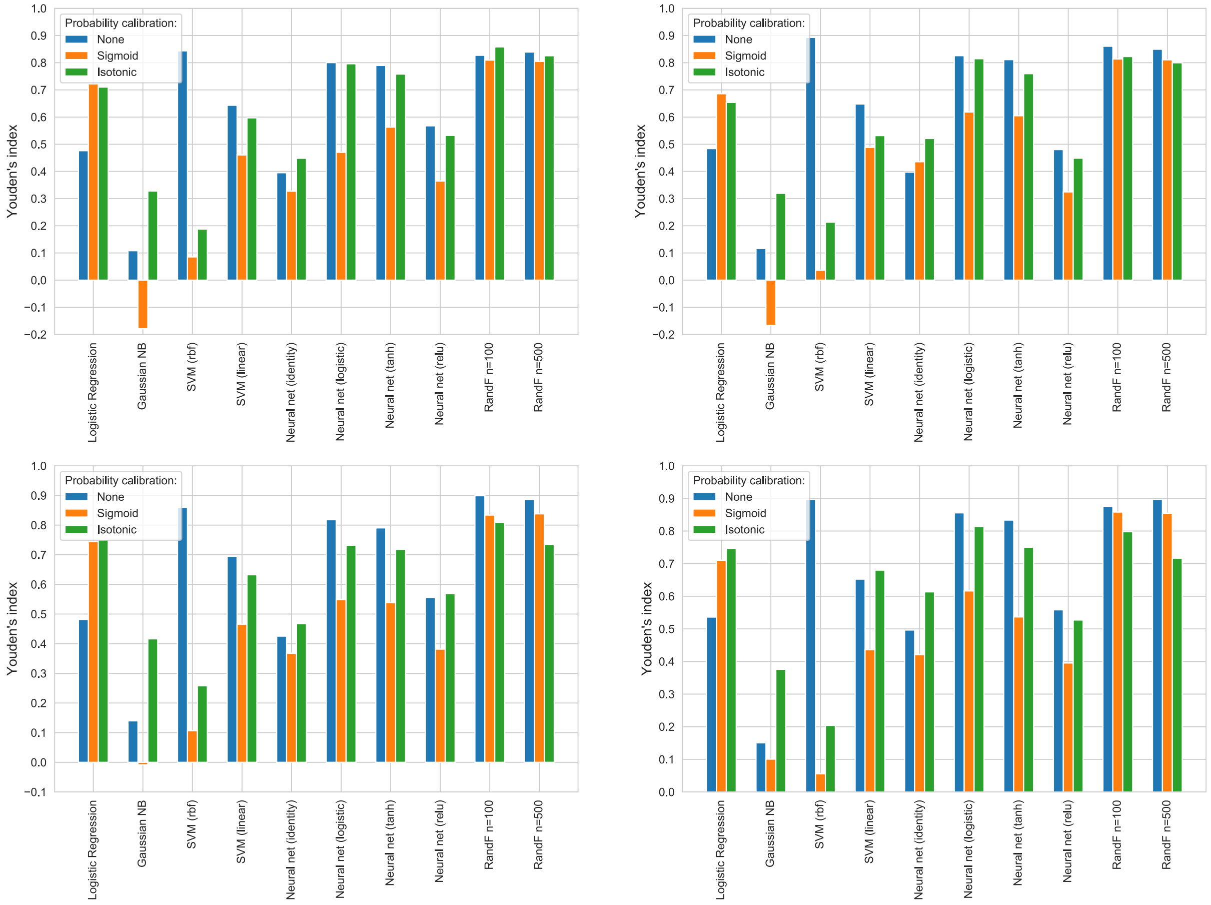


Figure 3.12: Probability calibration results. Top: with E. coli. Bottom: dropped E. coli. Left: full melt. Right: short melt. Calibrating probabilities improves novelty detection for Logistic Regression and Naïve Bayes. None of the calibrated methods outperform the best results (SVM, Neural Net, RandF) as outlined in Figure 3.11 and Table 3.3.

a non-parametric approach. Results are summarized in Figure 3.12. As expected, we see a large improvement for Naïve Bayes. We also see a significant improvement for Logistic Regression. None of the calibrated methods outperform the best results (SVM, Neural Net, RandF) as outlined in Figure 3.11 and Table 3.3 though.

Table 3.3: Overview of best results.

	Optimal specificity	Optimal sensitivity	Optimal Youden	Practical specificity	Practical sensitivity	Practical Youden
Full melt						
SVM (rbf)	0.93	0.94	0.87	0.94	0.92	0.86
RandF (n = 100)	0.98	0.93	0.91	0.96	0.94	0.9
RandF (n = 500)	0.98	0.93	0.91	0.93	0.95	0.89
Short melt						
SVM (rbf)	0.94	0.97	0.91	0.96	0.94	0.9
RandF (n = 100)	0.97	0.93	0.9	0.92	0.95	0.88
RandF (n = 500)	0.98	0.93	0.91	0.96	0.94	0.9

3.5 Discussion

Our work demonstrates the utility of probabilistic classification algorithms in resolving multiple bacterial organism melt curves, and in identifying previously unknown (novel) melt curves that are not represented in the database. The large amount of dPCR chip-generated melt curve data enabled the development of probabilistic classifier and novelty detection algorithms, which distinguishes this study from previous studies applying non-probabilistic methods to small datasets of melt curves [81, 82].

Previously used ML methods for dHRM analysis [81] are unsatisfactory for melt curve novelty detection, i.e., detecting when test melt curves are not represented in the melt curve database. The only other published method specifically aimed at melt curve novelty detection [26] aligns the melt curves to one specific temperature, losing the useful melt peak location information in the process. We have selected the most widely used ML methods in HRM analysis and have shown that they are all able to classify our dHRM database with very high accuracy. Interestingly, some drastically outperform others when it comes to novelty detection. We find that Neural Nets, SVMs, and Random Forest outperform the other ML methods, even after calibrating the probabilities. The SVM with the radial basis function (rbf) as a kernel outperforms the SVM with a linear kernel, which shows that our data cannot easily be separated by linear hyper-planes.

Random forests have been shown to perform well on classification tasks involving time series data sets [30] and are known to be an efficient way of performing outlier detection in high-dimensional datasets [80]. Here, we show that its well-calibrated probabilities are also particularly useful for conducting HRM novelty detection. One of the limitations of this work is that we used the most basic versions of each ML method and chose to not further optimize parameters for each one. Optimizing parameters could result in further improved novelty detection performance. Further improvements can be expected with more complex algorithms. It has been shown that the simplest way to gain improvement on time series problems is to transform the data into an alternative data space where discriminatory features are more easily detected [51]. The COTE classifier [51], which stands for Collective of Transformation-Based Ensembles is a collective of ensembles of classifiers using different data transformations. The collective contains classifiers constructed in the time, frequency, change, and shapelet transformation domains. It has outperformed any other previously published time series classification algorithm [51] and is expected to perform well doing novelty detection too.

