Development of a novel method for purity assessment of nucleic acid samples

Inaugural dissertation

for the attainment of the title of doctor
in the Faculty of Mathematics and Natural Sciences
at the Heinrich Heine University Düsseldorf

presented by

Conny Rosella Unger
from Kol. Neuland (Paraguay)

Düsseldorf, June 2019

from the Institute of Synthetic Microbiology
at the Heinrich-Heine-University of Düsseldorf

Published by permission of the
Faculty of Mathematics and Natural Sciences at
Heinrich Heine University Düsseldorf

Correspondents:

1. Jun.-Prof. Dr. Ilka Maria Axmann

2. Prof. Dr. Markus Kollmann

Date of the oral examination: November 15th 2019

Statement of authorship

I hereby declare that this dissertation is the result of my own work. No other person's work has been used without due acknowledgment. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Düsseldorf, June 4th 2019

Conny Unger

Summary

Standard methods for research, applied testing and molecular diagnostics, like quantitative Polymerase Chain Reaction (qPCR) or Next Generation Sequencing (NGS) require the application of high quality nucleic acids to obtain reliable and reproducible results. Nucleic acid quality is defined by concentration, integrity and purity. Methods for nucleic acid concentration and integrity determination have advanced over the past decades. In contrast, the current method for nucleic acid purity assessment, using A_{260}/A_{280} and A_{260}/A_{230} absorbance ratios, was first introduced in 1942 and its validity has been controversially discussed by the scientific community. In this study, a novel method for DNA purity assessment was developed using mathematical data modelling to predict the influence of possible DNA sample impurities on enzyme activities, by analyzing their absorbance spectra. After establishment of polymerase and ligase activity assays, influence of possible DNA sample impurities on absorbance spectra and enzyme activities was recorded and obtained data were used for algorithm training and testing. It was demonstrated, that a *K*-nearest-neighbor algorithm, assigning absorbance spectra into five classes, representing increasing DNA purity defined by measured enzyme activities, could predict DNA purity with higher accuracy compared to A_{260}/A_{280} and A_{260}/A_{230} absorbance ratios. Although the algorithm failed to correlate with qPCR outcome, this study shows that using mathematical data modelling to analyze absorbance spectra is a promising approach to develop a novel method for nucleic acid purity assessment.

Zusammenfassung

Qualitativ hochwertige Nukleinsäuren sind essentiell für den Erhalt vertrauenswürdiger und reproduzierbarer Ergebnisse bei der Anwendung von Standardmethoden in der molekularbiologischen Forschung, der Molekulardiagnostik und für angewandte Testverfahren, wie die quantitative Polymerase-Kettenreaktion (qPCR) oder Hochdurchsatzsequenzierung. Dabei wird die Qualität von Nukleinsäuren mittels Konzentration, Integrität und Reinheit beschrieben. Die Methoden zur Bestimmung von Nukleinsäure Konzentration und Integrität wurden in den letzten Jahrzehnten weiterentwickelt. Im Gegensatz dazu, wurde die aktuelle Methode zur Bestimmung der Reinheit einer Nukleinsäure Probe, basierend auf die A_{260}/A_{280} und A_{260}/A_{230} Absorptionsverhältnisse, schon 1942 eingeführt und ihre Aussagekraft wurde in der Wissenschaft über die Jahre kontrovers diskutiert. In dieser Arbeit wurde eine neue Methode zur Bestimmung der Reinheit von DNA entwickelt. Sie basiert auf die Verwendung mathematischer Modelle, die anhand von Absorptionsspektren den Einfluss möglicher Unreinheiten in DNA Proben auf Enzymaktivitäten vorhersagen. Dafür wurden nach der Etablierung von Polymerase und Ligase Aktivitätstest der Einfluss von möglichen Unreinheiten in DNA Proben auf Absorptionsspektren und Enzymaktivitäten gemessen und die erhaltenen Daten wurden angewandt um verschiedene Algorithmen zu trainieren und zu testen. Es wurde gezeigt, dass die *K*-nearest-neigbor Klassifikation, die die Absorptionsspektren in fünf Klassen unterteilte, die unterschiedliche DNA Reinheitsgrade basierend auf gemessene Enzymaktivitäten beschrieben, die Reinheit einer DNA Probe mit einer höheren Genauigkeit vorhersagen konnte als die A_{260}/A_{280} und A_{260}/A_{230} Absorptionsverhältnisse. Obwohl ein Zusammenhang zwischen den Ergebnissen der *K*-nearest-neigbor Klassifikation mit den Ergebnissen einer qPCR nicht nachgewiesen werden konnte, zeigt diese Arbeit, dass die Anwendung mathematischer Modelle zur Analyse von Absorptionsspektren eine vielversprechende Herangehensweise ist, um eine neue Methode zur Bestimmung der Reinheit von Nukleinsäure Proben zu entwickeln.

Table of Contents

Abbreviations

PhiX DNA	PhiX 174 DNA
DTT	dithiothreitol
dNTPs	deoxy-nucleoside triphosphate
EDTA	ethylenediaminetetraacetic acid
Hb	human hemoglobin
HSA	human serum albumin
IgG	Immunoglobulin G
SA	sodium azide
GITC	guanidine isothiocyanate
SC	sodium citrate dihydrate
% NA	percentage normalized area
PCR	polymerase chain reaction
qPCR	quantitative PCR
NGS	next generation sequencing
MOPS	3-(N-morpholino)propanesulfonic acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
nzv	near zero variance
PCA	principal component analysis
MLR	multiclass logistic regression
KNN	*K*-nearest-neighbor
Cq	quantification cycle
Taq	*Thermus aquaticus*
E. coli	*Escherichia coli*
T4 PNK	T4 Polynucleotide Kinase
ATP	adenosine triphosphate
tRNA	transfer RNA
DNA	deoxyribonucleic acid
RNA	ribonucleic acid
OD	optical density
RIN	RNA integrity number
mRNA	messenger RNA
rRNA	ribosomal RNA
RNases	Ribonucleases
UV/Vis	ultraviolet visible
SCP	spectral content profiling
dPCR	digital PCR

1 Introduction

1.1 Nucleic acid purity in molecular biological methods

Using high quality nucleic acids is critical for the successful application of modern molecular biological methods, like quantitative Polymerase Chain Reaction (qPCR) or Next Generation Sequencing (NGS), that have become standard methods for research, applied testing and molecular diagnostics [1]–[4]. Therefore, quality control of nucleic acids has become increasingly important in recent years. Efforts have been made to identify sample quality metrics and improve quality and reproducibility of results of labor intensive, time-consuming, and highly expensive downstream applications [3], [5]. As studies have shown, inaccurate or irreproducible results can be caused by impurities and contaminations from starting material or cleanup procedure of nucleic acid samples [6]–[12].

Common impurities in nucleic acid samples include proteins, such as Immunoglobulin G (IgG) and hemoglobin from sample material, or phenol, proteins, and salts from cleanup procedure [8]. These sample impurities inhibit or hamper downstream applications in a concentration dependent manner through different mechanisms [8], [13]. IgG from sample material for example, binds to single-stranded DNA inhibiting amplification thereof by polymerases, whereas hemoglobin affects polymerase activity by reacting with its cofactor and quenching the fluorescence signal in qPCR [8], [14]. Proteins, like proteases, from nucleic acid purification chemistry inhibit downstream reactions by degradation of applied enzymes. Phenol and salts, which are often used to remove nucleases during nucleic acid purification, lead to inhibition by denaturation of enzymes [6], [13], [15], [16].

Sample impurities in NGS workflow can cause poor library quality, which in turn leads to poor quality of sequencing results [17]. In qPCR, complete inhibition of polymerase by sample impurities or insufficient nucleic acid template lead to failed amplification and consequently to false negative results or no detection of a PCR product [3], [18] Partial inhibition of polymerase in qPCR or interaction of impurities with DNA template can lead to false results in form of quantification cycle (Cq) shifts. The Cq is used for relative quantification of PCR product and describes the qPCR cycle at which the fluorescence signal of a sample reaches a threshold. If less PCR product is detected, due to less template input in qPCR reaction or partial inhibition of polymerase, the threshold is reached in a later cycle and the Cq value is higher. For relative quantification within one qPCR run, the Cq value of an unknown sample is compared to the Cq value of a control sample; if Cq values differ from each other, a Cq shift or delta Cq values greater or smaller zero are observed [19].

In summary, impurities in nucleic acid samples can interact with DNA template applied in downstream applications or inhibit enzymes catalyzing different steps of molecular biological methods.

1.2 Enzymes in molecular biological applications

Many standard molecular biology methods like PCR and NGS are based on the amplification of nucleic acids by polymerases. Polymerases are enzymes that catalyze template-directed synthesis of DNA and are essential in all organisms for DNA replication and repair. The DNA polymerase was first discovered in 1955 [20], [21]. The isolation of thermostable DNA polymerase from thermophilic bacterium *Thermus aquaticus* (Taq polymerase) [22], the introduction of PCR by Kary Mullis [23] and further development to quantitative real-time PCR (qPCR), enabling real-time monitoring of amplification and more accurate quantification of a nucleic acid sequence in a sample, have led to a revolution of biological research and molecular diagnostics [7], [23], [24].

DNA Ligases are nucleotidyltransferases (NTases) and catalyze the formation of a phosphodiester by joining the 3'-hydroxyl and 5'-phosphate ends of DNA fragments. They are widely used in cloning assays as well as in the preparation steps of NGS libraries. Ligases are essential for DNA repair mechanisms, such as single strand breaks, and replication, to join Okazaki fragments [25]–[28]. DNA ligases were first discovered nearly simultaneously by five independent laboratories in 1967 in uninfected and T4 bacteriophage infected *Escherichia coli* (*E. coli*) [29]–[33]. Since their purification, ligation using the T4 DNA ligase for instance, has become an important tool for the development of molecular cloning and many molecular biology methods for nucleic acid editing *in vivo* and *in vitro* based on generation of recombinant DNA by ligation of two different DNA fragments [34], [35]. Besides standard methods like NGS, many less known methods like Ligase Chain Reaction (LCR) [36], [37] or Multiplex Ligation-dependent Probe Amplification (MLPA) [38] depend on ligase reaction.

Ligases are often applied in combination with a kinase. Kinases belong to the enzyme class of phosphotransferases and catalyze the transfer of a phosphate from a donor, usually ATP, to a substrate [39]. The T4 Polynucleotide Kinase (T4 PNK) is a Nucleotidekinase and transfers the γ-Phosphate from ATP to the 5'-OH termini of a nucleic acids [40]. The T4 PNK was first extracted from *E. coli*, that was infected with a T4 bacteriophage [41]. In T4 bacteriophages, the T4 PNK serves to restore transfer RNAs (tRNA) degraded by host enzymes by phosphorylation of 5'-ends for following ligation by phage RNA ligase [42], [43].

The ability of T4 PNK to support DNA or RNA repair is also used for molecular biology applications [44]. Thereby T4 PNK is often applied in combination with ligases, for example in molecular cloning protocols and NGS, to assure phosphorylation of 5'-termini of nucleic acids fragments that are to be joined by a ligase [45], [46].

With the development of molecular biology methods based on nucleic acid amplification by polymerase, ligation, or phosphorylation by kinases, assays have been developed to test and quantify the activity of these enzymes. Polymerase, ligase and kinase activity was first measured using radioactive labeled nucleotides or phosphate [34], [45], [47]. This approach is time-consuming, discontinuous and includes safety and health hazards due to work with radioisotopes [48], [49]. Thus, over the past decades, numerous non-radioactive enzyme activity assays have been developed. Polymerase activity is now quantified by measuring signal of fluorescence dyes binding double-stranded DNA [48], [50] or fluorescence signal of a molecular beacon [51], instead of incorporation of radiolabeled nucleotides [47].

For ligases, radioisotope-free assays were suggested by replacing radioactive labeled phosphate, detected by denaturing gel electrophoresis or autoradiography [34], by a fluorophore [52], or using two DNA fragments with sticky ends, labeled either with a fluorophore or a quencher, resulting in quenching upon ligation [53]. Tang and colleagues developed a real-time assay to continuously monitor ligase reaction, using a molecular beacon (MB) [54]. The majority of alternative kinase activity assays for T4 PNK, avoiding radioactive labeled phosphate, are coupled to a second enzyme, such as a nucleotidase coupled assay, using malachite green for detection of free phosphate generated as side product from ATP [55]. Kleman-Leyer *et al.* described an antibody based fluorescence polarization method to detect ADP as a side product of T4 PNK reaction [56]. Furthermore, many different assays coupled to λ exonuclease cleavage using diverse detection methods have been suggested [57]–[59]. Interestingly, according to citation numbers and manufacturer websites, none of these more recent assays have been able to replace the radioactive based approach to measure ligase or T4 PNK activity as standard methods.

Enzyme activity assays are generally used to test different assay parameters such as concentration in enzyme, substrate, salt, etc. on the enzymatic reaction. They can however, also be used to quantify the inhibition of different compounds on tested enzymes, such as potential sample impurities contaminating biological samples eventually having an impact on PCR, NGS or other downstream assays.

1.3 Detection of impurities in nucleic acid samples

Performing enzyme activity assays for each sample before its application on PCR or NGS to detect possible contaminants in nucleic acid samples, would be too expensive, tedious and time-consuming. Some authors report or recommend using qPCR assays for detection of possible contaminants in nucleic acids [3], [9], [60], [61], but purity of nucleic acid samples is usually determined using A_{260}/A_{280} and A_{260}/A_{230} ratios of a UV/Vis absorbance measurement. In molecular biology laboratories, classical UV/Vis spectrometers are primarily used to determine the concentration of proteins or nucleic acid samples based on the Lambert-Beer law stating that the absorbance intensity of a substance is proportional to its concentration [62]. The absorbance maximum for nucleic acids is at 260 nm wavelength, and an OD value of 1 corresponds to 50 ng/µL dsDNA. Salts, proteins or phenol show absorbance peaks at 230, 280 or 230 and 270 nm, causing a deformation of the absorbance spectrum of nucleic acid samples when those contaminants are present (Ref. [63], [64] and Figure 1). A_{260}/A_{280} ratio of ~2.0 and A_{260}/A_{230} ratio in the range of 1.8 – 2.2 are assumed to indicate pure nucleic acids. Contamination of salts in nucleic acids lowers the A_{260}/A_{230} ratios, and protein contamination results in lower A_{260}/A_{280} ratios [64]–[66].

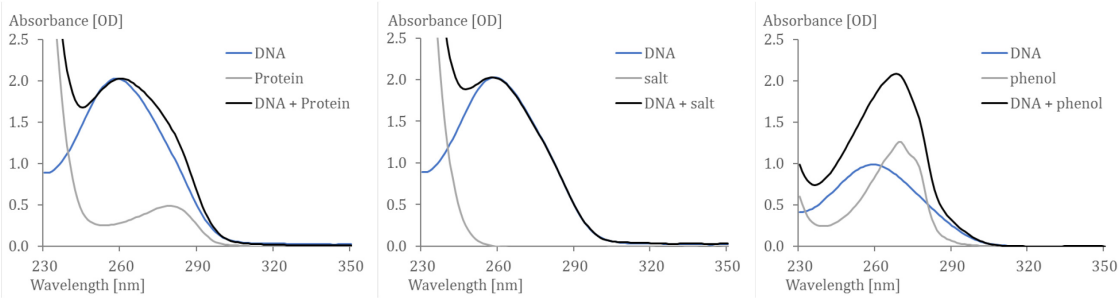


Figure 1: Absorbance spectra of DNA with (left) protein, (center) salt, and (right) phenol. Salts, proteins and phenol show absorbance peaks at 230, 280 or 230 and 270 nm, respectively, causing a deformation of DNA absorbance spectrum.

The A_{260}/A_{280} was first introduced by Warburg in 1942 to detect nucleic acid contamination in a protein solution [67] and its validity to assess nucleic acid purity has been controversially discussed in the scientific community over time [12], [62], [65], [68], [69]. However, no other method has been able to replace the ratios for detection of impurities in nucleic acid samples, even after several decades.

1.4 Application of mathematical modelling in biology

The application of complex mathematical data models or algorithms has become a powerful tool to get more information out of biological data. An example for the successful application

of algorithms to analyze available data to set new standards is the RNA Integrity Number (RIN).

RNA, as key molecules for protein biosynthesis, are important indicators for cell activities and cell stages, since cells respond to external or internal stimuli by translating information stored on their genome into proteins. Messenger RNA (mRNA), is the transcript of the genomic DNA transporting the nucleotide sequence that is translated into the amino acid sequence of the protein. Thus, the extraction and analysis of mRNA is an important tool in biomedical research to study changing gene expression associated with differentiation, transformation, or development of cells. Unfortunately, RNA molecules can also be quickly degraded by, elevated temperatures or ubiquitous Ribonucleases (RNases). Therefore, measurement of RNA integrity or fragmentation is essential before gene expression analysis to ensure integrity of RNA sequence of interest.

Although mRNA is usually the molecule of interest, total RNA is used to determine the integrity of extracted RNA, since about only 5% of total RNA consists of mRNA, while ribosomal RNA (rRNA), which embodies the main part of the ribosome that translates the mRNA into a protein, represents about 80% [70]. Ribosomal RNA consists of three sub-units defined by their molecular weight: the 5S, 18S and 28S RNA in eukaryotic cells. RNA integrity or fragmentation was traditionally determined by visual comparison of the 18S and 28S rRNA sub units, obtained after separation by traditional gel electrophoresis (Figure 2 left). This method is based on the observation that the 28S RNA is degraded faster than the 18S RNA, leading to a decreasing intensity ratio [71], [72]. Although authors of several studies have shown that the rRNA ratio does not always correlate with downstream assay success [2], [71], [73], this was the standard method for RNA integrity assessment for decades. The introduction of an electrophoresis system allowing digital data acquisition in the form of electropherograms in 1999 (Figure 2 right) allowed the development and training of an algorithm, developed to analyze electropherogram features, including areas before and between the 18S and 28S peak, to objectively assess integrity i.e. of RNA samples [74]. The RIN algorithm, provides users with an easy-to-use number ranging from one to ten that allows standardization of RNA samples' integrity both inter- and intra-laboratory and, has become the new standard for RNA integrity assessments and its validity has been shown by various authors [2], [3], [73], [75].