The performance of our approach was improved with the removal of data for the outlier species E. coli. One reason for the lower performance of E. coli compared to the other species could be that is has the lowest number of melt curves available (Table 3.1), which results in a smaller amount of training data available for the ML methods. We do not expect E. coli to inherently have more heterogeneity in its sequence compared to other species. Melt curve shape variance might be another contributor to its outlier behavior as it has the third most variance in shape (from the ten species) as measured by dynamic time warping (DTW), in our previous work [62].

No major differences were found between using the full length and short version of the melt curves, although for most methods the short version does outperform the full length, showing that there might not be any additional information in the low-temperature tail of the curve, and including it might even confound the novelty detection performance.

There are several advantages of probabilistic classification methods over non-probabilistic classifiers. First, unlike probabilistic classification models, with non-probabilistic models, uncertainty in classification cannot be modeled. This is important in high-stake situations, such as clinical disease diagnosis, where decisions rely strongly on the certainty of the classification [83]. As interest in the utilization of HRM technology in the clinical set-ting grows, probabilistic classification methods will likely also become increasingly necessary for these cost-sensitive situations, where the certainty of classification has to be sufficiently high for clinical management decisions to be made. Second, probabilistic classifiers can be more effectively combined with other classifiers, within a large machine learning framework [84].

In conclusion, advances in machine learning and 'big data' generation are opening up more opportunities for the advancement of HRM, which due to its speed, low cost and simplicity was already attractive. The opportunity to use HRM as a discovery tool as well as profiling technology will further advance HRM technology towards its application in research and clinical diagnostics.

3.6 Additional Figures

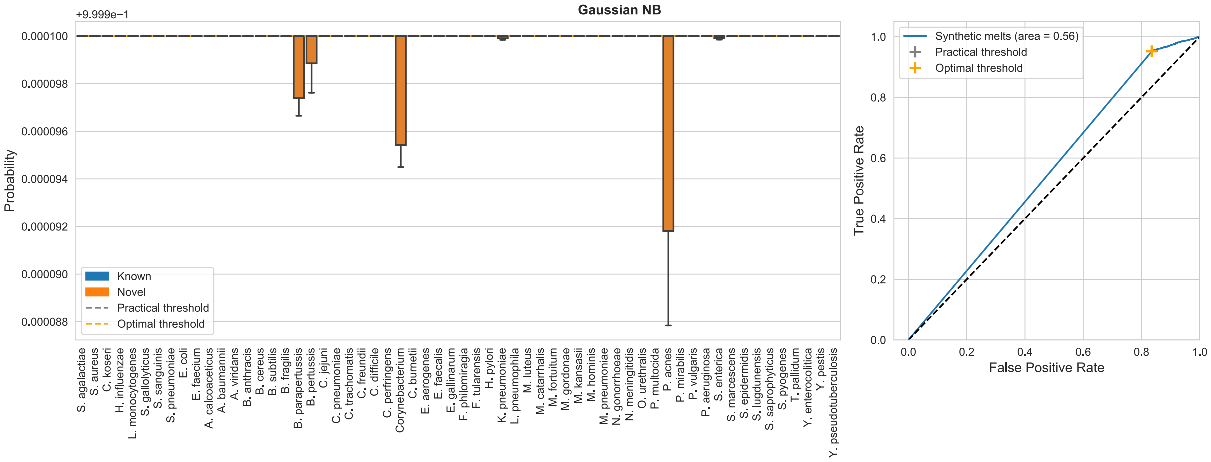


Figure 3.13: Gaussian NB.

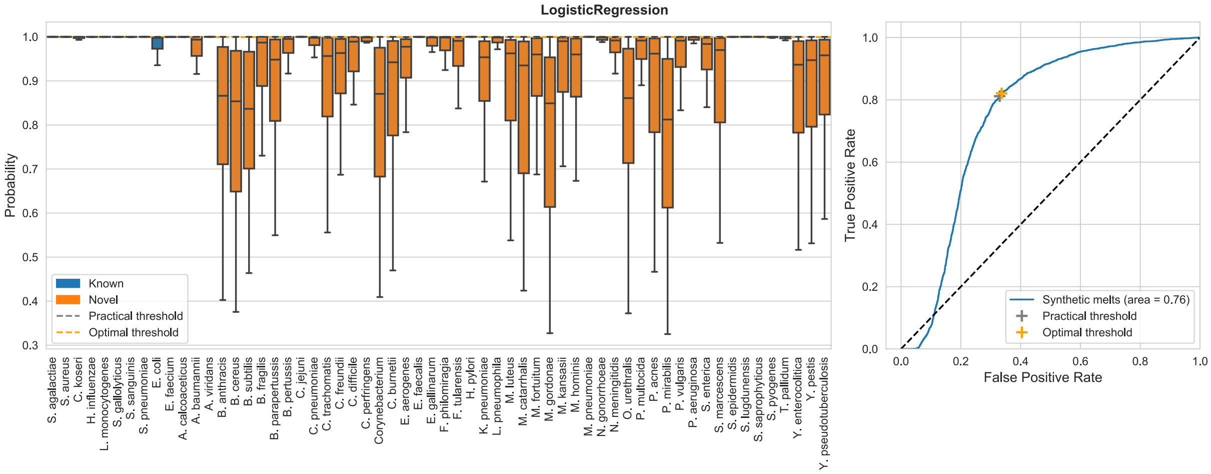


Figure 3.14: Logistic Regression.