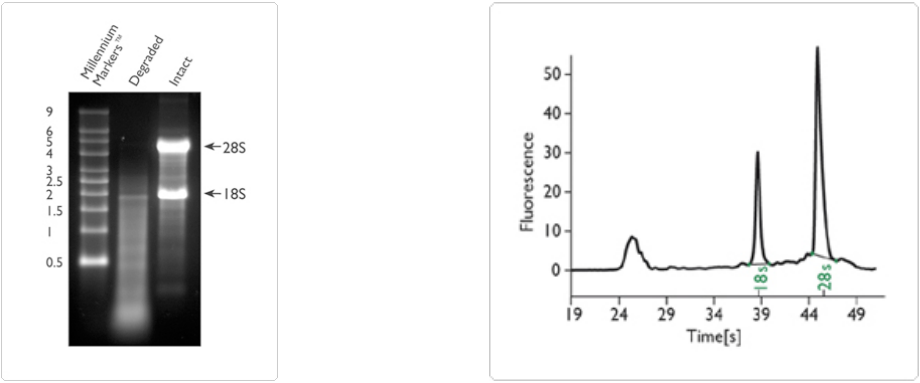


Figure 2: (left) Intact vs. Degraded RNA. (right) Agilent 2100 Bioanalyzer Data. (left) Two µg of degraded total RNA and intact total RNA were run beside Ambion's RNA Millennium Markers™ on a 1.5% denaturing agarose gel. The 18S and 28S ribosomal RNA bands are clearly visible in the intact RNA sample. The degraded RNA appears as a lower molecular weight smear. (right) Electropherogram of a high quality, eukaryotic, total RNA sample. The 18S and 28S peaks are clearly visible at 39 and 46 seconds, respectively. The microchannels of the Bioanalyzer are filled with a sieving polymer and fluorescence dye. Samples are detected by their fluorescence and translated into electropherograms or into gel-like images (data not shown). Reprinted from [76].

UV/Vis absorbance spectra of DNA samples have been recorded in a digital format for decades and over time instruments have been developed to use less sample volume or achieve accurate measurements with lower DNA concentrations. In 1999, Saurina and colleagues described a method to detect different nucleic acids components such as different nucleotides in a mixture by using multivariate curve resolution-alternating least squares (MCR-ALS) to analyze UV/Vis spectra between 230 and 350 nm wavelength [77]. Boonefaes and Luyssaert claimed a patent in 2011 for an approach to detect nucleic acids and other substances in complex mixtures by analysis of absorption spectra. The method was called Spectral Content Profiling (SCP) and consists of mathematically fitting reference spectra of possible components of a measured sample that absorb in the UV/Vis range, such as nucleic acids, salts, proteins, and phenol to the total measured UV/Vis spectrum [78], [79]. With resulting artificial spectra representing sample components and the Lambert-Beer law for mixtures, the abundance of these components in a sample can be quantified [80]. SCP was implemented as on board algorithm onto commercially available spectrophotometer, called Lunatic (Unchained Labs) and QIAxpert (QIAGEN). Although it has been shown that the results of SCP correlate with the amount of a contaminant added to a nucleic acid sample [12], a limitation of SCP is its inability to distinguish contaminants with similar absorbance spectra. Various salts or different proteins for example have very similar absorbance spectra, while their effect on downstream applications, on which nucleic acid samples are applied, vary significantly. Therefore, the quantity of detected impurities by SCP cannot be used to predict success of downstream assay or reliability of downstream assay results.

1.5 Aim of this thesis

The underlying hypothesis of this thesis was that, similar to the RIN, which categorizes RNA samples from one to ten according to their degradation level, absorbance spectra of nucleic acid samples could be categorized into several levels of purity using absorbance spectra. In addition, these purity levels should correlate with the success of different downstream applications. Therefore, the aim was to generate a novel method for purity assessment of nucleic acid samples based on UV/Vis absorbance spectra that correlates with quality of downstream assay results. Therefore, three steps were aimed:

I. Assays to measure activity of polymerase, ligase, and kinase should be established and used to investigate the influence of possible impurities on enzyme activity, by contaminating pure nucleic acid samples with known concentrations of possible impurities (Figure 3 A).

II. The resulting enzyme activities should be used as target value to develop a mathematical data model that would be able to predict enzyme activity based on the UV/Vis spectra or results of spectral content profiling of applied nucleic acid sample (Figure 3 A).

III. The best data model should be tested with data from real life samples applied to qPCR, to demonstrate the usefulness of the established algorithm (Figure 3 B).

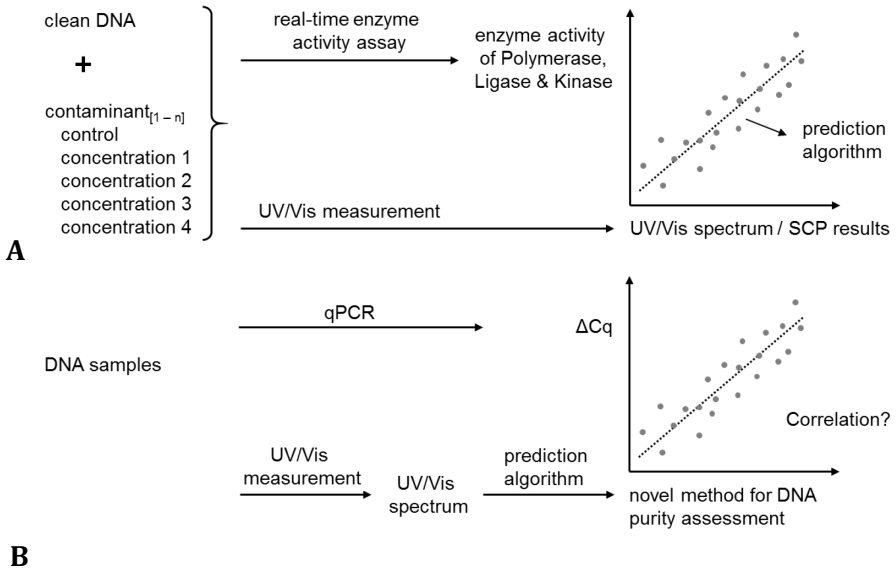


Figure 3: Schematic overview of strategy. (A) First, clean DNA should be spiked with defined contaminants and submitted to UV/Vis measurement and enzyme activity assays for polymerases, ligases and kinases. Second, results of step one should be used to generate a mathematical model able to predict enzyme activity based on UV/Vis measurement. (B) Third, real life DNA samples should be utilized to demonstrate usefulness of novel method for purity assessment of nucleic acid samples by comparing determined purity for each sample with delta Cq (ΔCq) values.

The novel method for purity assessment could enable researchers and clinicians to identify sample impurities before the nucleic acid sample is applied to a downstream assay and obtain more reliable results. Furthermore, it would allow inter- and intra-laboratory comparison and standardization of nucleic acid sample purity, and thus help to improve reproducibility in research.

2 Results

2.1 Selection of possible contaminants

As described in chapter 1.5, step one was to define possible contaminants or impurities, whose influence on UV/Vis absorbance spectra and enzyme activities were tested to collect data for algorithm development. Therefore, a pre-selection of substances found in cleanup or assay buffers and reagents was prepared and extended with further possible contaminants from sample origin or working environment.

2.1.1 Absorbing and non-absorbing contaminants

First, the absorbance in UV/Vis range was measured by recording UV/Vis spectra of possible contaminants diluted in RNase-free water on DropSense96 or QIAxpert, using the General UV/Vis or UV/Vis application. The absorbance spectra served to select final list of possible contaminants, since only contaminants absorbing light in UV/Vis range can be detected by absorbance measurement and consequently be used for development of novel method for purity assessment of nucleic acids based on absorbance spectra.

Figure 4 shows absorbance spectra of 12 discarded possible contaminants from 26 pre-selected substances from literature: 3-(N-morpholino)propanesulfonic acid (MOPS), sucrose, urea, glycerol, diluted isopropanol, guanidine hydrochloride (GuHCl), trizma base, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), β-mercaptoethanol, DNase 1, proteinase K, and RNase A. Chloroform and ethanol were part of pre-selected substances from literature and were not presented in Figure 4, although they were excluded from final list of possible contaminants, since they could not be loaded into the measurement chip due to their hydrophobicity. Chloroform is used in combination with phenol for denaturation of proteins during nucleic acid extraction and facilitates separation of aqueous and organic phase [81], while ethanol is used to precipitate nucleic acids [82].

MOPS, sucrose, urea, glycerol, isopropanol, and guanidine hydrochloride (GuHCl) showed no absorbance between 230 and 350 nm wavelength (Figure 4) and therefore were removed from final list of possible contaminants. MOPS is a buffering compound for near-neutral pH commonly used for gel electrophoresis [83], [84]. The sugar sucrose is added to cell lysis buffers to increase osmotic pressure outside cells, aiding cell rupture [85]. Cell lysis buffers for nucleic acid purification can also contain urea or GuHCl to

denature proteins [45], [86]. Isopropanol, like ethanol, is used to precipitate nucleic acids during purification [87].

The absorbance spectra of trizma base, HEPES, and β-mercaptoethanol had a peak at ≤ 230 nm, followed by a steep decline of absorbance, also called an A_{230} shoulder (Figure 4). Trizma base is a component of buffer solutions used for nucleic acid purification or gel electrophoresis, usually applied in concentrations between 10 and 100 mM in DNA samples [88]. Since a 1 M solution showed an absorbance maximum of only ~0.3 OD, it would not be detectable at working concentrations and was not added to the final list of possible contaminants. HEPES is widely used in as buffer system in cell culture or during nucleic acid purification to maintain a pH between 6.8 and 8.2 [83]. During RNA isolation, β-mercaptoethanol, showing high absorbance at low concentration, is often used for denaturation of ribonucleases by reducing their bisulfide bonds to prevent degradation of RNA [89]. However, this thesis focuses on DNA and therefore β-mercaptoethanol was not applied as possible DNA contaminant.

DNase 1, proteinase K, and RNase A, had an A230 shoulder and a second absorbance maximum at A280 (Figure 4). The proteins DNase I, Proteinase K and RNase A are used during cleanup procedure of RNA and DNA to degrade unwanted DNA, nucleases or RNA. They were not added to final list of possible contaminants, since they would degrade DNA or enzymes in enzyme activity assays, resulting in no detectable enzyme activity.

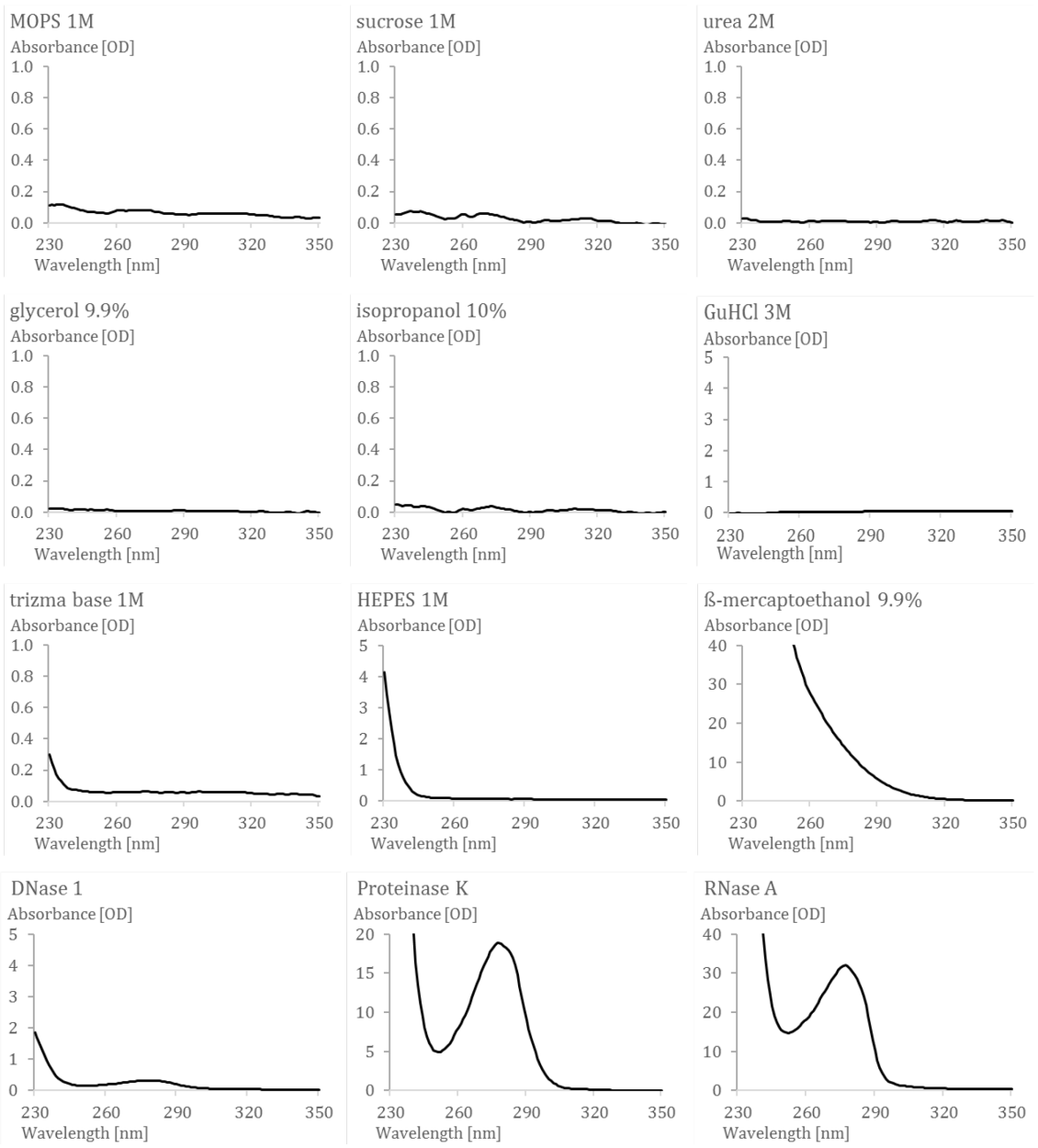


Figure 4: Absorbance spectra of measured pre-selected possible contaminants that were discarded from final list of possible contaminants.

Remaining 12 pre-selected substances from literature were added to the final list of possible contaminants: sodium citrate, betaine, sodium azide, EDTA, DTT, GITC, glycogen, dNTPs, phenol, HSA, IgG, and hemoglobin (Figure 5).

The absorbance spectra of sodiumcitrate (SC), betaine, sodiumazide (SA), EDTA, DTT, and guanidine isothiocyanate (GITC) had an A_{230} shoulder and were added to the final list of possible contaminants (Figure 5). Sodium citrate (SC) is used during DNA purification to neutralize negative charge of DNA for dissociation from water [90], [91], whereas betaine is often used in PCR reactions to enhance amplification of GC-rich sequences by denaturing secondary structures of DNA [92] and can be carried over from one step of a workflow to

another. During DNA purification or in elution buffers sodium azide (SA) is applied as preservative preventing the microbial growth [93]. Another substance found in nucleic acid elution buffers is EDTA, binding metal ions to inhibit nucleases and prevent nucleic acid degradation [94], [95]. DTT is found in enzyme solutions and reaction buffers to stabilize enzymes [96], while GITC is a salt applied for denaturation of nucleases in lysis buffers to prevent nucleic acid degradation during cleanup procedure [97].

Furthermore, glycogen, dNTPs and phenol were to be used as possible contaminants in this thesis. Glycogen showed a broad absorbance range from 230 – 350 nm wavelength (Figure 5). During ethanol precipitation of DNA, it can be used to trap DNA creating a visible pellet for easier handling [98]. As expected, dNTPs and phenol had absorbance maxima at 260 or 270 nm, respectively (Figure 5). For nucleic acid amplification with polymerases, dNTPs are used as single building blocks, while phenol is often used during purification of nucleic acids for denaturation of nucleases [99].

Like DNase 1, proteinase K, and RNase A, the proteins human serum albumin (HSA), IgG, and human hemoglobin (Hb) had an A_{230} shoulder and a second absorbance maximum at A_{280}. In addition, Hb showed a third absorbance peak at 410 nm wavelength (Figure 5). HSA, IgG and hemoglobin were applied as representative proteins, since they can be carried over from sample material.

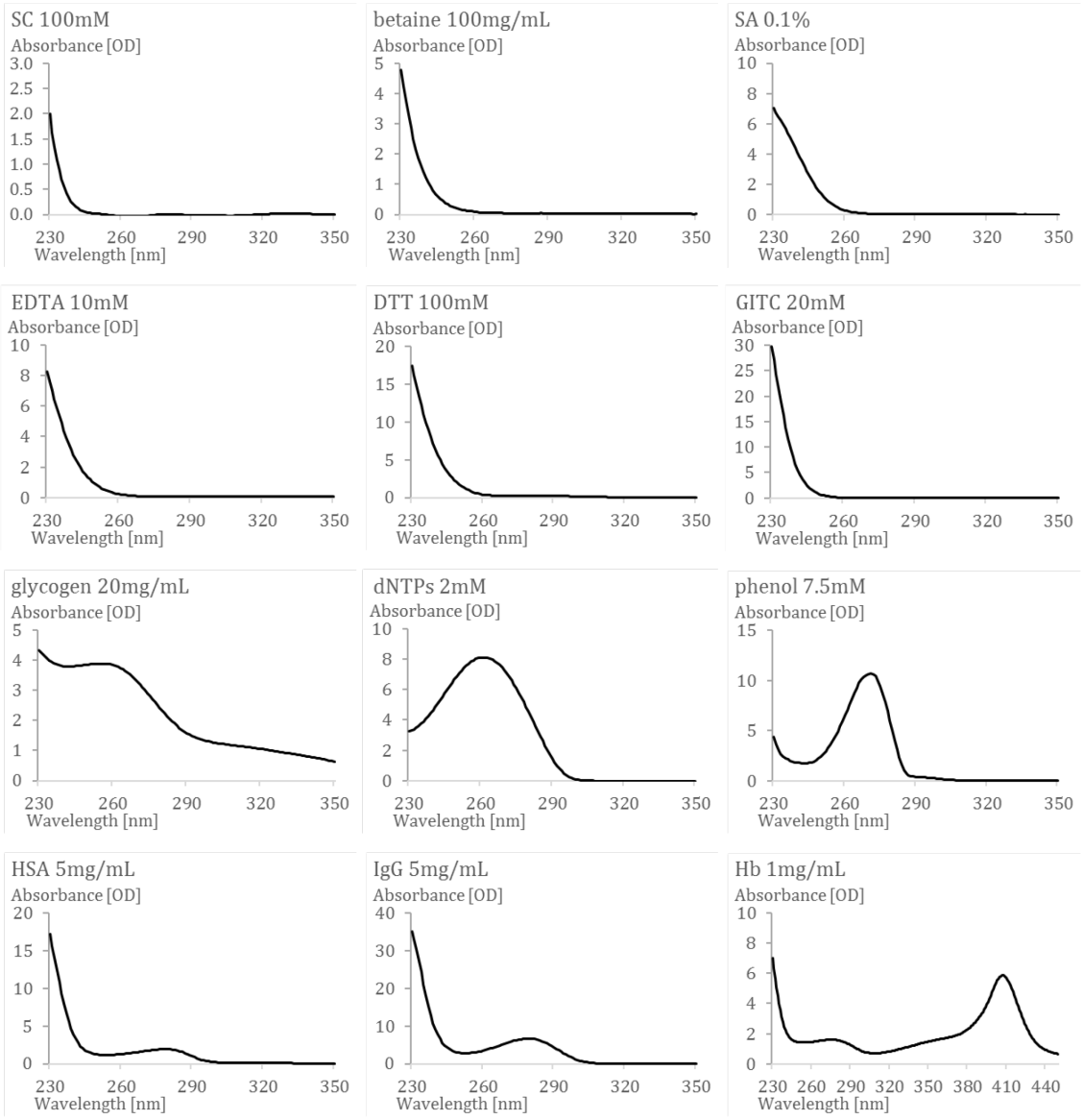


Figure 5: Absorbance spectra of 12 substances added to final list of possible DNA contaminants. Note different wavelength range on X-axis of Hb 1mg/mL.