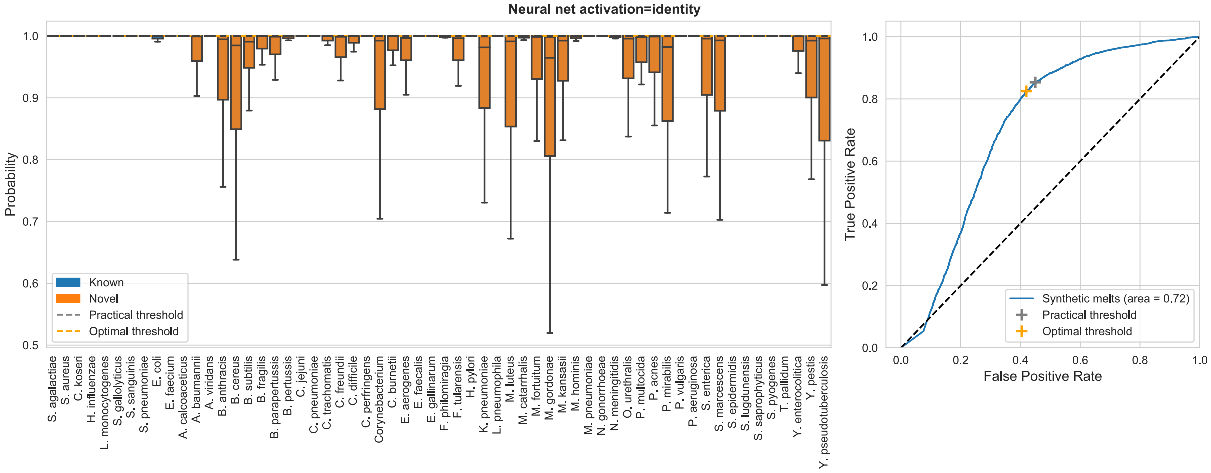


Figure 3.15: Neural net activation=identity.

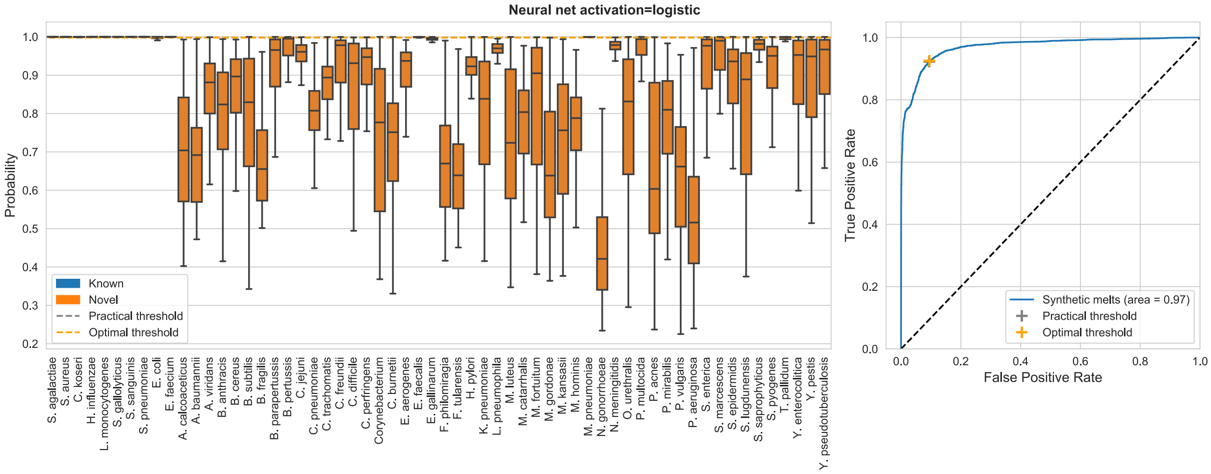


Figure 3.16: Neural net activation=logistic.

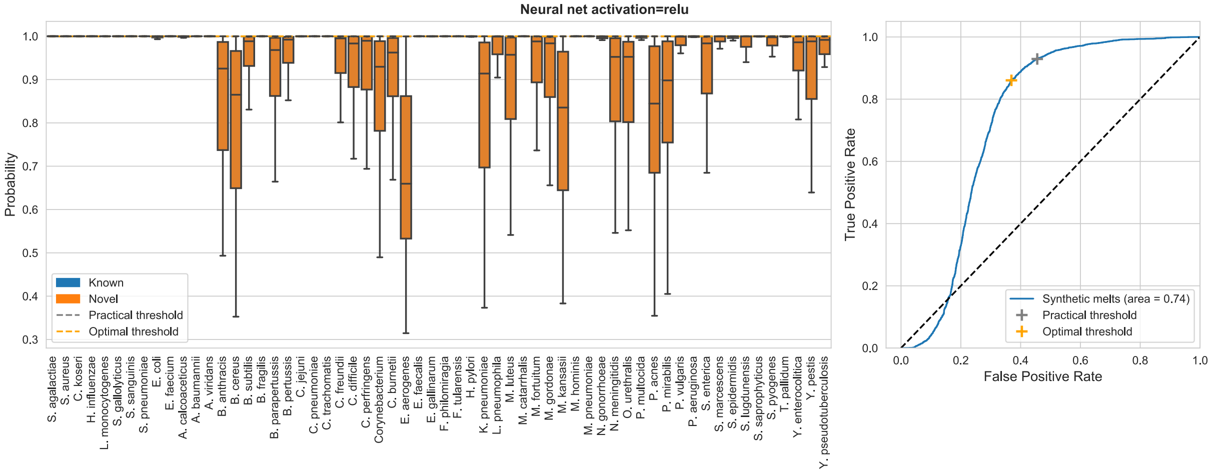


Figure 3.17: Neural net activation=relu.

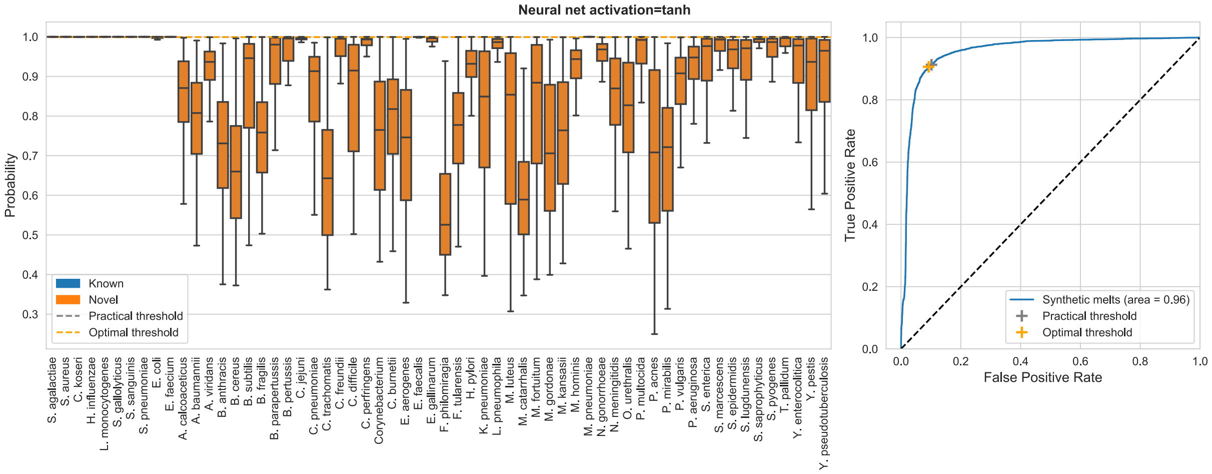


Figure 3.18: Neural net activation=tanh.