Twelve selected possible contaminants were added to pure DNA samples, in four decreasing concentrations, between common working concentration and UV/Vis absorbance maximum above 0.03 OD, to record UV/Vis absorbance spectra, measure DNA concentration of contaminated sample, and quantify enzyme activities for algorithm development.

2.1.2 Influence of possible contaminants on DNA concentration

Different concentrations of possible contaminants were added to DNA samples and measured with DNA QIAsymphony application of the QIAxpert to determine volume of contaminated DNA to be applied on enzyme activity assays. Target DNA concentrations for

polymerase and ligase inhibition assay were 45 or 30 ng/µL, respectively. Measured DNA concentrations were plotted for each enzyme activity assay and means of all samples for each contaminant and enzyme assay were compared using ANOVA and Tukey-Kramer HSD test.

Results showed that betaine, and EDTA had no influence on measured DNA concentrations. DNA concentrations for polymerase assay samples with SA were constantly and independent of contaminant concentration higher than 45 µL, indicating a pipetting error during preparation of DNA pre-dilution. SA had no influence on DNA concentration of kinase assay samples (Figure 6 A - C).

GITC and sodium citrate (SC) at highest concentration led to lower DNA concentration detected for DNA samples of polymerase and higher DNA concentrations for kinase assay samples (Figure 6 D - E). The addition of DTT had no influence on measured DNA concentrations for polymerase assay samples, but led to increasing DNA concentrations estimated for kinase assay DNA samples (Figure 6 F). DTT can cause single stranded breaks in double stranded DNA [96], which in turn could lead to denaturation of double stranded DNA and higher UV/vis absorbance.

All other contaminants applied on enzyme activity assays, glycogen, hemoglobin, IgG, HSA, phenol, and dNTP, had a significant influence ($p < 0.05$) on DNA concentration estimation and led to increased DNA levels detected for higher contaminant concentrations (Figure 6 G - L). Overestimation of DNA concentration led to lower DNA volume added to enzyme activity assays as described in chapter 4.2.9.

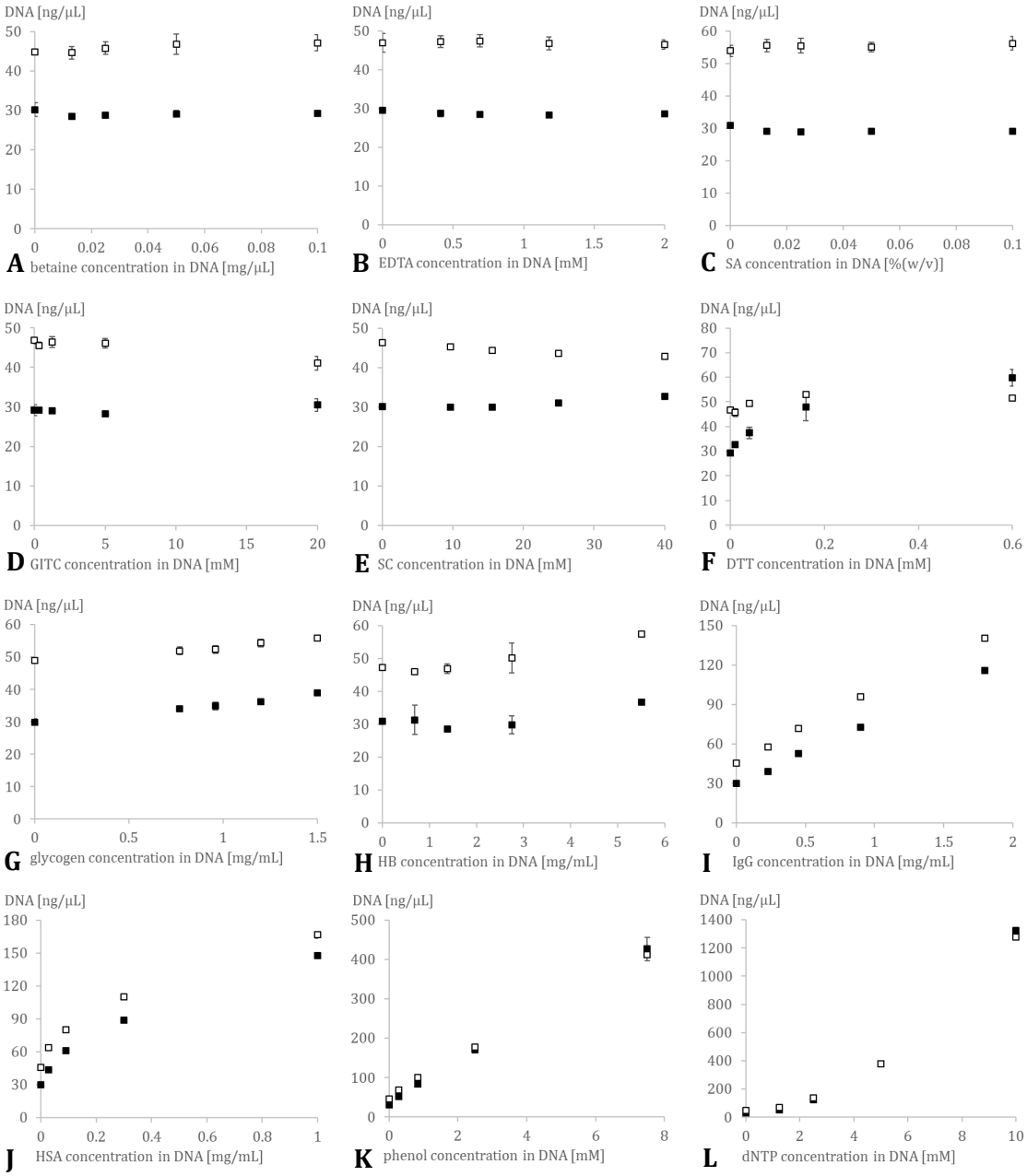


Figure 6: Measured DNA concentrations of contaminated DNA samples for enzyme activity assays. Measured nucleic acid concentrations [ng/µL] of contaminated ssDNA for polymerase assay in white and those of contaminated dsDNA fragments for ligase assay in black. Target concentrations were 45 ng/µL or 30 ng/µL, respectively. N = 6 measurement replicates, error bar = standard deviation.

2.2 Measurement of polymerase inhibition

2.2.1 Phi-Inhibition-Assay to measure Taq DNA polymerase activity

The Phi-Assay, used to determine activity of polymerases by comparing their activity to a reference enzyme was developed at QIAGEN. It is based on measurement of fluorescence intensity, resulting from a linear amplification of the ssDNA PhiX 174 plasmid (PhiX DNA)

using only 1 primer and EvaGreen as intercalating dye to detect double stranded DNA (dsDNA) (Figure 7). Polymerase activity is determined, using decreasing concentrations of a reference enzyme to create a standard curve by plotting theoretical enzyme activity of dilutions as function of slope of fluorescence signal over several cycles. The slope of fluorescence signal of tested enzyme was then used to calculate its activity.

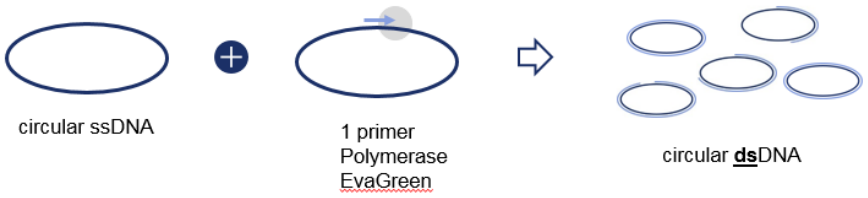


Figure 7: Schematic drawing of Phi-Inhibition-Assay principle.

To determine the effect of different contaminants on Taq polymerase activity, it was decided to use the Phi-Inhibition-Assay, adapted from Phi-Assay. A standard curve of decreasing enzyme concentrations was created with Taq polymerase. The influence of a contaminant on Taq polymerase was determined by adding contaminated PhiX DNA to standard with highest enzyme concentration and calculating relative enzyme activity based on standard curve.

2.2.2 Establishment of Taq DNA polymerase standard curve

To quantify polymerase activity in presence of contaminants (chapter 2.2.3), first a reproducible standard curve consisting of four decreasing concentrations of Taq polymerase and a no-enzyme negative control was established. The theoretical enzyme activity of dilutions (based on enzyme activity claimed by manufacturer) was plotted against slope of fluorescence signal over cycles 10 to 25 (Figure 8).

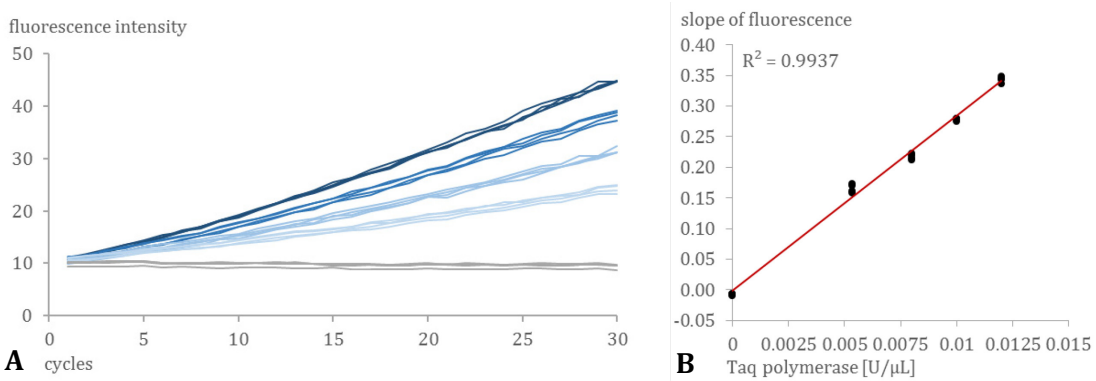


Figure 8: Representative example of fluorescence signal recorded for Phi-Inhibition-Assay standard curve (A) and plot of slope of fluorescence signal vs. theoretical enzyme activity in U/µL (B). (A) Standard 1 to 4 are presented in dark to light blue, and negative control in grey. N = 4 technical replicates on one Phi-Inhibition-Assay run.

At first 0.0019, 0.0016, 0.0012, and 0.0008 U/µL Taq polymerase used in Phi-Assay for reference enzyme were tested. Over all, results of all four runs were similar and comparable. In Figure 8, the fluorescence signal over time for standards and negative control of run 1 are presented as representative results. Looking at the fluorescence intensity of standards, a clear decrease of slope was observed with decreasing enzyme concentration. The resulting standard curves of all 4 runs had R^2s ≥ 0.99 (Supplementary Table 1). After calculation of actual activity, using standard curve, mean of slopes, and dilution factor all standards had activities of 5 ± 0.3 U/µL, resulting in a standard deviation ≤ 0.2 U/µL around expected theoretical activity of 5 U/µL (Figure 9).

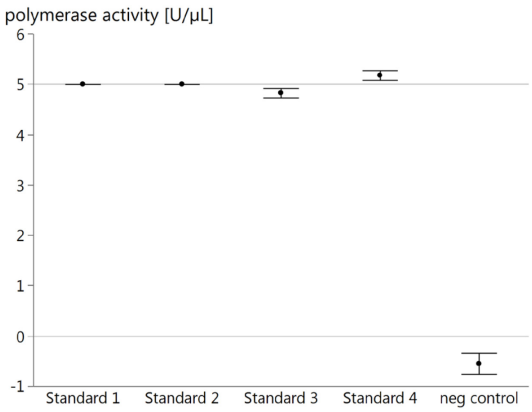


Figure 9: Calculated enzyme activities of standards and negative control. Presented are mean enzyme activities calculated using standard curve and slope of fluorescence signal. N = 4 independent runs, error bar = standard deviation. Grey lines indicate expected values of 5 U/µL for all standards and 0 U/µL for negative control.

2.2.3 Influence of contaminants on Taq DNA polymerase activity

To investigate the influence of possible impurities on polymerase activity, PhiX DNA was combined with four decreasing concentrations of each contaminant or buffer EB for clean DNA control and applied on Phi-Inhibition-Assay. UV/Vis absorbance spectra and DNA concentrations to calculate DNA input volume of same contaminated DNA samples in Phi-Inhibition-Assay were beforehand recorded on QIAxpert (Figure 6).

Mean percentage of six replicates of measured Taq polymerase activity was plotted against contaminant concentration in PhiX DNA. Using one-way Analysis of Variance (ANOVA) and Tukey-Kramer Honest Significant Difference (Tukey-Kramer HSD) test, significant differences in average enzyme activities of different contaminant concentrations were identified. Results showed that overall standard deviation (StDev) of replicates was ≤10% for measured Taq activity, except for 12 samples were StDev was between 10 and 15% (Figure 10 and Supplementary Table 2). Betaine and DTT had no influence on Taq polymerase activity in applied concentrations (Figure 10 A - B). Decreased Taq polymerase activity was observed for increasing concentration of dNTPs, EDTA, human hemoglobin (Hb), HSA, and IgG; whereby for IgG and Hb second highest concentration resulted in complete inhibition of enzyme, whereas lowest concentration of dNTPs, EDTA, and Hb had no effect on enzyme activity (Figure 10 C - G). Figure 10 H and I show that only highest concentration applied for sodium azide (SA) and GITC had inhibitory effect on enzyme and all concentrations applied of glycogen reduced Taq polymerase activity by about 30%. Sodium citrate (SC) led to complete Taq inhibition for all applied concentrations (Figure 10 K). Interestingly, lowest concentration of phenol led to improved Taq activity, whereas all other concentrations showed increasing inhibition with increasing phenol concentration (Figure 10 L).

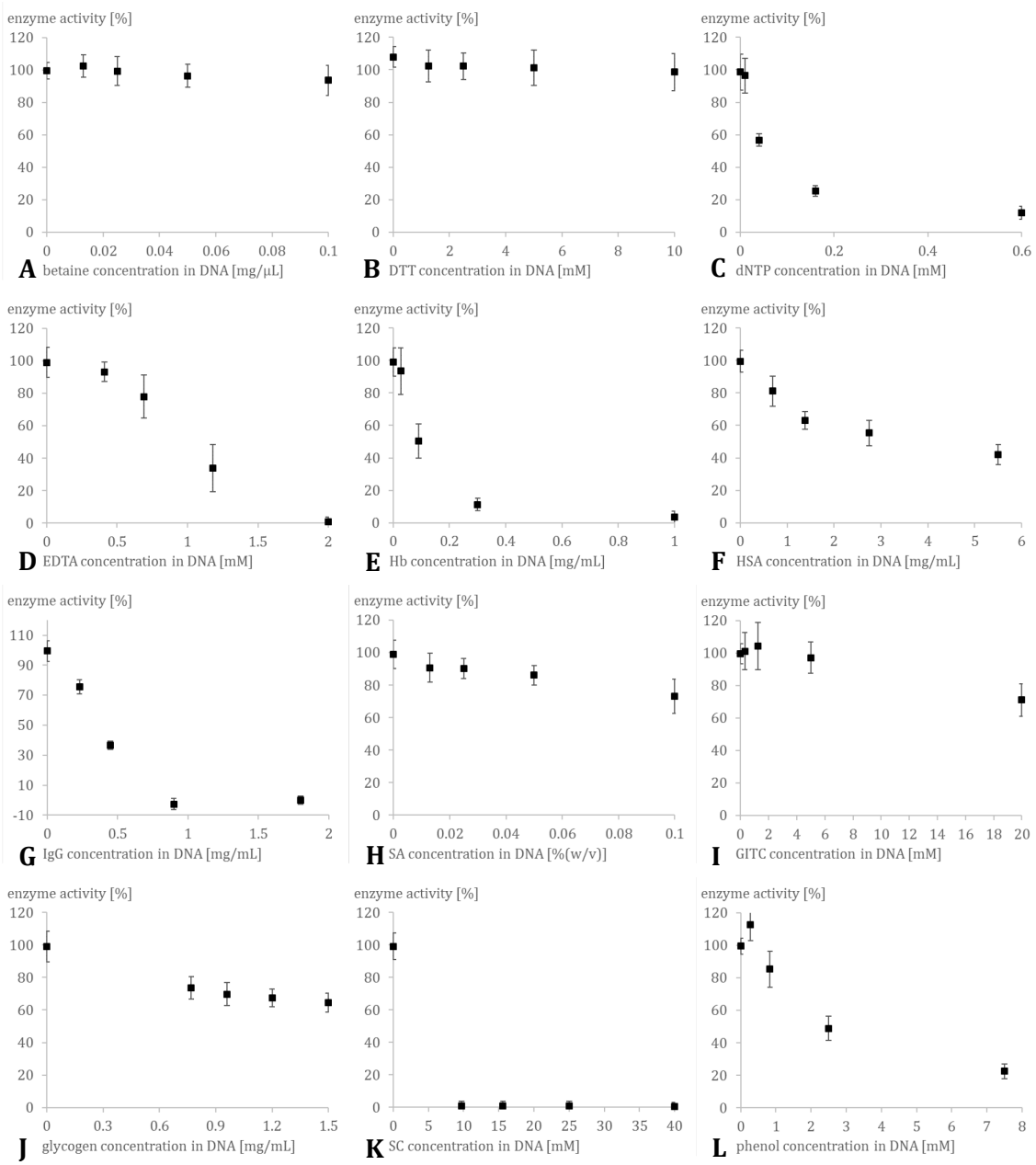


Figure 10: Mean percentage of Taq activity plotted against contaminant concentration in DNA. Presented are results for all tested contaminants: (A) Betaine, (B) DTT, (C) dNTPs, (D) EDTA, (E) human hemoglobin, (F) HAS, (G) IgG, (H) sodium azide, (I) GITC, (J) glycogen, (K) sodium citrate, and (L) phenol. N = 6 replicates from independent Phi-Inhibition-Assay runs, error bar = standard deviation, note different y-axis scale for (G) IgG.