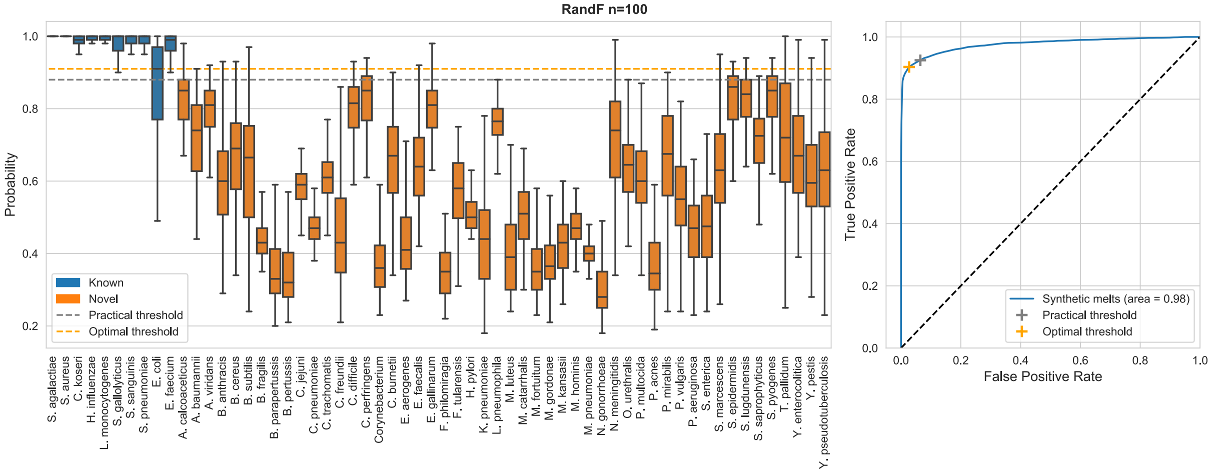


Figure 3.19: RandF n=100.

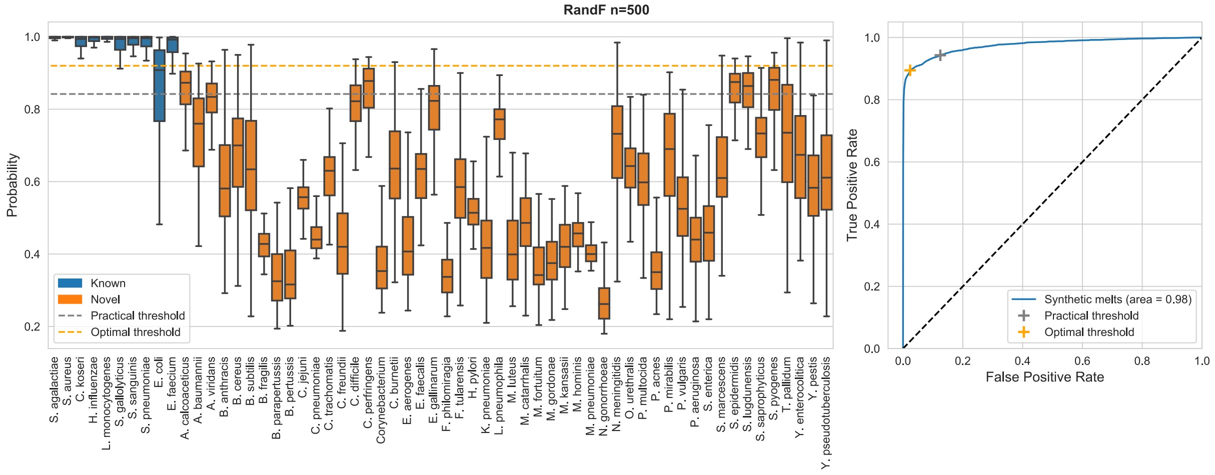


Figure 3.20: RandF n=500.

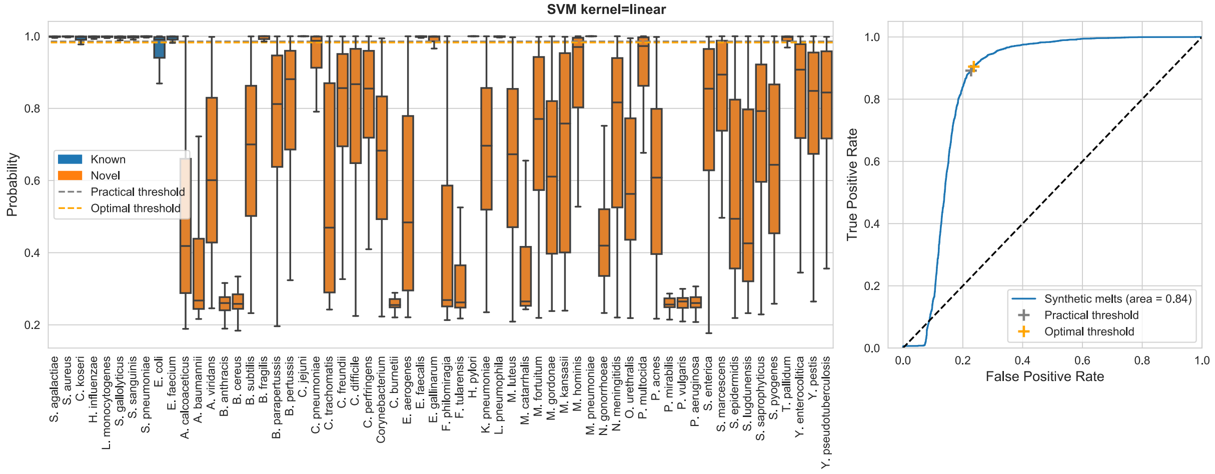


Figure 3.21: SVM kernel=linear.