2.3 Measurement of ligase inhibition

2.3.1 Gel electrophorese based assay to measure T4 DNA Ligase activity

To measure T4 DNA Ligase activity in presence or possible contaminants, an assay was designed to measure T4 DNA Ligase activity based on amount of ligated fragment detected by capillary gel electrophoresis, using the QIAxcel Advanced as detection instrument. The region flanking XhoI restriction site on plasmid pCMVbeta was selected to serve as template

for PCR to obtain dsDNA fragments for gel electrophoresis based ligase assay. Restriction digest of PCR product with XhoI would result in a 4 nt 5' overhang, which would subsequently be ligated by T4 DNA ligase (Figure 11).

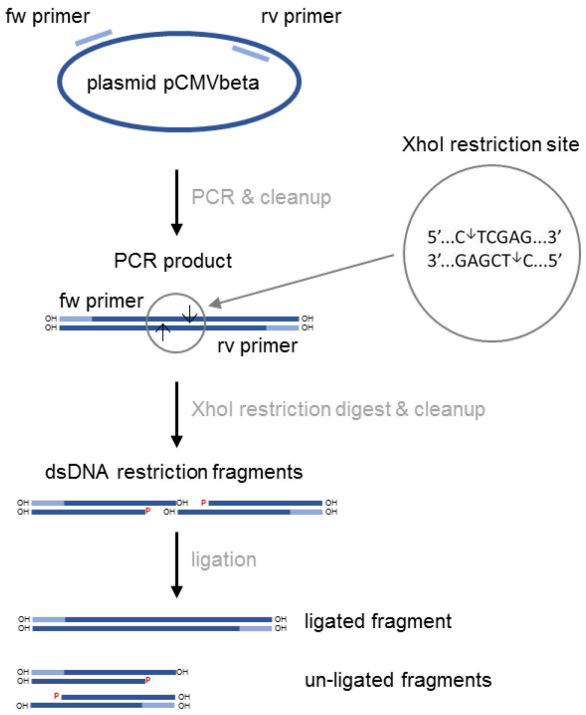


Figure 11: Schematic drawing of ligase assay workflow and experimental setup.

First, primers were designed flanking the XhoI restriction site and resulting in a PCR product of 549 bp. Results of first run through the experimental workflow (Figure 11), showed that selected primers generated a specific PCR product of about 549 bp (Figure 12 A). Subsequently, restriction digest with XhoI was carried out and resulted in two fragments of about 271 and 274 bp, although only one band was observed in digital gel image due to resolution (Figure 12 B).

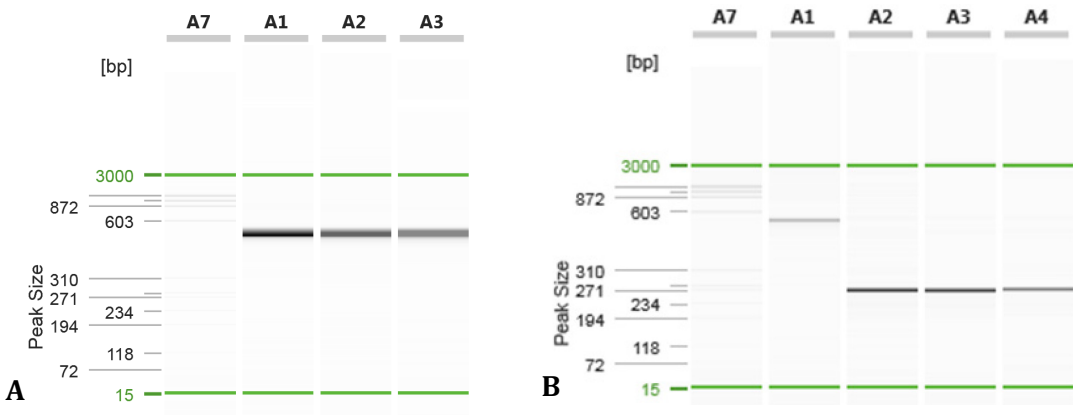


Figure 12: Digital Gel image of (A) PCR product and (B) restriction fragments after cleanup. Presented were in (A): Size marker (A7) 3 lanes showing purified 549 bp PCR product (A1 - A3), and (B): Size marker (A7), diluted PCR product (A1), and 3 lanes of purified XhoI digested fragments (A2 – A4). Gel images were recorded with QIAxcel Advanced, using a High Resolution Cartridge, Alignment Marker 15bp – 3kb, Size Marker FX 174/HaeIII, and Method OM500.

To investigate, whether XhoI digested dsDNA fragments could be ligated by T4 DNA Ligase and whether different concentrations of the enzyme would lead to different amounts of ligated and non-ligated fragments detected on QIAxcel, a T4 DNA Ligase dilution series containing 1.2, 0.5, 0.4, 0.3 or 0.2 U/µL, and a Ligase negative control were prepared and incubated with restriction fragments.

The restriction site for XhoI on the 549 bp PCR product was near the middle, resulting in a 271 bp and a 274 bp fragment with 4 nt symmetrical overhangs (Figure 11). In theory, ligation of these fragments would lead to three possible fragments: 546 bp, 549 bp, or 552 bp. With the QIAxcel High Resolution Kit a distinct detection of these fragments would be possible, but if analysis parameter in ScreenGel Software were set to detect peaks with a minimum distance of three seconds, ligated and un-ligated fragments would be summarized in one peak each. As expected, the QIAxcel gel image, presented in Figure 13 A, showed one band for ligated and one for un-ligated DNA fragments, while the electropherogram showed that each gel-band was composed of at least two peaks (Figure 13 B). Furthermore, gel image showed that decreasing Ligase concentration led to decreasing amount of ligated DNA fragments (Figure 13).

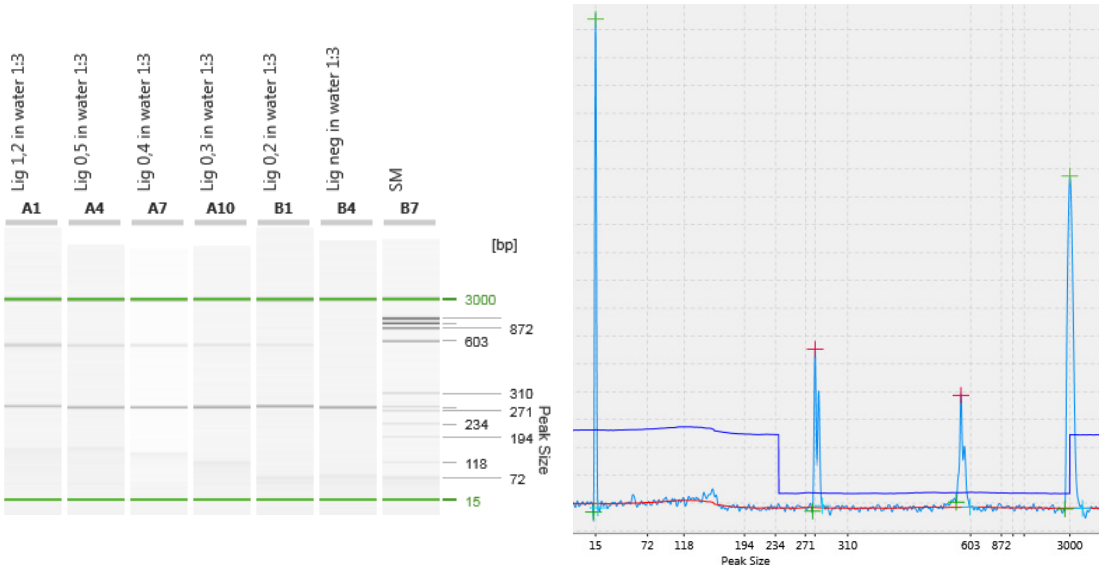


Figure 13: Digital Gel image of ligation reactions (left) and corresponding electropherogram of lane A1 (right). Presented in gel image were 1 representative lane for each of applied enzyme concentration, diluted 1 in 3 in RNase-free water, and recorded immediately after reaction.

The amount of ligated and un-ligated fragments based on evaluation of percent normalized area (% NA) values were compared in Figure 14. The % NA value describes the percentage of the area under the curve of each peak in an electropherogram, where the sum of the areas under the curve of all detected peaks, excluding alignment markers, equals 100%. As expected, results showed that with decreasing T4 ligase concentration, the % NA of peaks representing un-ligated fragments increased, while % NA of peaks representing ligated fragments decreased.

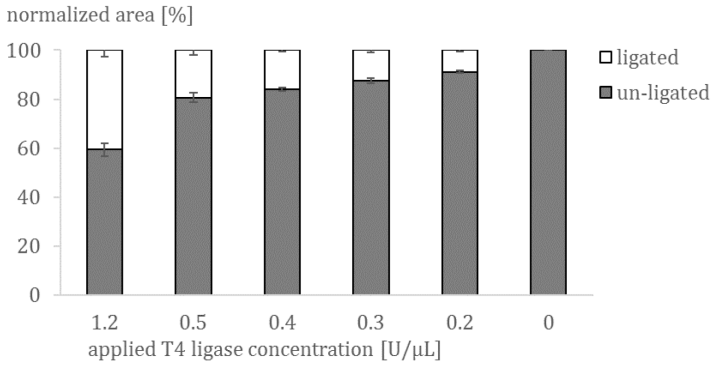


Figure 14: Mean % NA values detected for different T4 ligase concentrations. Ligated and un-ligated DNA was quantified using % NA values for detected long or short DNA fragments after incubation with decreasing concentrations of T4 ligase. N = 3 replicates applied on QIAxcel, error bar = standard deviation.

2.3.2 Establishment of T4 DNA Ligase standard curve

To investigate whether applied enzyme concentrations could be used as standard curve to quantify T4 ligase activity in presence of contaminated DNA, enzyme concentrations were plotted against percentage normalized area of ligated fragments. The resulting standard curve had an $R^2 = 0.98$ (Figure 15 A). Mean enzyme activities of standards, after subtracting dilution factor, were expected to be 120 U/µL and measured values were in range of 116 - 130 U/µL, with standard deviation ≤ 14 U/µL (Figure 15 B).

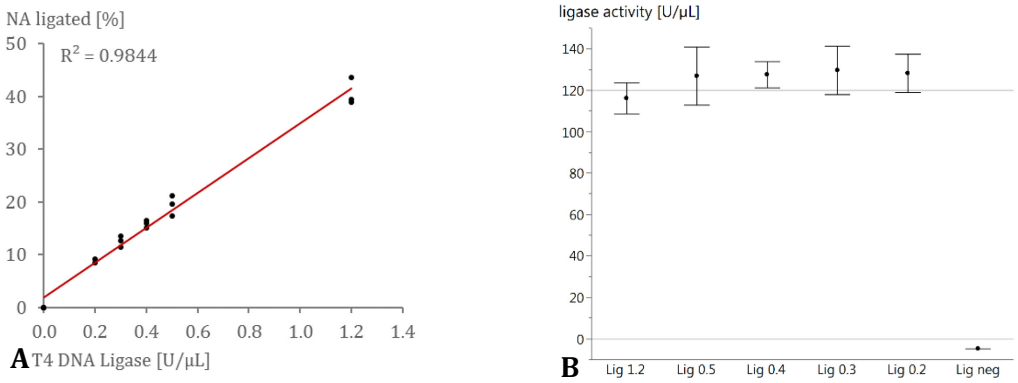


Figure 15: (A) Standard curve for T4 DNA Ligase on gel electrophorese based ligase assay and (B) activities of standards after subtraction of dilution factor. (A) Standard curve was generated by plotting measured % NA values of ligated fragments against applied T4 DNA Ligase dilutions. (B) Standard curve and dilution factor were used to calculate enzyme activity of standards. Grey lines indicate expected values of 120 U/µL for all standards and 0 U/µL for ligase negative control. N = 3 technical replicates applied on real-time ligase assay, error bar = standard deviation.

The standard curve was repeated in 3 runs with 1.2, 0.8, 0.5, and 0.3 or 0.2 U/µL T4 DNA ligase dilutions as well as a ligase negative control to test reproducibility. The 0.4 U/µL dilution was exchanged for a 0.8 U/µL dilution to obtain an even distribution over the whole range of applied concentrations. On each run, three replicates were applied of each sample. The results showed that standard curve of first run had an R^2 or 0.99 (Figure 16 A). The ligase dilution with 1.2 U/µL was below regression line in run 2 and 3, indicating that it was beyond linear range. R^2 values for these runs were 0.95 or 0.96. When highest ligase concentration was excluded from evaluation for run 2 and 3, R^2 values were 0.99 for both runs (Figure 16 B and C).

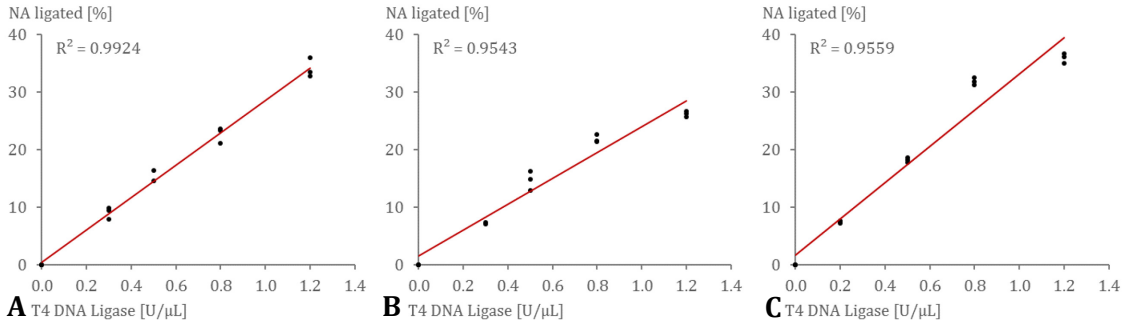


Figure 16: Standard curves for T4 DNA ligase of 3 independent runs (A) run 1, (B) run 2, and (C) run 3. Measured percentage normalized area of ligated DNA fragments was plotted against applied enzyme dilution. N = 3 replicates on each run.

To optimize the standard curve for T4 DNA ligase activity quantification, another 4 standard curves were prepared using enzyme concentrations of 1.0, 0.8, 0.5, 0.3, and 0.2 U/µL, representing Standard 1 – 5, as well as a no ligase control. R^2 values for all runs were ≥ 0.97. After subtraction of dilution factor, mean ligase activities of 4 independent runs were expected to be 120 U/µL and measured values were between 113 and 134 U/µL with standard deviation ≤ 7 U/µL (Figure 17).

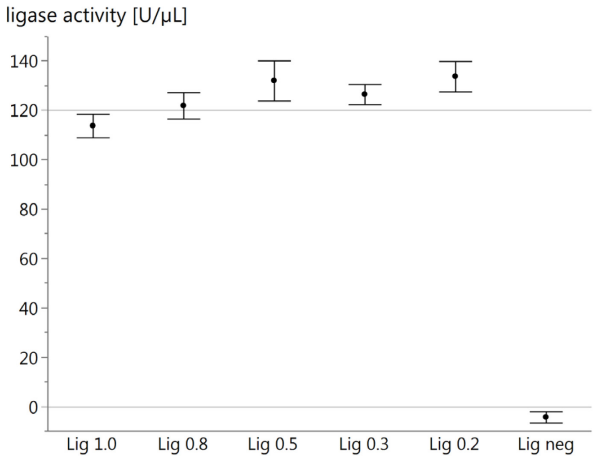


Figure 17: Calculated T4 DNA ligase activities of standards and negative control. Presented are mean enzyme activities calculated using standard curve and percentage normalized area for ligated DNA fragments. Grey lines indicate expected values of 120 U/µL for all standards and 0 U/µL for ligase negative control. N = 4 independent runs, error bar = standard deviation.

Based on presented results, the enzyme dilutions with 1.0, 0.8, 0.5, 0.3, and 0.2 U/µL T4 DNA ligase were chosen for standard curve to test influence of possible contaminants on ligase activity.

2.3.3 Influence of contaminants on T4 DNA Ligase activity

In order to investigate the influence of contaminants on T4 DNA Ligase activity, four decreasing concentrations of contaminants or buffer EB for clean DNA control were added to restriction fragments. Before contaminated DNA samples were applied on gel electrophoresis based ligase assay, same contaminated DNA samples were applied on QIAxpert to measure DNA concentration to calculate DNA input volume in ligase assay (Figure 5), and to record UV/Vis absorbance spectra for data modelling.

Mean percentage of T4 ligase activity measured for each sample in six replicates, were plotted against contaminant concentration in DNA (Figure 18). Overall results showed StDev of replicates ≤10%, except for 6 samples with StDev between 10 and 15%, and 1 sample with StDev of 20% (Figure 18 and Supplementary Table 2). To determine significant differences in mean ligase activities, ANOVA and Tukey-Kramer HSD test were applied. EDTA had no influence on T4 DNA Ligase activity (Figure 18 A), whereas betaine in highest concentration applied had a positive effect, leading to about 120% activity (Figure 18 B). Interestingly, low concentrations of human hemoglobin (Hb) and HSA also had a positive effect on ligase activity, leading to about 140% or 150% activity measured. Low concentrations of these contaminants were comparable to control without contaminant (Figure 18 C and D). Highest concentration of sodium azide (SA) led to inhibition of T4 ligase by about 20%, while all other concentrations of same contaminant had no influence (Figure 18 E). Results for GITC and DTT showed increasing inhibition with increasing concentration, with a maximum inhibition of about 30% for the highest concentrations (Figure 18 F and G). For glycogen, all applied concentrations led to an inhibition of about 20% (Figure 18 H). IgG, dNTPs, sodium citrate (SC), and phenol caused decreased activity for increasing contaminant concentration, with complete inhibition of T4 ligase for second lowest concentration of IgG, and highest two concentrations of dNTPs, sodium citrate, and phenol (Figure 18 I - L).

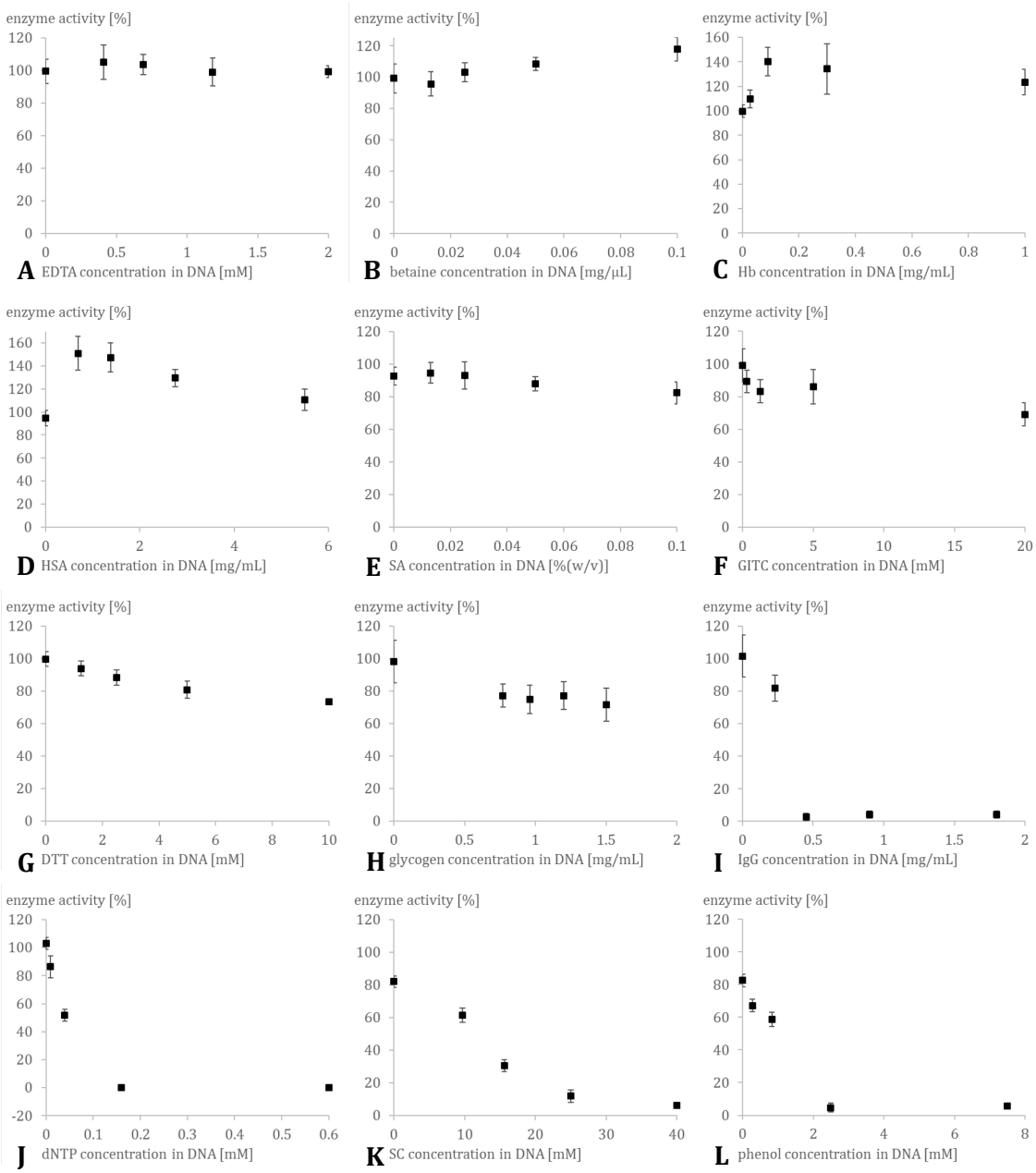


Figure 18: Mean percentage of T4 ligase activity plotted against contaminant concentration in DNA. Presented are results for all tested contaminants: (A) EDTA, (B) Betaine, (C) human hemoglobin, (D) HSA, (E) sodium azide, (F) GITC, (G) DTT, (H) glycogen, (I) IgG, (J) dNTPs, (K) sodium citrate, and (L) phenol. N = 6 replicates from independent ligase inhibition assay experiments, error bar = standard deviation, note different y-axis scale for (C) human hemoglobin, (D) HSA, and (J) dNTP.

2.4 Measurement of kinase inhibition

2.4.1 Radiometric assay to measure T4 PNK actvity

The radiometric kinase assay, based on a transfer of radioactive labeled phosphate and first described by Sambrook *et al.* [45], was selected to measure activity of T4 PNK in presence of possible contaminants. To determine enzyme activity, the amount of radioactive P^{32}

transferred by the enzyme from [γ-^{32}P]ATP to the 5'end of a dsDNA fragment was measured, using a Beckman LS 6500 scintillation counter. A purified PCR product was applied as dsDNA fragment (Figure 19).

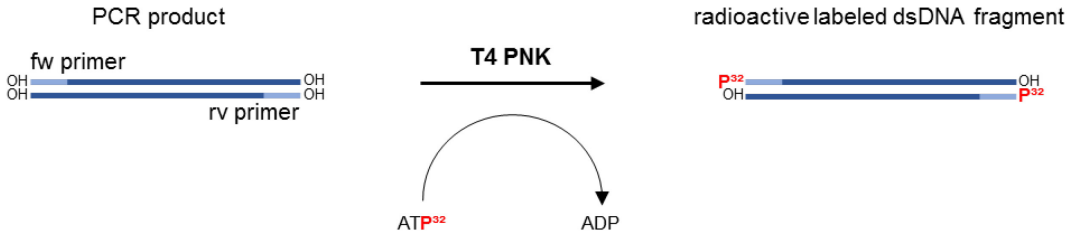


Figure 19: Schematic drawing of radiometric T4 PNK assay principle.

For relative quantification, a standard curve was to be established as previously done and described for polymerase and ligase assay. Therefore, T4 PNK was applied in 200, 150, 100, and 50 U/µL to reaction mix containing reaction buffer, dsDNA fragments, and [γ-^{32}P]ATP. Measured counts per minute (CPM) were plotted against theoretical enzyme activities of dilutions.

Results for measured CPM varied strongly between independent runs and showed poor correlation with enzyme activity of T4 PNK dilutions (Figure 20).

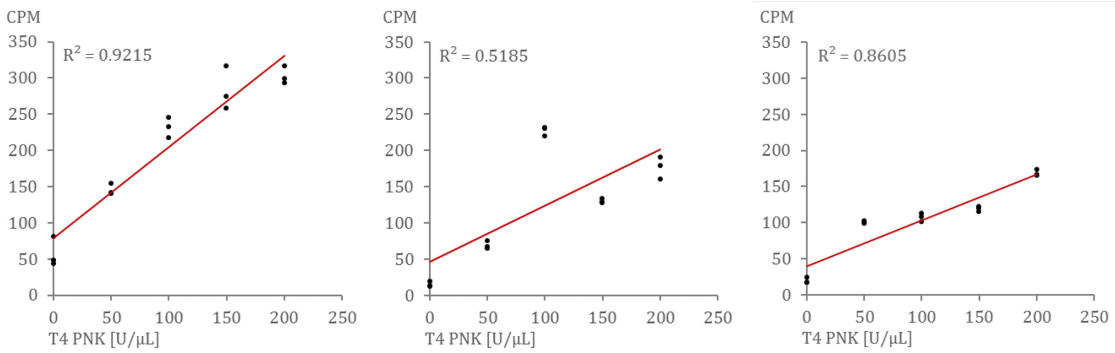


Figure 20: Standard curves for T4 PNK of 3 independent runs. Measured counts per minute (CPM) were plotted against applied enzyme dilution. N = 3 replicates on each run.

Due to poor correlation between measured CPM and applied T4 PNK concentrations, enzyme activity data investigating the influence of possible contaminants on T4 PNK were not recorded.

2.5 Purity assessment of DNA samples

For purity assessment of DNA samples 5 classes were created to be used as target values for data evaluation and algorithm training and testing, based on measured enzyme

activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, where DNA was considered pure, when measured enzyme activity was >80%. From here on, classes based on enzyme activities will be referred to as "actual classes". For evaluation of different methods for purity assessment, actual classes will be considered true and will be plotted against values describing nucleic acid purity to visualize results.

If purity values are continuous, actual classes will be plotted against box plots, representing the distribution of measured purity values for each class. In an ideal case, where actual class and purity values correlate perfectly with each other, the median of purity values would climb for increasing actual class and variance of purity values would not overlap between different classes (Figure 21).

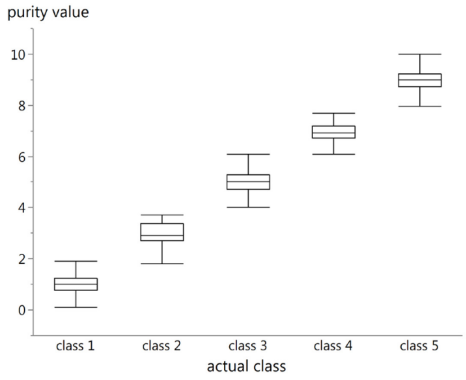


Figure 21: Demonstrative plot of actual class vs. continuous purity value. Presented are data for a fictive case, where purity value and actual class correlate perfectly with each other. Actual classes were based on measured enzyme activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5

When purity values were predicted as classes resulting from classification algorithms, confusion matrices will be used as performance measurement. A confusion matrix is presented as a table with four different combinations of predicted and actual values (Figure **22**):

- True positives (TP): a predicted 11positive value is an actual positive value
- False positive (FP): a predicted positive value is an actual negative value
- False negative (FN): a predicted negative value is an actual positive value
- True negative (TN): a predicted negative value is an actual negative value

These four categories are used to determine performance measures such as Recall, Precision, F-measure, and Accuracy. Recall describes how many instances were predicted correctly out of all actual positive values (Figure 22 yellow), whereas precision is a measure for how many values were predicted correctly out of all predicted positive values (Figure **22** green). The F-measure combines Recall and Precision by building their harmonic mean.

The proportion of overall correctly predicted values is defined by the Accuracy (Figure 22 red). Higher values for all performance values indicate better results.

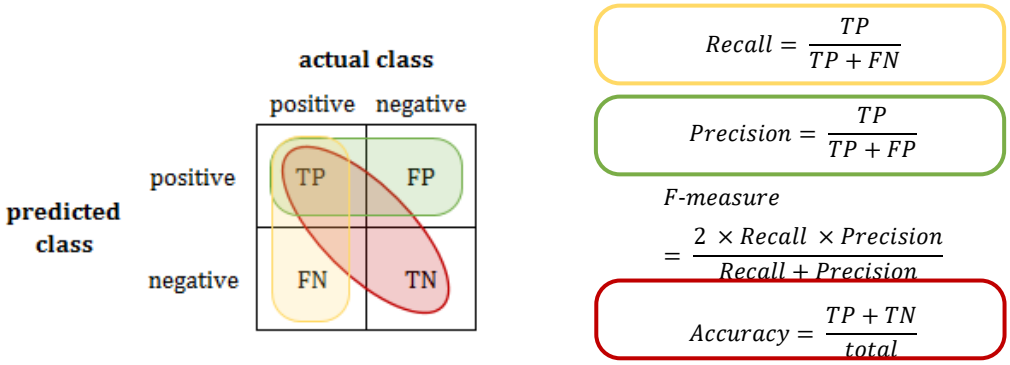


Figure 22: (left) Schematic presentation of confusion matrix and (right) performance measures with formulas. Formulas to determine performance measures and areas of confusion matrix representing performance measures were color-coded. Adapted from [100].

2.5.1 Absorbance ratios for evaluation of DNA sample purity

A_{260}/A_{280} and A_{260}/A_{230} absorbance ratios are commonly used to determine sample purity of nucleic acid samples. Absorbance values at A_{230}, A_{260}, and A_{280} were recorded for all samples of development dataset using the QIAxpert UV/Vis application, ratios were calculated and plotted against actual classes. The results showed no correlation between actual class and absorbance ratios. For A_{260}/A_{280} ratios, median of ratios were ~ 1.7 for all actual classes, whereas median of A_{260}/A_{230} ratios were highest for class 3 and went down towards class 1 and class 5 (Figure 23).

Interestingly, A_{260}/A_{280} ratios above 2 in class 3 to 5 and A_{260}/A_{230} ratios greater 2.2 in all 5 actual classes were observed (Figure 23). Absorbance ratios above approximately 2.0 or 2.2 are unusual for nucleic acid samples, since absorbance maximum of nucleic acids is at A_{260}, while impurities have absorbance maxima at A_{230}, A_{270}, or A_{280} (Ref. [63], [64], Figure 1, and Figure 5). Therefore, elevated impurity concentration in nucleic acid samples would lead to higher denominator values of absorbance ratios and thus to lower overall ratio values.

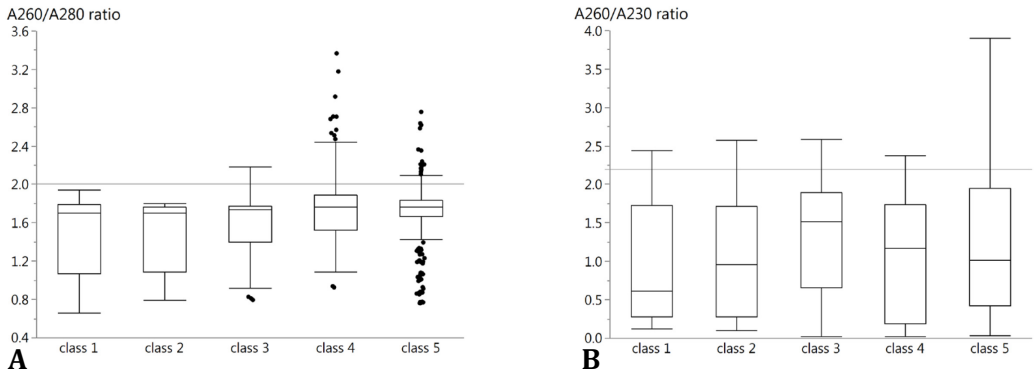


Figure 23: Actual classes plotted against box plots of (A) A_{260}/A_{280} and (B) A_{260}/A_{230} ratios. Actual classes were created based on enzyme activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, absorbance ratios were calculated using measurements of QIAxpert UV/Vis application, and outlier box plots were generated with JMP, showing median, 1st /3rd quartile and 1st /3rd quartile -/+ 1.5*interquartile range. Grey lines indicate A_{260}/A_{280} = 2.0 or A_{260}/A_{280} = 2.2.

Inspection of samples with A_{260}/A_{280} ratios above 2 revealed that they contained sodium azide (SA) or DTT, which are both possible contaminants with an absorbance maximum at ≤ A_{230}. Representative UV/Vis absorbance spectra of these samples and their control samples without contamination were plotted and showed that indeed the right shoulder of the absorbance peak shifted to the right, resulting in an increased A_{260} absorbance with increasing concentration of SA or DTT, while A_{280} values were not affected (Figure 24 A and B). A_{260}/A_{230} ratios > 2.2 in classes 1 through 4, were observed for samples contaminated with dNTPs. Representative absorbance spectra of DNA samples contaminated with increasing concentrations of dNTPs showed that absorbance at A_{260} climbed faster than absorbance at A_{230}, resulting in higher A_{260}/A_{230} ratios (Figure 24 C).

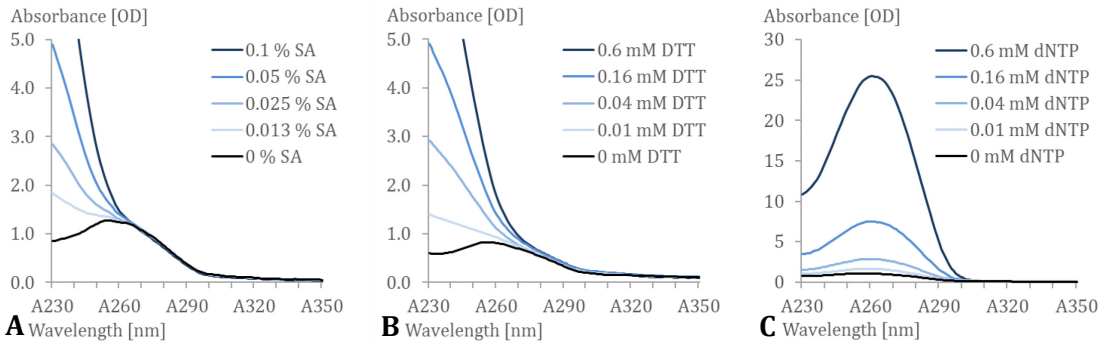


Figure 24: UV/Vis absorbance spectra of samples with (A and B) A_{260}/A_{280} ratios > 2.0 or (C) A_{260}/A_{230} ratios > 2.2. Presented were representative spectra of DNA contaminated with (A) sodium azide, (B) DTT, or (C) dNTP recorded with QIAxpert UV/Vis application.

For easier comparison of absorbance ratios with classification algorithms, ratios were also divided into 5 classes, with class 5 representing pure DNA with A_{260}/A_{280} values between 1.8 and 2.0, and A_{260}/A_{230} ratios in range of 1.8 – 2.2, based on reference [64]. Sample purity

and classes were assumed to decrease with increasing or decreasing ratios in both directions (Figure 25).

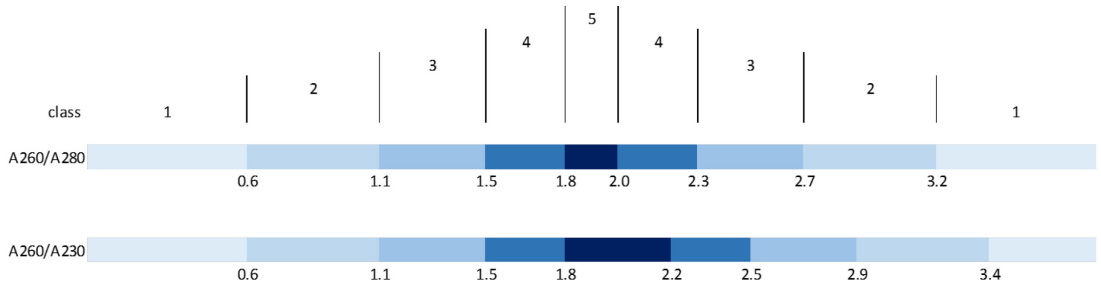


Figure 25: Classes generated using absorbance ratios. Class 5 was defined as recommended in literature and guidelines.

All samples were then assigned to classes based on A_{260}/A_{280} or A_{260}/A_{230} ratios from UV/Vis absorbance measurements as described above (Figure 25). The overall class based on both absorbance ratios was then determined for each sample by choosing the lower of both classes, since purity of a DNA sample would be considered poor, if one of both ratios was below expected range. The classes based on absorbance ratios were then plotted against actual classes in a confusion matrix, displayed in Figure 26. The result showed that with classes based on absorbance ratios, an accuracy of 17% was achieved and 570 of 684 were misclassified (Figure 26).