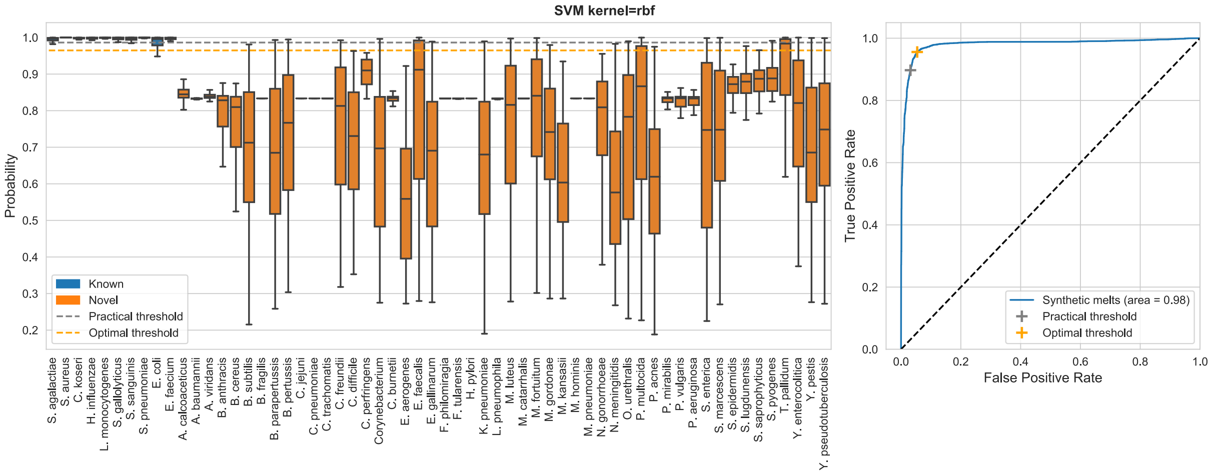


Figure 3.22: SVM kernel=rbf.

3.7 Acknowledgments

Chapter 3, in part, has been submitted for publication as it may appear in Bioinformatics Advances 2021. Lennart Langouche, Augustine C. Obirieze, Mridu Sinha, Hannah Mack, William Leineweber, April Aralar, David T. Pride, Todd P. Coleman, Stephanie I. Fraley. The dissertation author was the primary investigator and author of this material.

This work was supported by a grant from the NIH/NIAID, award number R01AI134982 and a Burroughs Wellcome Fund Career Award at the Scientific Interface to S.I.F., award number 1012027. Fellowship support was provided to A.C.O. by an NSF GRFP award under Grant number DGE-1650112.

Chapter 4

Conclusion

Rapid and precise molecular diagnostics are revolutionizing the clinical care of patients with suspected infections. As increasing numbers of molecular diagnostic technologies emerge from research laboratories into the clinical domain, it is essential to assess their clinical impact and cost-effectiveness prior to adoption in the clinical setting. The goal of this PhD was to explore the limitations of the dHRM platform, which allows it to be compared to other emerging technologies, and directly affects its clinical potential.

First, we developed a computational framework for estimating the resolving power of dHRM technology for defined sequence profiling tasks. We found that, despite noise inherent to dHRM, single-nucleotide resolution can be achieved in certain cases. Having such a high resolving power means that dHRM can potentially be used to differentiate closely related species or screen for genotypic markers of antimicrobial resistance. Second, we presented an advancement in universal microbial high resolution melting (HRM) analysis that is capable of accomplishing both known genotype identification and novel genotype detection. Specifically, we achieved this novel surveillance functionality through probabilistic modeling of sequence-defined HRM curves, uniquely enabled by the large-scale melt curve datasets generated on our high-throughput digital HRM platform. Our hope is that in the future, the dHRM platform can translate into a near-point of care, cost-effective tool for infectious disease screening.

Bibliography

[1] K. F. Smith, M. Goldberg, S. Rosenthal, L. Carlson, J. Chen, C. Chen, and S. Ramachandran, "Global rise in human infectious disease outbreaks.," *Journal of the Royal Society, Interface*, vol. 11, p. 20140950, dec 2014.

[2] L. E. Huerta and T. W. Rice, "Pathologic difference between sepsis and bloodstream infections.," *The journal of applied laboratory medicine*, vol. 3, pp. 654–663, jan 2019.

[3] C. Rhee, R. Dantes, L. Epstein, D. J. Murphy, C. W. Seymour, T. J. Iwashyna, S. S. Kadri, D. C. Angus, R. L. Danner, A. E. Fiore, J. A. Jernigan, G. S. Martin, E. Septimus, D. K. Warren, A. Karcz, C. Chan, J. T. Menchaca, R. Wang, S. Gruber, M. Klompas, and C. P. E. Program, "Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009-2014.," *The Journal of the American Medical Association*, vol. 318, pp. 1241–1249, oct 2017.

[4] C. for Disease Control and C. Prevention, "Clinical information sepsis," aug 2018.

[5] C. for Disease Control and C. Prevention, "Sepsis surveillance toolkit," aug 2018.

[6] J. M. Miller, M. J. Binnicker, S. Campbell, K. C. Carroll, K. C. Chapin, P. H. Gilligan, M. D. Gonzalez, R. C. Jerris, S. C. Kehl, R. Patel, B. S. Pritt, S. S. Richter, B. Robinson-Dunn, J. D. Schwartzman, J. W. Snyder, S. Telford, E. S. Theel, R. B. Thomson, M. P. Weinstein, and J. D. Yao, "A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2018 update by the infectious diseases society of america and the american society for microbiology.," *Clinical Infectious Diseases*, vol. 67, pp. e1–e94, aug 2018.

[7] E. J. Baron, J. M. Miller, M. P. Weinstein, S. S. Richter, P. H. Gilligan, R. B. Thomson, P. Bourbeau, K. C. Carroll, S. C. Kehl, W. M. Dunne, B. Robinson-Dunn, J. D. Schwartzman, K. C. Chapin, J. W. Snyder, B. A. Forbes, R. Patel, J. E. Rosenblatt, and B. S. Pritt, "A guide to utilization of the microbiology laboratory for diagnosis of infectious diseases: 2013 recommendations by the infectious diseases society of america (IDSA) and the american society for microbiology (ASM)(a).," *Clinical Infectious Diseases*, vol. 57, pp. e22–e121, aug 2013.

[8] K. Messacar, S. K. Parker, J. K. Todd, and S. R. Dominguez, “Implementation of rapid molecular infectious disease diagnostics: the role of diagnostic and antimicrobial stewardship.,” *Journal of Clinical Microbiology*, vol. 55, no. 3, pp. 715–723, 2017.

[9] A. J. Blaschke, C. Heyrend, C. L. Byington, M. A. Fisher, E. Barker, N. F. Garrone, S. A. Thatcher, A. T. Pavia, T. Barney, G. D. Alger, J. A. Daly, K. M. Ririe, I. Ota, and M. A. Poritz, “Rapid identification of pathogens from positive blood cultures by multiplex polymerase chain reaction using the FilmArray system.,” *Diagnostic Microbiology and Infectious Disease*, vol. 74, pp. 349–355, dec 2012.