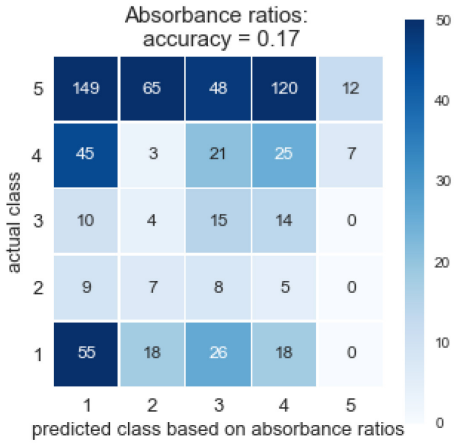


Figure 26: Accuracy and color encoded confusion matrix for classes predicted based on absorbance ratios vs. actual classes. Actual classes were created based on enzyme activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, and predicted classes based on absorbance ratios were determined as described in Figure 25. Color scale of heat map on right.

2.5.2 Development of novel method for assessment of DNA purity

To investigate whether whole spectra between A_{230} and A_{410} of contaminated DNA, recorded with UV/Vis spectrometer, could be used to predict enzyme activities of same samples and consequently, predict nucleic acid sample purity, mathematical data modelling methods were applied. Therefore, recorded absorbance spectra were used as input data in three different variations and an example for each variation was presented in Figure 27 to demonstrate differences between different input spectra.

- **raw spectra**: These minimally processed spectra recorded on QIAxpert should be comparable to minimally processed spectra of same samples recorded on other instruments (Figure 27 A). The resulting data model would be able to use raw spectra directly from the instrument, without any further manipulation.
- **A_{260} normalized spectra**: Named raw spectra from QIAxpert were normalized to A_{260} = 1 OD (Figure 27 B). Therefore, the resulting algorithm considered ratio of DNA and contaminant concentration. The A_{260} normalization would be done for sample spectra that were to be analyzed with the resulting data model.
- **delta spectra**: By spectral content profiling generated nucleic acids spectra, through integrated spectral content profiling (SCP) on QIAxpert DNA QIAsymphony application, were subtracted from raw spectra. Consequently, delta spectra represented absorbance of sample impurities only (Figure 27 C), and resulting data model would be independent of DNA concentration. Delta spectra would be calculated for all sample spectra that were to be analyzed with resulting algorithm and thus require Spectral Content Profiling (SCP).

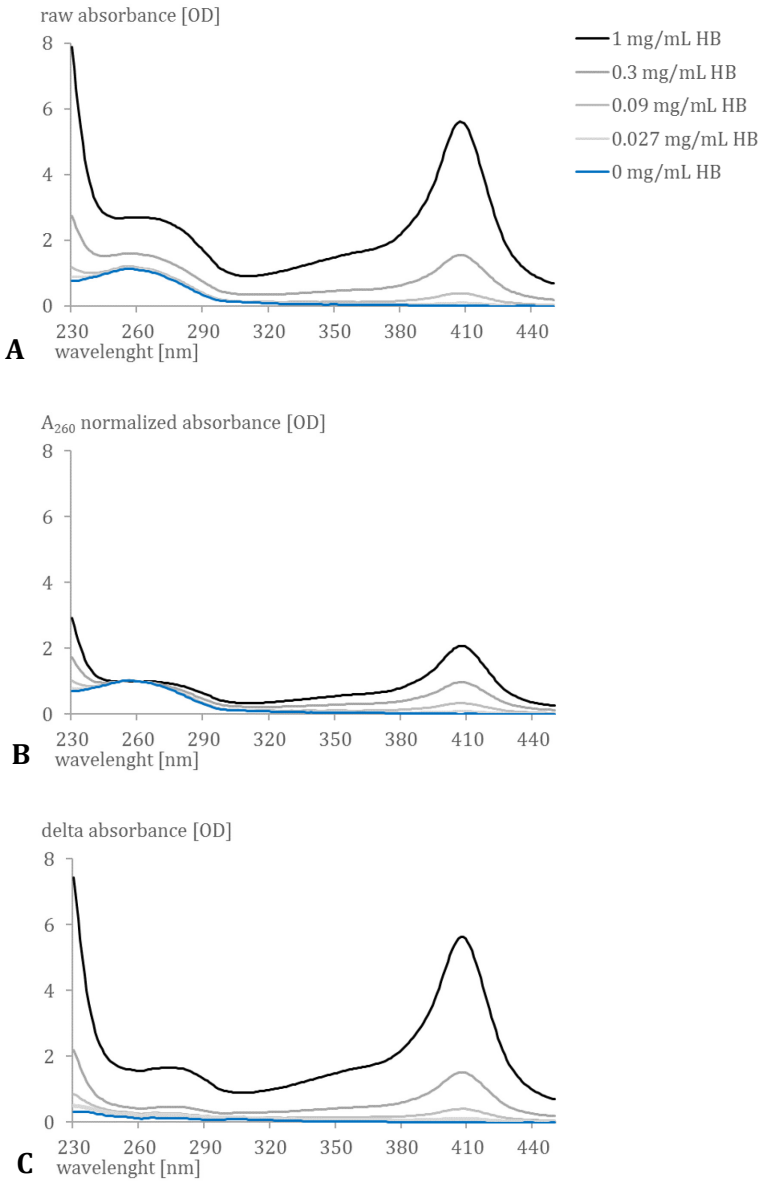


Figure 27: Absorbance spectra of DNA contaminated with human hemoglobin (HB). (A) Raw, (B) A_{260} normalized, and (C) delta spectra of 1 representative measurement replicate for five different DNA samples with or without human hemoglobin.

Datasets with different input spectra were each divided into three subsets for data model training and testing with same proportion of samples belonging to actual class one through five in each subset (for class description see chapter 2.5). The (1) trainings dataset consisted of 70% of instances and was used to train algorithms. The (2) development dataset, containing 15% of instances, was applied to test and optimize different data pre-processing and algorithm parameter settings. Finally, the (3) test dataset, containing remaining 15% of instances, was used to compare optimized algorithms using different input spectra.

The input spectra showed high collinearity, meaning that various input variables had similar values throughout the complete dataset, due to the similar shape of absorbance spectra of many possible contaminants (Figure 5). These input variables carried redundant information that could lower the performance measures of algorithms using absorbance spectra to predict actual classes. Therefore, data were pre-processed, using principal component analysis (PCA) or near zero variance (nzv), to eliminate redundant input variables and reduce the number of input data for subsequent algorithm. In PCA new, abstract values are generated from original input data, that explain the variance of input data using less features. These values are called principal components and serve as input data for subsequently applied classification algorithm. The nzv method reduces the count of input variables by eliminating original input values with variances smaller than a defined value, called threshold. If for example threshold is set to 0.1 OD, only wavelengths with a variance ≤ 0.1 OD over all instances of trainings dataset will be discarded. Therefore, increasing thresholds lead to reduced number of input data for algorithms. The optimal threshold should reduce the amount of input data without affecting algorithm performance.

Multiclass logistic regression (MLR) and *K*-nearest-neighbor (KNN), were applied as classification methods to predict purity of sample based on spectra. The F-measures and accuracies were used to evaluate performance of resulting data models (Figure 28).

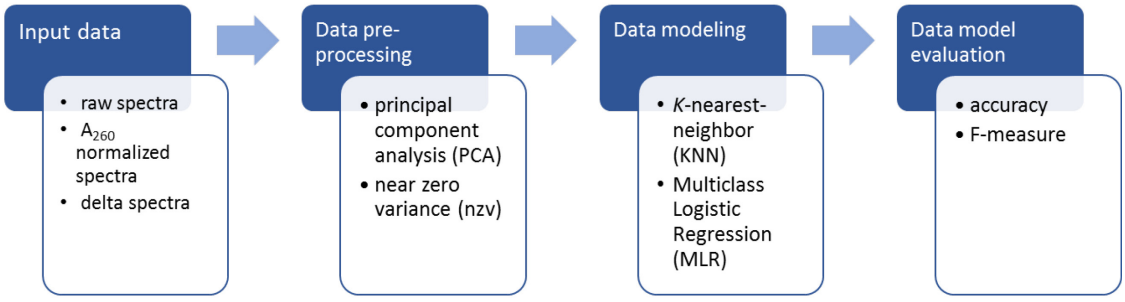


Figure 28: Overview of mathematical data modelling process.

2.5.2.1 Multiclass logistic regression for DNA purity estimation

Logistic regression is a probabilistic classification model, which is commonly utilized for datasets with binary target values. In this study, the goal was to assign absorbance spectra as input data to 1 out of 5 classes, representing DNA sample purity, defined by measured enzyme activities: c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5. For such cases with multiple categorical target values, multiclass logistic regression (MLR) can be applied. The approach used here is called one-over-rest, where the probability for five binary problems

were determined for each observation or input spectrum, where one binary problem had two possible outcomes, for example “class 1” or “not-class 1”.

Before multiclass logistic regression (MLR) was applied to classify measured absorbance spectra to predict enzyme activity, the input data were pre-processed to reduce input features. The unprocessed input data consist of 181 wavelength values or features for each absorbance measurement between A_{230} and A_{410}. The goal of data pre-processing was to reduce the number of features to accelerate algorithm calculation time without losing accuracy of algorithm outcome. Two methods were applied for data per-processing, principal component analysis (PCA) and near zero variance (nzv). Each pre-processing method was optimized for three different versions of input spectra, since the distribution of their OD values varied. For data pre-processing with PCA, default parameters were applied, where several principal components were generated that retained 99, 97.5, 95 or 90% variance of unprocessed input data. The threshold for nzv pre-processing was set to 0, 0.01, 0.02, 0.03, 0,06, 0.1, 0.2, 0.3, 0.6, 1, 2, 3, 6, 10 or 20 OD. With nzv threshold set to 0, input data were used without data pre-processing. Resulting input data with reduced number of features were then submitted to MLR algorithm with default parameters and best results were summarized in Table 1. In addition, all obtained accuracies and F-measures of development datasets were plotted against threshold for nzv or retained variance for PCA (Supplementary Figure 3 and Supplementary Figure 4).

Overall, results showed similar performance for MLR after data pre-processing using nzv or PCA, and slightly better results for raw spectra compared to A_{260} normalized or delta spectra (Table 1). Best results for raw spectra were obtained with accuracy and F-measure of 69% and 43% for nzv threshold at 0.03 OD and 99% variance retained in PCA (Table 1, Supplementary Figure 3, and Supplementary Figure 4). After nzv, 140 of 181 features were retained (Supplementary Figure 3 A), and 6 principal components were generated with PCA (Supplementary Figure 4 A).

For A_{260} normalized spectra, best accuracy and F-measure obtained for nzv pre-processing were 63% and 35% (Table 1 and Supplementary Figure 3 B). Interestingly, these results were obtained for nzv threshold set to 0 OD, were all wavelengths were retained as input features for MLR algorithm, except A_{260}, which was 1 OD for all measurements due to normalization (Supplementary Figure 3 B). With PCA, best MLR performance results were found where 99% or 97.5% of variance were retained, resulting in 6 or 5 principal components, 63% accuracy, and a F-measure of 34% (Table 1 and Supplementary Figure **4** B).

MLR with delta spectra after nzv pre-processing, showed best performance results where nzv threshold was set to 0 OD and all 181 absorbance wavelengths were used as input features. The resulting accuracy and F-measure were 67% and 41% (Table 1 and Supplementary Figure 3 C). When 6 input features for MLR were generated with PCA, containing 99% of variance form original input data, best performance results with PCA pre-processing were obtained with 63% accuracy and a F-measure of 36% (Table 1 and Supplementary Figure 4 C).

Table 1: Summary of multiclass logistic regression performance for development datasets with optimized data pre-processing, using nzv and PCA. MLR with different input spectra were run with non-weighted classes and C = 1 as default settings, after feature reduction with increasing thresholds for nzv or decreasing variance retained in PCA. "t" indicates threshold, and "v" retained variance.

input spectra	pre-processing	accuracy	F-measure
raw spectra	nzv, with t = 0.03 OD	69%	43%
	PCA, with v = 99%	68%	42%
A_{260} normalized spectra	nzv, with t = 0 OD	63%	35%
	PCA, with v = 99% or 97.5%	63%	34%
delta spectra	nzv, with t = 0 OD	67%	41%
	PCA, with v = 99%	63%	36%

In order to further optimize the performance of MLR algorithms, all actual classes of the trainings set were weighted with one (non-weighted) or inversely proportional to class frequencies (balanced). Since the dataset used for this study had more observations assigned to actual class 5 compared to other actual classes, weighting or balancing classes could lead to better MLR algorithm performance, by compensating the imbalance of the input data. In addition, the inverse regularization factor C, in default settings 1.0, was set to 0.01, 0.05, 0,1.0, 0.5, 1, 5, 10, 50, 100. Smaller values specify stronger regularization and avoid overfitting. To find the best parameter settings for MLR algorithms with different input spectra, the algorithms were tested with best obtained nzv pre-processing parameters. Tested MLR parameter settings with best performance results for development datasets were summarized in Table 2.

The results showed, that accuracy and F-measure of MLR algorithms could overall be improved by adjusting regularization strength. The comparison of non-weighted classes in default settings and balanced classes, showed no improvement for balanced classes. MLR algorithms based on raw and A_{260} normalized spectra showed best performance for C = 50, with accuracies of 70% and 65% and F-measures with 43% and 42%. Performance of MLR

algorithm using delta spectra showed highest accuracy and F-measure of 72% and 47% for C = 5 (Table 2).

Table 2: Summary of MLR classification performance for non-weighted and balanced classes with optimized C for development datasets. MLR algorithms with different input spectra were run after feature reduction with optimized nzv thresholds of 0.03 OD for raw, and 0 OD for A_{260} normalized as well as delta spectra. Non-weighted and balanced classes were applied and the inverse regularization factor C was set to 0.01, 0.05, 0,1, 0.5, 1, 5, 10, 50, 100.

input spectra	class	C	accuracy	F-measure
raw spectra	non-weighted	50	70%	43%
	balanced	5	63%	40%
A_{260} normalized spectra	non-weighted	50	65%	42%
	balanced	0.1	54%	41%
delta spectra	non-weighted	5	72%	47%
	balanced	50	62%	49%

Finally, the results obtained from three different input spectra, raw, A_{260} normalized and delta spectra, were compared. Therefore, best thresholds for nzv, non-weighted classes, and best C were applied to test datasets. The results showed accuracies between 68% and 72% for MLR algorithms using test data of different input spectra. Accuracies and confusion matrices were presented in Figure 29.

The results showed overall somewhat better performance results for MLR algorithm using A_{260} normalized spectra, compared to raw or delta spectra. However, none of the three algorithms correctly classified any samples of actual class 2 or 3, whereas most correctly classified samples by all three algorithms belonged to actual class 5. In total, 65 and 58 out of 204 samples were misclassified for raw and A_{260} normalized spectra, while 32 out of 102 samples were misclassified for delta spectra by MLR algorithms (Figure 29).

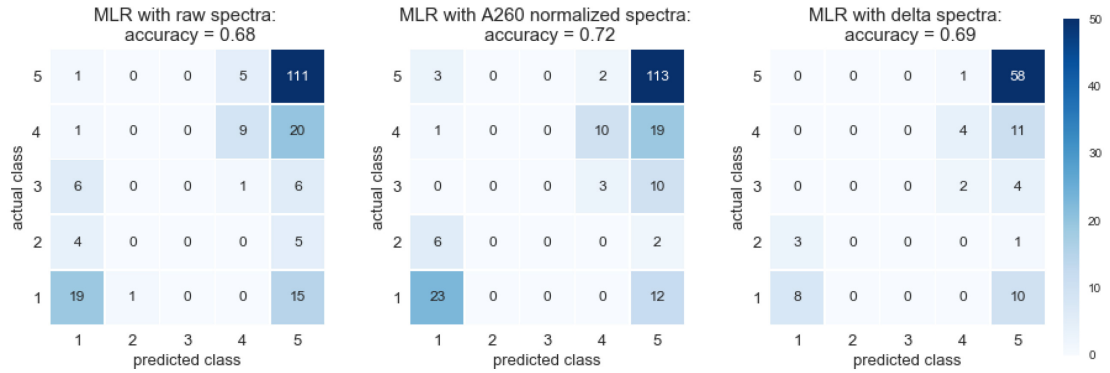


Figure 29: Accuracies and color encoded confusion matrices for MLR algotihms based on (left) raw spectra, (center) A_{260} normalized spectra and (right) delta spectra. Actual classes were created based on enzyme activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, and predicted classes represent outcome of MLR algorithms, which were run using nzv threshold 0.03 OD for raw spectra, or 0 OD for A_{260} normalized and delta spectra, non-weighted classes, and C = 50, 50, or 5, respectively. Color scale of heat map on right.

2.5.2.2 *K*-nearest-neighbor classification for DNA purity estimation

The *K*-nearest-neighbor (KNN) method is a supervised machine learning algorithm for classification. To learn a function for prediction of unknown data, it needs labeled input data. For this thesis, 5 classes based on measured enzyme activities: c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, were used as labels or target values, also called "actual class". The data used to predict these classes were different versions of measured UV/Vis absorbance spectra.

As described in chapter 2.5.2.1 for multiclass regression, input data were pre-processed to reduce input features, before the KNN algorithm was applied to classify measured absorbance spectra to predict enzyme activity. Same parameters were applied for data pre-processing using principal component analysis (PCA) and near zero variance (nzv) and resulting input data with reduced number of features were then submitted to KNN algorithm with default parameters. Obtained results with best accuracies and F-measures of development datasets were summarized in Table 3. In addition, all obtained accuracies and F-measures were plotted against threshold for nzv (Supplementary Figure 1) or retained variance for PCA (Supplementary Figure 2).

For raw spectra, best F-measure and accuracy with 67% and 81% for development dataset were obtained when threshold for nzv was set to 0.1 OD (Supplementary Figure 1 A). Instead of 181 wavelengths between A_{230} and A_{410}, only 93 wavelength values of raw spectra were retained as input data after all wavelengths with variance ≤ 0.1 OD were discarded (Supplementary Figure 1 A). With PCA, best F-measure and accuracy with 65% and 80%

were obtained when 99% of variance from unprocessed data were retained, resulting in 4 principal components as KNN input features (Supplementary Figure 2 A).

A threshold of 0.03 OD led to 56 input features and best performance results for KNN algorithm using development dataset with A_{260} normalized spectra with F-measure at 78% and an accuracy of 85% (Supplementary Figure 1 B). For data pre-processing with PCA, best performance results were obtained when 3 principal components were created retaining 99% variance on unprocessed data, leading to F-measure and accuracy of 72% and 83% (Supplementary Figure 2 B).