[10] M. A. Poritz, A. J. Blaschke, C. L. Byington, L. Meyers, K. Nilsson, D. E. Jones, S. A. Thatcher, T. Robbins, B. Lingenfelter, E. Amiott, A. Herbener, J. Daly, S. F. Dobrowolski, D. H.-F. Teng, and K. M. Ririe, “FilmArray, an automated nested multiplex PCR system for multi-pathogen detection: development and application to respiratory tract infection.,” *Plos One*, vol. 6, p. e26047, oct 2011.

[11] S. N. Buss, A. Leber, K. Chapin, P. D. Fey, M. J. Bankowski, M. K. Jones, M. Rogatcheva, K. J. Kanack, and K. M. Bourzac, “Multicenter evaluation of the BioFire FilmArray gastrointestinal panel for etiologic diagnosis of infectious gastroenteritis.,” *Journal of Clinical Microbiology*, vol. 53, pp. 915–925, mar 2015.

[12] A. Rohatgi, C. Wattal, N. Goel, Y. Sagvekar, and N. Pandita, “Evaluation of BioFire film array meningitis/encephalitis (ME) panel in testing of CSF specimen of patients of meningoencephalitis in indian scenario (p4. 9-025),” 2019.

[13] E. C. Claas, C.-A. D. Burnham, T. Mazzulli, K. Templeton, and F. Topin, “Performance of the xTAG® gastrointestinal pathogen panel, a multiplex molecular assay for simultaneous detection of bacterial, viral, and parasitic causes of infectious gastroenteritis.,” *Journal of Microbiology and Biotechnology*, vol. 23, no. 7, pp. 1041–1045, 2013.

[14] B. W. Buchan, C. C. Ginocchio, R. Manii, R. Cavagnolo, P. Pancholi, L. Swyers, R. B. Thomson, C. Anderson, K. Kaul, and N. A. Ledeboer, “Multiplex identification of gram-positive bacteria and resistance determinants directly from positive blood culture broths: evaluation of an automated microarray-based nucleic acid test.,” *PLoS Medicine*, vol. 10, p. e1001478, jul 2013.

[15] H. Stender, “PNA FISH: an intelligent stain for rapid diagnosis of infectious diseases.,” *Expert Review of Molecular Diagnostics*, vol. 3, pp. 649–655, sep 2003.

[16] E. Mylonakis, C. J. Clancy, L. Ostrosky-Zeichner, K. W. Garey, G. J. Alangaden, J. A. Vazquez, J. S. Groeger, M. A. Judson, Y.-M. Vinagre, S. O. Heard, F. N. Zervou, I. M. Zacharioudakis, D. P. Kontoyiannis, and P. G. Pappas, “T2 magnetic resonance assay for the rapid diagnosis of candidemia in whole blood: a clinical trial.,” *Clinical Infectious Diseases*, vol. 60, pp. 892–899, mar 2015.

[17] P. Seng, M. Drancourt, F. Gouriet, B. La Scola, P.-E. Fournier, J. M. Rolain, and D. Raoult, "Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry.," *Clinical Infectious Diseases*, vol. 49, pp. 543–551, aug 2009.

[18] S. N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, J. Bouquet, A. L. Greninger, K.-C. Luk, B. Enge, D. A. Wadford, S. L. Messenger, G. L. Genrich, K. Pellegrino, G. Grard, E. Leroy, B. S. Schneider, J. N. Fair, M. A. Martínez, P. Isa, J. A. Crump, J. L. DeRisi, T. Sittler, J. Hackett, S. Miller, and C. Y. Chiu, "A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples.," *Genome Research*, vol. 24, pp. 1180–1192, jul 2014.

[19] P. Ramanan, A. L. Bryson, M. J. Binnicker, B. S. Pritt, and R. Patel, "Syndromic panel-based testing in clinical microbiology.," *Clinical Microbiology Reviews*, vol. 31, no. 1, 2018.

[20] R. Schlaberg, C. Y. Chiu, S. Miller, G. W. Procop, G. Weinstock, P. P. Committee, C. on Laboratory Practices of the American Society for Microbiology, and M. R. C. of the College of American Pathologists, "Validation of metagenomic next-generation sequencing tests for universal pathogen detection.," *Archives of Pathology & Laboratory Medicine*, vol. 141, pp. 776–786, jun 2017.

[21] K. M. Ririe, R. P. Rasmussen, and C. T. Wittwer, "Product differentiation by analysis of DNA melting curves during the polymerase chain reaction.," *Analytical Biochemistry*, vol. 245, pp. 154–160, feb 1997.

[22] M. Sinha, J. Jupe, H. Mack, T. P. Coleman, S. M. Lawrence, and S. I. Fraley, "Emerging technologies for molecular diagnosis of sepsis.," *Clinical Microbiology Reviews*, vol. 31, apr 2018.

[23] D. O. Velez, H. Mack, J. Jupe, S. Hawker, N. Kulkarni, B. Hedayatnia, Y. Zhang, S. Lawrence, and S. I. Fraley, "Massively parallel digital high resolution melt for rapid and absolutely quantitative sequence profiling.," *Scientific Reports*, vol. 7, p. 42326, feb 2017.

[24] P. Athamanolap, V. Parekh, S. I. Fraley, V. Agarwal, D. J. Shin, M. A. Jacobs, T.-H. Wang, and S. Yang, "Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants.," *Plos One*, vol. 9, p. e109094, oct 2014.

[25] S. I. Fraley, P. Athamanolap, B. J. Masek, J. Hardick, K. C. Carroll, Y.-H. Hsieh, R. E. Rothman, C. A. Gaydos, T.-H. Wang, and S. Yang, "Nested machine learning facilitates increased sequence content for large-scale automated high resolution melt genotyping.," *Scientific Reports*, vol. 6, p. 19218, jan 2016.

[26] N. Andini, B. Wang, P. Athamanolap, J. Hardick, B. J. Masek, S. Thair, A. Hu, G. Avornu, S. Peterson, S. Cogill, R. E. Rothman, K. C. Carroll, C. A. Gaydos, J. T.-H. Wang, S. Batzoglou, and S. Yang, "Microbial typing by machine learned DNA melt signatures.," *Scientific Reports*, vol. 7, p. 42097, feb 2017.