In comparison, performance of KNN algorithm using delta spectra was inferior to results of KNN algorithms using raw or A_{260} normalized spectra. The best performance for development dataset was obtained at threshold 0.02 OD, when 23 of 181 wavelength values were discarded from input data, resulting in F-measure and accuracy of 66% and 75%, respectively (Supplementary Figure 1 C). When PCA was applied for feature reduction, highest F-measure and accuracy with 60% and 72% were obtained when 3 principal components were created retaining 97.5% or 95% variance of unprocessed delta spectra (Supplementary Figure 2 C).

Overall, somewhat higher accuracies and F-measures were obtained for KNN algorithm when using nzv for data pre-processing compared to PCA (Table 3). Therefore, feature reduction with PCA was excluded from further KNN algorithm testing.

Table 3: Summary of KNN classification performance for development datasets with optimized data pre-processing, using nzv and PCA. KNN algorithms with different input spectra were run with non-weighted distances and k = 5 as default settings, after feature reduction with increasing thresholds for nzv or decreasing variance retained in PCA. "t" indicates threshold, and "v" retained variance.

input spectra	pre-processing	accuracy	F-measure
raw spectra	nzv, with t = 0.1 OD	81%	67%
	PCA, with v = 99%	80%	65%
A_{260} normalized spectra	nzv, with t = 0.03 OD	85%	78%
	PCA, with v = 99%	83%	72%
delta spectra	nzv, with t = 0.02 OD	75%	66%
	PCA, with v = 97.5% or 95%	72%	60%

To further optimize the outcome of a KNN algorithm for presented input spectra, the algorithms were tested with different neighbor counts as well as with weighted and non-weighted distances. The distances between and unknown sample and a classified sample were calculated as Euclidean distance (equation (19)). For non-weighted distance, only the calculated Euclidean distances were used, whereas the order of distances for

classified neighbors was considered for weighted distances, giving higher influence to closer neighbors compared to more distant neighbors (see chapter 4.2.26). The "k" in a KNN algorithm is the number of classified neighbors used to classify an unknown sample. To find the best neighbor count for KNN algorithms with different input spectra, the algorithms were tested with best obtained nzv pre-processing parameters, and weighted or non-weighted distances, using k in range of 1 to 15. Tested KNN parameter settings with best performance results for development datasets were summarized in Table 4.

Overall, by adjusting the neighbor count and comparing non-weighted and weighted distances, accuracy and F-measure could be improved for KNN algorithms with different input spectra. KNN algorithms with raw and A_{260} normalized spectra showed better performance with weighted distances. Highest accuracy and F-measure with 84% and 71% for KNN with raw spectra was obtained for k = 4. KNN algorithm using A_{260} normalized spectra showed overall best performance with accuracy of 88% and F-measure of 82% with k = 6. Performance of KNN algorithm using delta spectra showed highest accuracy and F-measure of 81% and 76% for weighed and non-weighted distances and k = 1 or 2, respectively (Table 4).

Table 4: Summary of KNN classification performance for non-weighted and weighted distances with optimized neighbor count for development datasets. KNN algorithms with different input spectra were run after feature reduction with optimized nzv thresholds of 0.1 OD for raw, 0.03 OD for A_{260} normalized and 0.02 OD for delta spectra. Non-weighted and weighted distances were applied and k was set from 1 through 15.

input spectra	distance	neighbor count	accuracy	F-measure
raw spectra	non-weighted	3	82%	70%
	weighted	4	84%	71%
A_{260} normalized spectra	non-weighted	4	87%	82%
	weighted	6	88%	82%
delta spectra	non-weighted	1	81%	76%
	weighted	2	81%	76%

To compare results obtained from three different input spectra, best thresholds for nzv, weighted distances, and best k were applied to test datasets. Results for KNN algorithms, showed accuracies between 75% and 89% for test data using different input spectra. Accuracies and confusion matrices showed better results for KNN data models using raw or A_{260} normalized spectra, compared to delta spectra (Figure 30).

When using raw spectra as input values for KNN, 29 of 204 instances were misclassified, and for 4 instances actual and predicted class varied more than 1 class from each other (Figure 30 left). For KNN data model trained and tested with A_{260} normalized spectra, 23

instances were misclassified and 5 of 204 instances differed more than 1 class from actual value (Figure 30 center). Using delta spectra as input values for KNN data model led to overall less accurate results, with 25 of 102 instances classified too low or too high. Of misclassified instances, 8 differed more than 1 class from actual value (Figure 30 right).

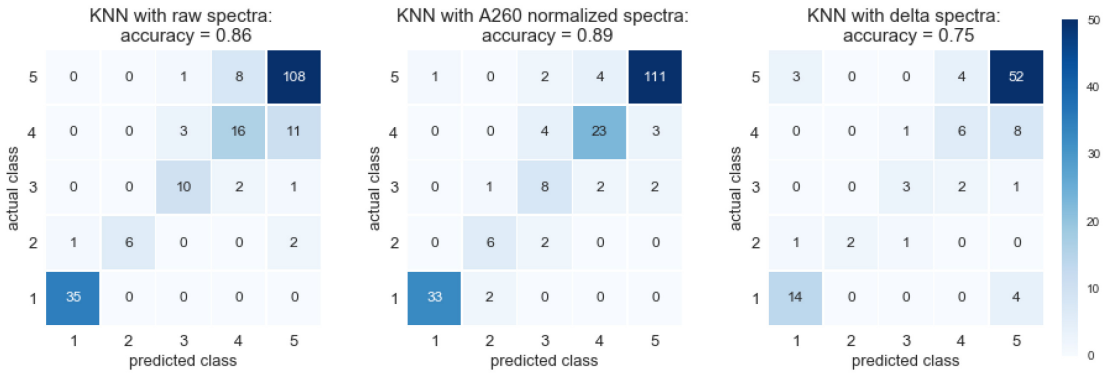


Figure 30: Accuracies and color encoded confusion matrices for KNN data models based on (left) raw spectra, (center) A_{260} normalized spectra and (right) delta spectra. Actual classes were created based on enzyme activities [%], with c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, and predicted classes represent outcome of KNN algorithms, which were run using nzv threshold 0.1 OD for raw spectra, 0.03 OD for A_{260} normalized spectra, or 0.02 OD for delta spectra, weighted distances, and k = 4, 6, or 1, respectively. Color scale of heat map on right.

2.6 *K*-nearest-neighbor algorithm testing

Of applied methods, the KNN algorithm led to best results for classification of DNA absorbance spectra according to sample purity. To evaluate the usefulness of the established data model, additional test datasets were recorded and spectra were classified using the KNN algorithm.

2.6.1 Classification of pure DNA samples with varying concentration

In order to test whether KNN data models trained and tested with three different input spectra were able to correctly classify pure DNA samples with varying concentrations, since DNA was applied at 30 ng/µL of 45 ng/µL for all measurements of enzyme activity dataset, a DNA dilution test dataset was recorded and classified with KNN algorithms. The test dataset consisted of 30 spectra obtained from a dilution series of 10 decreasing concentrations of calf thymus DNA. Raw, A_{260} normalized, and delta spectra were determined for DNA dilution test dataset and presented in Figure 31 to demonstrate the differences between different input spectra for KNN algorithms. Raw spectra of DNA

dilution series showed increased peak at A_{260} with increasing concentration (Figure 31 A). The A_{260} peak for A_{260} normalized spectra was 1 OD for all samples. However, for lower concentrations, the A_{260} normalized absorbance below and above A_{260} increased with decreasing DNA concentration (Figure 31 B). As expected, delta spectra showed no absorbance, independent of DNA concentration (Figure 31 C).

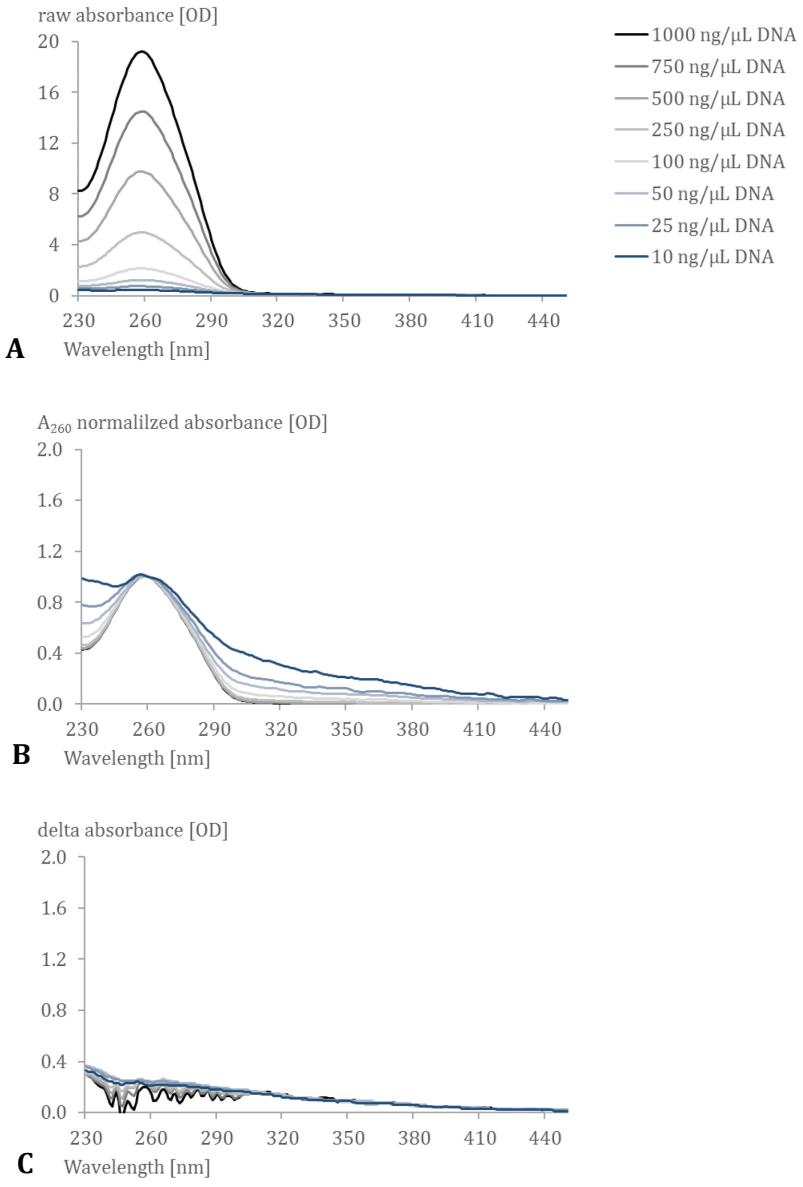


Figure 31: (A) Raw, (B) A260 normalized, and (C) delta spectra of DNA dilution series. Presented were 1 representative measurement replicate of 8 decreasing DNA concentrations.

Raw, A_{260} normalized, and delta spectra were then used to test the corresponding KNN algorithm and predicted classes were plotted against DNA concentration obtained from QIAxpert DNA QIAsymphony application (Figure 32). Since pure DNA was used for all measurements, all samples were expected to be in class 5, defined as samples containing impurities that inhibit enzyme activities by ≤ 20% (see chapter 2.5). Results showed that,

as expected, KNN algorithms trained with raw classified all or all but 2 samples containing ≤ 50 ng/µL DNA as class 5, and classes decreased with increasing DNA concentrations. As seen in Figure 31 A, the shape of raw spectra changes with increasing DNA concentration. Interestingly, results for classification of pure DNA with KNN based on A_{260} normalized spectra were similar to those of raw spectra; DNA samples with high DNA concentrations were assigned to lower classes (Figure 32), although A_{260} normalized spectra of DNA samples with high concentrations had typical shape of pure DNA (Figure 31 B). By comparing A_{260} normalized spectra of DNA dilution series to those of enzyme assay trainings dataset, it was found that DNA samples contaminated with dNTPs had similar shapes: with decreasing concentrations of dNTPs in DNA samples, absorbance below and above A_{260} increased, whereas high concentrations of dNTPs in DNA samples were comparable to high DNA concentrations of DNA dilution series (Supplementary Figure 5). Results of enzyme activity assays had shown that increasing concentrations of dNTPs in DNA samples led to reduced enzyme activity and thus to decreasing actual classes (Figure 10, Figure 18, and Supplementary Figure 5). The KNN model trained and tested with delta spectra, recognized all but 4 DNA samples as class 5, independent of their concentration (Figure 32).

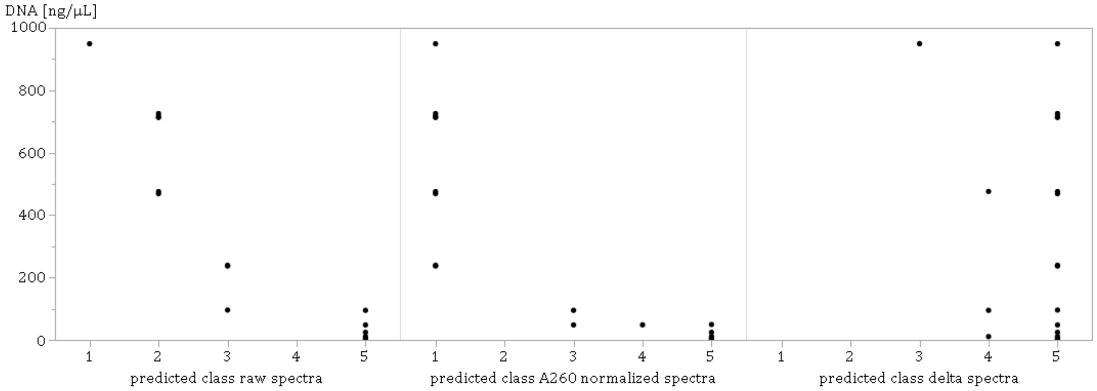


Figure 32: Predicted class of pure DNA samples obtained with KNN algorithms plotted against DNA concentration. KNN algorithms were run using near zero variance threshold 0.1 OD for raw spectra, 0.03 OD for A_{260} normalized spectra, or 0.02 OD for delta spectra for feature reduction, weighted distances, and k = 4, 6, or 1, respectively. Higher predicted classes indicate better DNA sample purity. DNA concentrations were obtained from DNA QIAsymphony application on QIAxpert.

2.6.2 Classification of qPCR samples and correlation with qPCR results

To investigate, whether novel method for DNA purity assessment correlated with the outcome of a downstream application, a qPCR was performed using DNA from human saliva samples showing impurities in absorbance spectra. Absorbance spectra were recorded, classified with KNN algorithm to assess DNA purity and compared to outcome of qPCR.

To record absorbance spectra of saliva DNA samples the DNA QIAsymphony application on QIAxpert was used and raw, A_{260} normalized and delta spectra were determined for all measurements. In Figure 33, 4 representative raw absorbance spectra were displayed, showing different DNA concentrations and high absorbance values at 230 nm wavelength, indicating sample impurities (Figure 33 black and grey spectra, and ref. [64]–[66]). For comparison, an absorbance spectrum of clean calf thymus DNA was added, showing expected DNA absorbance, with lower absorbance at A_{230} and an absorbance peak at A_{260} (Figure 33 red spectrum).

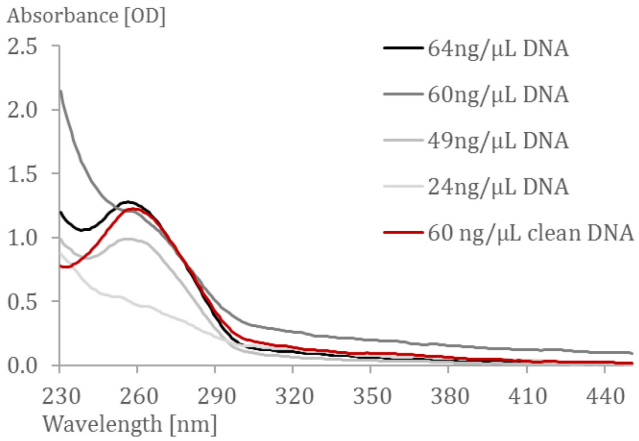


Figure 33: Raw absorbance spectra of a clean calf thymus DNA sample and four representative DNA samples from human saliva. DNA concentrations in legend were calculated by multiplying absorbance at A_{260} of raw spectra with 50.

In addition, the absorbance ratios A_{260}/A_{280} and A_{260}/A_{230}, currently used to assess DNA purity, were plotted for each sample (Figure 34). DNA is considered pure, when A_{260}/A_{280} values are between 1.8 and 2.0, and A_{260}/A_{230} ratios range from 1.8 to 2.2 [64]. The results showed poor DNA purity for all saliva DNA samples, with A_{260}/A_{280} values ranging from 1.1 to 1.8, with only 1 measurement replicate of 1 sample at 1.8, and A_{260}/A_{230} ratios between 0.4 and 1.1 (Figure 34).

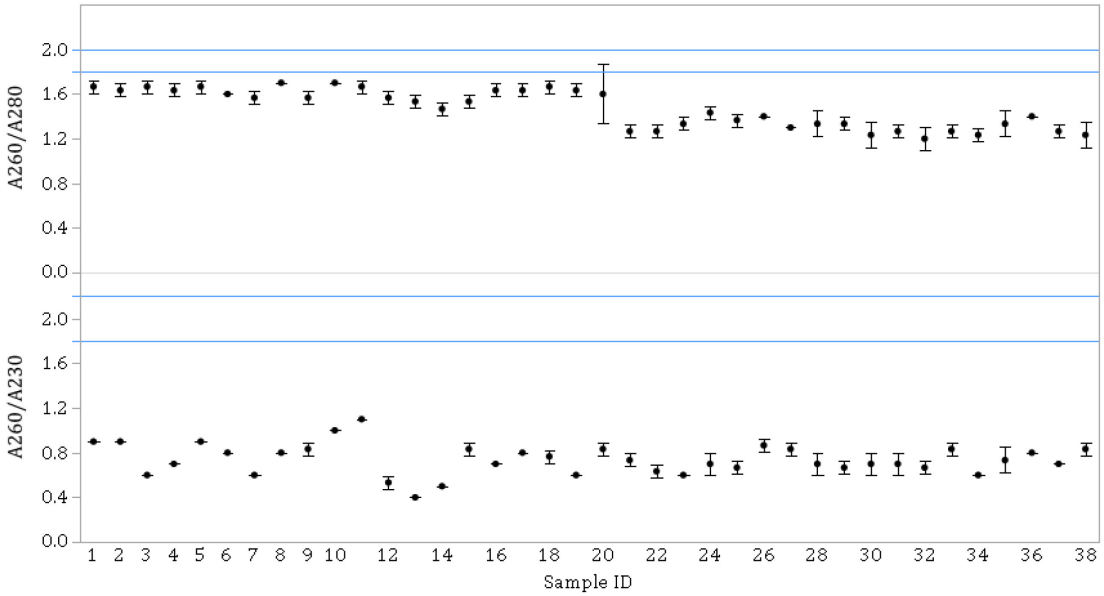


Figure 34: Absorbance ratios (top) A_{260}/A_{280} and (bottom) A_{260}/A_{230} for each saliva DNA sample. Sample IDs were assigned to saliva DNA samples without specific order. Absorbance ratios were calculated from raw spectra. Presented were mean values or 3 measurement replicates with error bar = standard deviation. Blue lines indicate range of pure DNA.