[27] S. I. Fraley, J. Hardick, B. J. Masek, P. Athamanolap, R. E. Rothman, C. A. Gaydos, K. C. Carroll, T. Wakefield, T.-H. Wang, and S. Yang, “Universal digital high-resolution melt: a novel approach to broad-based profiling of heterogeneous biological samples.,” *Nucleic Acids Research*, vol. 41, p. e175, oct 2013.

[28] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE transactions on acoustics, speech, and signal processing*, vol. 26, pp. 43–49, feb 1978.

[29] S. Lu, G. Mirchevska, S. S. Phatak, D. Li, J. Luka, R. A. Calderone, and W. A. Fonzi, “Dynamic time warping assessment of high-resolution melt curves provides a robust metric for fungal identification.,” *Plos One*, vol. 12, p. e0173320, mar 2017.

[30] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh, “The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances.,” *Data mining and knowledge discovery*, vol. 31, no. 3, pp. 606–660, 2017.

[31] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, “The UCR time series classification archive,” jul 2015.

[32] C. A. Ratanamahatana and E. Keogh, “Three myths about dynamic time warping data mining,” in *Proceedings of the 2005 SIAM International Conference on Data Mining* (H. Kargupta, J. Srivastava, C. Kamath, and A. Goodman, eds.), (Philadelphia, PA), pp. 506–510, Society for Industrial and Applied Mathematics, apr 2005.

[33] L. Breiman, “Random forests,” *Machine learning*, 2001.

[34] J. J. Rodríguez, L. I. Kuncheva, and C. J. Alonso, “Rotation forest: A new classifier ensemble method.,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1619–1630, oct 2006.

[35] F. V. Pedregosa, G. Gramfort, A. Michel, V. Thirion, B. Grisel, O. Blondel, M. Prettenhofer, P. Weiss, R. Dubourg, V. Vanderplas, J. Passos, A. Cournapeau, D. Brucher, M. Perrot, and M. d. Duchesnay, “Scikit-learn: Machine learning in python,” *The Journal of Machine Learning Research*, jan 2011.

[36] W. Meert, K. Hendrickx, and T. V. Craenendonck, “wannesm/dtaidistance v2.0.0,” *Zenodo*, 2020.

[37] D. Silva, G. Batista, and E. Keogh, “Prefix and suffix invariant dynamic time warping,” *2016 IEEE 16th International . . .* , 2016.

[38] A. Bagnall, M. Flynn, J. Large, J. Line, and A. Bostrom, “Is rotation forest the best classifier for problems with continuous features?,” *arXiv preprint arXiv . . .* , 2018.

[39] M. Liew, R. Pryor, R. Palais, C. Meadows, M. Erali, E. Lyon, and C. Wittwer, “Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons.,” *Clinical Chemistry*, vol. 50, pp. 1156–1164, jul 2004.

[40] C. T. Wittwer, G. H. Reed, C. N. Gundry, J. G. Vandersteen, and R. J. Pryor, “High-resolution genotyping by amplicon melting analysis using LCGreen.,” *Clinical Chemistry*, vol. 49, pp. 853–860, jun 2003.

[41] S. Yang, P. Ramachandran, R. Rothman, Y.-H. Hsieh, A. Hardick, H. Won, A. Kecojevic, J. Jackman, and C. Gaydos, “Rapid identification of biothreat and other clinically relevant bacterial species by use of universal PCR coupled with high-resolution melting analysis.,” *Journal of Clinical Microbiology*, vol. 47, pp. 2252–2255, jul 2009.

[42] J. Hester, “primerTree: Visually assessing the specificity and informativeness of primer pairs,” 2020.

[43] J. Ye, G. Coulouris, I. Zaretskaya, I. Cutcutache, S. Rozen, and T. L. Madden, “Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.,” *BMC Bioinformatics*, vol. 13, p. 134, jun 2012.

[44] Z. Dwight, R. Palais, and C. T. Wittwer, “uMELT: prediction of high-resolution melting curves and dynamic melting profiles of PCR products in a rich web application.,” *Bioinformatics*, vol. 27, pp. 1019–1020, apr 2011.

[45] R. D. Blake, J. W. Bizzaro, J. D. Blake, G. R. Day, S. G. Delcourt, J. Knowles, K. A. Marx, and J. SantaLucia, “Statistical mechanical simulation of polymeric DNA melting with MELTSIM.,” *Bioinformatics*, vol. 15, pp. 370–375, may 1999.

[46] O. Gotoh, “Prediction of melting profiles and local helix stability for sequenced DNA,” *Advances in biophysics*, 1983.

[47] N. R. Markham and M. Zuker, “DINAMelt web server for nucleic acid melting prediction.,” *Nucleic Acids Research*, vol. 33, pp. W577–81, jul 2005.

[48] G. Steger, “Thermal denaturation of double-stranded nucleic acids: prediction of temperatures critical for gradient gel electrophoresis and polymerase chain reaction.,” *Nucleic Acids Research*, vol. 22, pp. 2760–2768, jul 1994.

[49] J. SantaLucia, “A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, pp. 1460–1465, feb 1998.

[50] C. Ani, S. Farshidpanah, A. Bellinghausen Stewart, and H. B. Nguyen, “Variations in organism-specific severe sepsis mortality in the united states: 1999-2008.,” *Critical Care Medicine*, vol. 43, pp. 65–77, jan 2015.

[51] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, “Time-series classification with COTE: The collective of transformation-based ensembles,” *IEEE transactions on knowledge and data engineering*, vol. 27, pp. 2522–2535, sep 2015.

[52] Y. T. Hwang, H. J. Zuccola, Q. Lu, and C. B. C. Hwang, “A point mutation within conserved region VI of herpes simplex virus type 1 DNA polymerase confers altered drug sensitivity and enhances replication fidelity.,” *Journal of Virology*, vol. 78, pp. 650–657, jan 2004.

[53] R. A. Sturm, D. L. Duffy, Z. Z. Zhao, F. P. N. Leite, M. S. Stark, N. K. Hayward, N. G. Martin, and G. W. Montgomery, “A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color.,” *American Journal of Human Genetics*, vol. 82, pp. 424–431, feb 2008.

[54] M. Lecuit and M. Eloit, “The potential of whole genome NGS for infectious disease diagnosis.,” *Expert Review of Molecular Diagnostics*, vol. 15, pp. 1517–1519, nov 2015.

[55] G. H. Reed and C. T. Wittwer, “Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis.,” *Clinical Chemistry*, vol. 50, pp. 1748–1754, oct 2004.

[56] P. Bidet, S. Liguori, C. Plainvert, S. Bonacorsi, C. Courroux, C. d’Humières, C. Poyart, A. Efstratiou, and E. Bingen, “Identification of group a streptococcal emm types commonly associated with invasive infections and antimicrobial resistance by the use of multiplex PCR and high-resolution melting analysis.,” *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 31, pp. 2817–2826, oct 2012.

[57] A. L. Roth and N. D. Hanson, “Rapid detection and statistical differentiation of KPC gene variants in gram-negative pathogens by use of high-resolution melting and ScreenClust analyses.,” *Journal of Clinical Microbiology*, vol. 51, pp. 61–65, jan 2013.

[58] M. R. Zianni, M. R. Nikbakhtzadeh, B. T. Jackson, J. Panescu, and W. A. Foster, “Rapid discrimination between anopheles gambiae s.s. and anopheles arabiensis by high-resolution melt (HRM) analysis.,” *Journal of Biomolecular Techniques*, vol. 24, pp. 1–7, apr 2013.

[59] B. S. Pritt, P. S. Mead, D. K. H. Johnson, D. F. Neitzel, L. B. Respicio-Kingry, J. P. Davis, E. Schiffman, L. M. Sloan, M. E. Schriefer, A. J. Replogle, S. M. Paskewitz, J. A. Ray, J. Bjork, C. R. Steward, A. Deedon, X. Lee, L. C. Kingry, T. K. Miller, M. A. Feist, E. S. Theel, R. Patel, C. L. Irish, and J. M. Petersen, “Identification of a novel pathogenic borrelia species causing lyme borreliosis with unusually high spirochaetaemia: a descriptive study.,” *The Lancet Infectious Diseases*, vol. 16, pp. 556–564, feb 2016.

[60] J.-C. Cheng, C.-L. Huang, C.-C. Lin, C.-C. Chen, Y.-C. Chang, S.-S. Chang, and C.-P. Tseng, “Rapid detection and identification of clinically important bacteria by high-resolution melting analysis after broad-range ribosomal RNA real-time PCR.,” *Clinical Chemistry*, vol. 52, pp. 1997–2004, nov 2006.

[61] S. Bowman, D. McNevin, S. J. Venables, P. Roffey, A. Richardson, and M. E. Gahan, “Species identification using high resolution melting (HRM) analysis with random forest classification,” *Australian Journal of Forensic Sciences*, pp. 1–16, apr 2017.

[62] L. Langouche, A. Aralar, M. Sinha, S. M. Lawrence, S. I. Fraley, and T. P. Coleman, "Data-driven noise modeling of digital DNA melting analysis enables prediction of sequence discriminating power.," *Bioinformatics*, dec 2020.

[63] J. Adelman, W. McKay, J. Lillis, and K. Lawson, "High-resolution melt curve classification using neural networks,"

[64] P. Athamanolap, K. Hsieh, C. M. O'Keefe, Y. Zhang, S. Yang, and T.-H. Wang, "Nanoarray digital polymerase chain reaction with high-resolution melt for enabling broad bacteria identification and pheno-molecular antimicrobial susceptibility test.," *Analytical Chemistry*, vol. 91, pp. 12784–12792, oct 2019.

[65] J. C. Rolando, E. Jue, J. T. Barlow, and R. F. Ismagilov, "Real-time kinetics and high-resolution melt curves in single-molecule digital LAMP to differentiate and study specific and non-specific amplification.," *Nucleic Acids Research*, vol. 48, p. e42, apr 2020.

[66] A. Aralar, Y. Yuan, K. Chen, Y. Geng, D. Ortiz Velez, M. Sinha, S. M. Lawrence, and S. I. Fraley, "Improving quantitative power in digital PCR through digital high-resolution melting.," *Journal of Clinical Microbiology*, vol. 58, may 2020.

[67] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal processing*, vol. 99, pp. 215–249, jun 2014.

[68] G. Klinger, I. Levy, L. Sirota, V. Boyko, B. Reichman, L. Lerner-Geva, and I. N. Network, "Epidemiology and risk factors for early onset sepsis among very-low-birthweight infants.," *American Journal of Obstetrics and Gynecology*, vol. 201, pp. 38.e1–6, jul 2009.

[69] B. J. Stoll, N. I. Hansen, P. J. Sánchez, R. G. Faix, B. B. Poindexter, K. P. Van Meurs, M. J. Bizzarro, R. N. Goldberg, I. D. Frantz, E. C. Hale, S. Shankaran, K. Kennedy, W. A. Carlo, K. L. Watterberg, E. F. Bell, M. C. Walsh, K. Schibler, A. R. Laptook, A. L. Shane, S. J. Schrag, A. Das, R. D. Higgins, E. K. S. N. I. of Child Health, and H. D. N. R. Network, "Early onset neonatal sepsis: the burden of group b streptococcal and e. coli disease continues.," *Pediatrics*, vol. 127, pp. 817–826, may 2011.

[70] R. Palais and C. T. Wittwer, "Mathematical algorithms for high-resolution DNA melting analysis.," *Methods in Enzymology*, vol. 454, pp. 323–343, 2009.

[71] A. Pant, "Introduction to logistic regression (https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148)."

[72] W. Monroe, "Logistic regression (https://web.stanford.edu/class/archive/cs/cs109-/cs109.1178/lectureHandouts/220-logistic-regression.pdf)," aug 2017.

[73] S. Yildirim, "Naive bayes classifier - explained (https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed)."

[74] G. Chauhan, “All about naive bayes (https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf).”

[75] R. Pupale, “Support vector machines (SVM) - an overview (https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989).”

[76] “Neural network models (https://scikit-learn.org/stable/modules/neural-networks-supervised.html).”

[77] T. Yiu, “Understanding random forest (https://towardsdatascience.com/understanding-random-forest-58381e0602d2).”

[78] J. Platt, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, 1999.

[79] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” *… of the 22nd international conference on …*, 2005.

[80] F. Liu, K. Ting, and Z. Zhou, “Isolation forest,” *2008 eighth ieee international …*, 2008.

[81] P. Athamanolap and K. Hsieh, “Integrated bacterial identification and antimicrobial susceptibility testing for polymicrobial infections using digital PCR and digital high-resolution melt in a microfluidic …,” *2018 40th Annual …*, 2018.

[82] C. M. OrKeefe and T.-H. l. Wang, “Digital high-resolution melt platform for rapid and parallelized molecule-by-molecule genetic profiling.,” *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2018, pp. 5342–5345, jul 2018.

[83] J. Lafferty, A. McCallum, and F. Pereira, “Proceedings of the 18th international conference on machine learning,” 2001.

[84] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, pp. 226–239, mar 1998.