After absorbance measurement and determination of DNA concentrations, all samples were submitted to a qPCR targeting the human *β-actin* gene. Therefore, 10 ng DNA per PCR reaction were applied in triplicates of each sample. For relative quantification of DNA with qPCR, a standard curve with 0.1, 1, 10 or 100 ng DNA per PCR reaction was applied on same qPCR run. The hypothesis was that sample impurities would lead to inhibition of qPCR reaction, resulting in lower DNA concentrations detected by qPCR, although 10 ng DNA per reaction were applied for all samples. Detection of lower DNA concentration would be indicated by higher delta Cq values. To determine delta Cq values, the mean Cq values of standard with 10 ng DNA per reaction, representing clean DNA, was subtracted from mean Cq value of each sample. The delta Cq values for each sample were plotted in Figure 35 and showed variability between samples, indicating that sample impurities seen in absorbance spectra indeed led to inhibition of qPCR and lower DNA concentrations detected by qPCR.

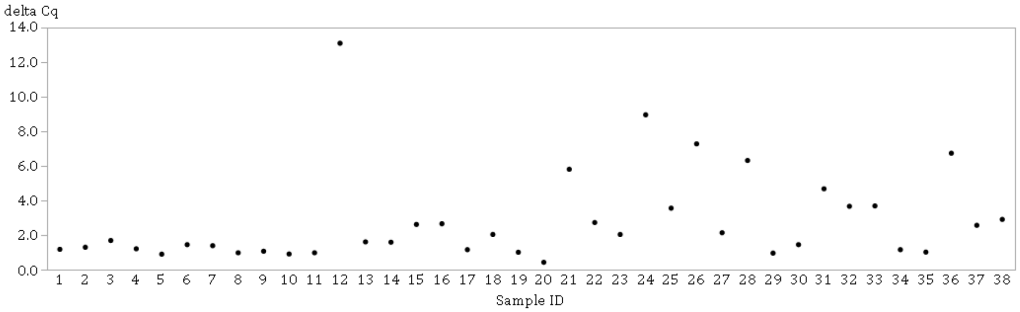


Figure 35: Delta Cq values of saliva DNA samples obtained with qPCR. Sample IDs were assigned to saliva DNA samples without specific order. Delta Cq values were obtained by subtracting mean Cq value of standard with 10 ng DNA per reaction, representing clean DNA, from mean Cq values obtained for each sample.

In order to test whether novel method for DNA purity assessment using KNN algorithm to classify DNA absorbance spectra according to sample purity, raw, A_{260} normalized and delta spectra of all absorbance measurements from saliva DNA samples were used to test the corresponding KNN algorithm. The class predicted by KNN algorithm for each absorbance spectrum was then plotted against delta Cq value of corresponding sample (Figure 36). Saliva DNA samples with high delta Cq values were expected to be assigned to a low purity class, while samples with low delta Cq values were expected to be assigned to a high purity class.

The results showed that classes predicted by KNN algorithm, describing purity of DNA samples, failed to correlate with delta Cq values obtained from qPCR. For purity classes predicted by KNN algorithms based on raw or delta spectra, all classes contained high and low delta Cq values. When A_{260} normalized spectra were used to predict purity classes, most samples were in class 5, considered pure DNA, independent of their delta Cq values (Figure 36).

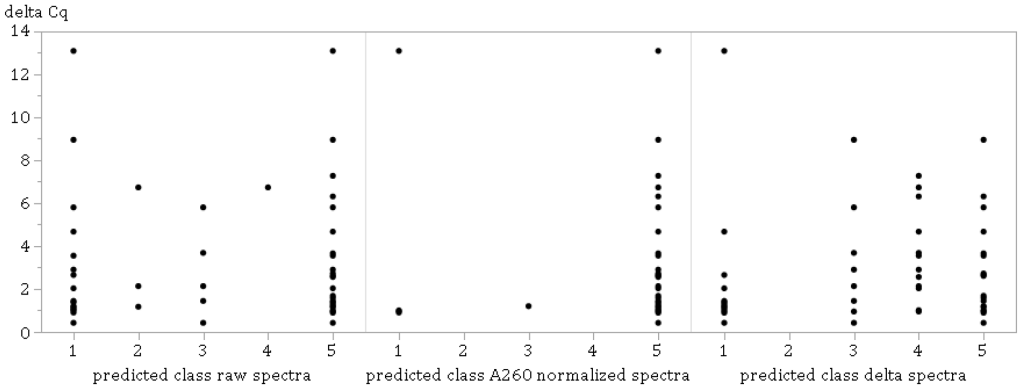


Figure 36: Predicted class of saliva DNA samples obtained with KNN algorithms plotted against delta Cq values. KNN algorithms were run using near zero variance threshold 0.1 OD for raw spectra, 0.03 OD for A_{260} normalized spectra, or 0.02 OD for delta spectra for data pre-processing, weighted distances, and k = 4, 6, or 1, respectively. Higher predicted classes indicate better DNA sample purity. Delta Cq values were obtained by subtracting mean Cq value of standard with 10 ng DNA per reaction, representing clean DNA, from mean Cq values obtained for each sample.

To compare the usefulness of novel method with current method for DNA purity assessment, delta Cq values were plotted against classes based on absorbance spectra (Figure 37). Classes based on absorbance spectra were generated as described in chapter 2.5, with c1 ≤ A260/A280 or A260/A230 ratio 0.6 < c2 ≤ A260/A280 or A260/A230 ratio 1.1 < c3 ≤ A260/A280 or A260/A230 ratio 1.5 < c4 ≤ A260/A280 or A260/A230 ratio 1.8 < c5, using raw spectra to calculate absorbance ratios. Similar to classes based on KNN algorithms, higher delta Cq values were expected to have lower classes and vice versa.

The results showed no correlation between classes based on absorbance spectra and delta Cq values. All but 3 samples were in class 2, since their A_{260}/A_{230} ratios were between 0.6 and 1.1 (Figure 37 and Figure 34).

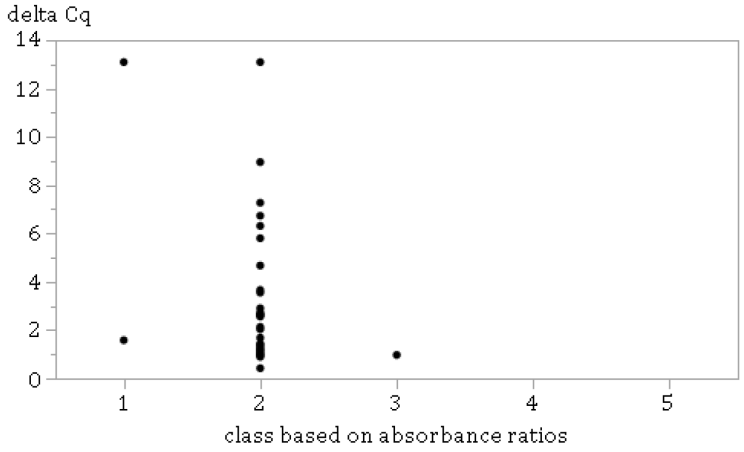


Figure 37: Classes based on absorbance spectra of saliva DNA samples plotted against delta Cq values. Absorbance ratios were calculated from raw spectra and classes of absorbance ratios were generated as described in chapter 2.5, with c1 ≤ ratios 0.6 < c2 ≤ ratios 1.1 < c3 ≤ ratios 1.5 < c4 ≤ ratios 1.8 < c5. Delta Cq values were obtained by subtracting mean Cq value of standard with 10 ng DNA per reaction, representing clean DNA, from mean Cq values obtained for each sample.

3 Discussion

The goal of this thesis was to develop a novel method for purity assessment of DNA, using absorbance spectra to predict whether a DNA sample was suitable for a downstream application. Since all downstream applications consist at least partially of enzyme reactions, the influence of possible DNA sample impurities on absorbance spectra and enzyme activities were recorded. Therefore, (1) enzyme activity assays to measure polymerase, ligase, and kinase activity were established, DNA was contaminated with possible impurities, absorbance spectra of contaminated DNA were recorded for DNA concentration estimation and data modelling, and contaminated DNA samples were applied on enzyme activity assays, using previously determined DNA concentration to adjust DNA template volume on assays. (2) Recorded data were used to train and test different algorithms to predict enzyme activity based on absorbance spectra, and (3) the usefulness of novel method for purity assessment of DNA was tested with qPCR data.

3.1 Enzyme inhibition by nucleic acid sample impurities

In order to measure the influence of possible contaminants on enzyme activities, polymerase, ligase, and kinase activity assays were established. However, high variability of radiometric kinase assay results and irreproducible standard curves, led to exclusion of T4 PNK activity data from further analysis and data modelling. Higher concentrations of [γ-^{32}P]ATP or lower concentrations of unlabeled ATP might have led to reproducible results, but due to lack of time this could not be tested. The effect of possible contaminants on Taq polymerase and T4 DNA ligase activities were analyzed, before obtained data were applied to train and test different algorithms to predict enzyme activity based on absorbance spectra. Results showed that of twelve applied contaminants, some had no influence on Taq polymerase and T4 DNA ligase activity, whereas others had an inhibiting or enhancing effect on either one or both enzymes.

As expected, betaine and DTT had no influence on polymerase activity. Betaine is a PCR additive denaturing secondary structures of DNA and therefore making the template more accessible for polymerase [92], while DTT is used to stabilize enzymes in enzyme solutions and reaction mixtures [96].

Surprisingly, dNTPs, added to amplification reactions as single DNA building blocks, led to lower polymerase activity with increasing concentrations. Part of this effect could be due to an imbalance of the reaction or oversaturation of dNTPs in reaction mix, but the larger part

was probably due to low DNA template input applied to reaction because of an overestimation of the DNA concentrations measured on QIAxpert. As described in sections 2.1.2 and 4.2.9 concentration of contaminated DNA was measured on QIAxpert and used to calculate volume of DNA applied on enzyme activity assays. Since dNTPs are the building blocks of DNA, the SCP algorithm of DNA QIAsymphony App is not able to distinguish between enzyme activity assay DNA template and dNTPs, leading to higher DNA concentration results for DNA samples contaminated with dNTPs. To circumvent the influence of DNA overestimation on enzyme activities, only contaminants with reference spectra in SCP algorithm could be used, either by testing only those possible contaminants for which reference spectra are already present in SCP algorithm, or by adding reference spectra of more contaminants to SCP algorithm. Another option could be to drop DNA concentration estimation of contaminated DNA samples and use target concentration of samples to calculate DNA input volume in enzyme activity assays; or to determine DNA concentration with another method such as fluorescence dyes that specifically bind nucleic acids, and therefore perform robustly in the presence of DNA sample impurities [101], [102].

EDTA, hemoglobin, HSA, and IgG all led to reduced polymerase activity with increasing concentrations. EDTA is a chelator, binding metal ions like Mg^{2+}, which is an essential cofactor for Taq polymerase. Thus inactivation of polymerase by EDTA was expected, similar to hemoglobin, which is also known to inhibit PCR reaction by interaction with the polymerase cofactor [14]. The inhibition of polymerase by HSA was unexpected, since bovine serum albumin (BSA), the same protein from bovine is a commonly added to PCR reactions for stabilization [103], and was probably caused by lower template input due to higher DNA concentrations measured on QIAxpert. The DNA QIAsymphony application does not recognize proteins as impurities, therefore the absorbance of proteins at A_{260} leads to higher DNA concentrations estimated by the SCP algorithm. IgG led to stronger inhibition of polymerase compared to HSA; partially it was probably due to overestimation of DNA concentration by QIAxpert as explained for HSA, but IgG is also known to inhibit PCR by binding ssDNA and hindering polymerase access to template DNA [14].

Interestingly, sodium azide (SA) and guanidine isothiocyanate (GITC) both led to about 20% inhibition of polymerase at highest applied concentration, while SA was expected to have no inhibiting effect on polymerase, whereas GITC was expected to lead to complete enzyme inhibition. However, highest concentration of SA applied in Phi-Inhibition-Assay reaction, was 0.05% and thus higher than the common concentration of 0.04%, in elution buffers [104]. GITC on the other hand was applied at a highest concentration of 10 mM in enzyme

assays due to its high absorbance, although it's usually applied in much higher concentrations of 4 M in lysis buffers for denaturation of nucleases during purification [97].

The polymerase inhibition observed for glycogen, was probably partially due to overestimation of DNA concentration and partially due to glycogen binding to DNA template [98]. As described for proteins above, the SCP algorithm of DNA QIAsymphony application does not recognize glycogen as impurity. Similar results were obtained for phenol; increasing concentrations of phenol led to reduced polymerase activity, although it has been shown, that inhibition of PCR reactions by phenol starts at a concentration above 0.01% [12], [13]. Sodium citrate (SC) led to complete inhibition of polymerase at all concentrations. It neutralizes the negative charge of DNA and lead to dissociation of DNA from water [90], therefore the template would not be accessible for polymerases.

In T4 DNA ligase activity assay, EDTA showed no influence on enzyme activity, since T4 DNA ligase is independent of metal ions as cofactor. Interestingly, betaine led to improved ligase activity at highest concentration. Betaine has a denaturing effect on DNA [92] and short non-ligated dsDNA fragments separate into single strands faster compared to long ligated dsDNA fragments. Ethidium bromide used for detection in gel electrophoresis specifically binds dsDNA. Therefore, denaturation of short non-ligated dsDNA fragments by betaine could result in lower detection of non-ligated compared to ligated dsDNA fragments, indicating higher ligase activity. Hemoglobin and HSA also had positive effects on T4 DNA ligase activity. Possibly these proteins have a stabilizing effect on ligation reaction, comparable to that of BSA on PCR [103].

Results for influence of sodium azide (SA), guanidine isothiocyanate (GITC), glycogen, IgG, dNTPs, sodium citrate (SC) and phenol on ligase were similar to those on polymerase, as described above. DTT is used to stabilize T4 DNA ligase and reduce its activity in storage solution and led to reduced ligase activity with increasing concentration.

In conclusion, observed influences of possible contaminants on enzyme activities were due to either direct interaction of contaminant with enzymes or DNA in enzyme activity assays, or influence of contaminant on DNA concentration estimation.

3.2 Application of mathematical data modelling

To train and test different algorithms for novel method of purity assessment for DNA samples, recorded absorbance spectra and enzyme activities were used. Absorbance spectra served as input variables to predict enzyme activities as target values. Based on

obtained enzyme activities, five classes were created, where higher classes represented higher enzyme activities indicating higher DNA purity. Instead of two classes or a simple "pure" or "not pure" classification, five classes were chosen to enable distinction between downstream applications that are to a greater or lesser extend susceptible to DNA sample impurities. It has for example been shown that digital PCR (dPCR) less affected by inhibitors that might be present nucleic acid samples compared to qPCR [105], [106], therefore reliable results for a DNA sample with a purity of class 4 might still be obtained for dPCR but not for qPCR.

At first, actual classes based on measured enzyme activities were plotted against absorbance ratios, to investigate whether the current method for nucleic acid purity assessment correlated with the results of enzyme activity assays. The results showed no correlation between actual classes and absorbance ratios, indicating that absorbance ratios are insufficient for nucleic acid purity evaluation. Although their validity has been controversially discussed over half a century [62], [65], [68], [69], and various recent publications have shown that they fail to correlate with downstream assay success [12], [17], they are still the standard method for purity assessment of nucleic acid samples. Absorbance ratios use 3 out of over 100 wavelengths recorded in a UV/Vis absorbance measurement, A_{230}, A_{260} and A_{280}, to estimate nucleic acid sample purity. Therefore, a lot of information contained in UV/Vis absorbance spectra remains unused.

The goal of this thesis was to develop a novel method for nucleic acid purity assessment, using all wavelengths between A_{230} and A_{410} of absorbance spectra to estimate purity of DNA samples. Multiclass logistic regression and *K*-nearest-neighbor were applied to classify absorbance spectra of contaminated DNA samples according to enzyme activity measured for same samples. Overall performance of KNN algorithm was better and less samples of test dataset were misclassified when using KNN compared to MLR.

Reducing the number of input features by near zero variance or PCA and adjusting regularization strength, led to improved MLR algorithm performance. However, a large number of samples were misclassified by MLR and none of the samples in actual class 2 or 3 were recognized as such, indicating that MLR algorithm with applied parameter settings was probably not the appropriate mathematical approach to create a novel method for DNA purity assessment, as described in this study.

Performance of KNN algorithm was improved by reduction of input features, applying weighted distances and adjusting neighbor count used for classification of unknown samples. After KNN optimization, the algorithm was applied to three test datasets: the enzyme activity test dataset, a dilution series of pure DNA samples, and saliva DNA samples

applied on qPCR. As input data, three different versions of absorbance spectra were applied: raw, A_{260} normalized, and delta spectra. Delta spectra represented absorbance spectra of sample impurities only. Performance of KNN algorithm varied for different test datasets and input spectra. For enzyme activity test dataset, best results were obtained with A_{260} normalized spectra, whereas delta spectra led to lowest KNN performance measures. Results for enzyme activity test datasets were promising with accuracies between 75% and 89%, when delta or A_{260} normalized spectra were used as input data for KNN algorithms. In other words, between 25% and 11% of samples were misclassified. Compared to current purity evaluation with absorbance ratios, where 83% of samples were misclassified, this was a notable improvement.

Inferior performance results for enzyme activity test dataset using delta spectra compared to raw or A_{260} normalized spectra, could be due to variation from spectral content profiling (SCP) algorithm of DNA QIAsymphony application on QIAxpert. Absorbance spectra of nucleic acids were detected in a contaminated DNA sample by SCP. Delta spectra were obtained by subtracting nucleic acid spectra from raw spectra. To detect absorbance spectra of single components of a complex sample, SCP algorithm uses reference spectra [78], [79]. Therefore, SCP is only able to accurately identify all components of a complex sample, when reference spectra of all components are available. The DNA QIAsymphony application on QIAxpert however, contains a limited number of reference spectra. Analysis of DNA samples with absorbing contaminants, for which there is no reference spectrum available to the SCP algorithm, leads to inaccurate results with higher or lower DNA concentrations detected in a sample or detection of absent impurities with a similar absorbance spectrum. For enzyme activity assays, several absorbing contaminants were used, for which reference spectra in DNA QIAsymphony application are missing. Therefore, nucleic acid absorbance spectra detected by DNA QIAsymphony application could be inaccurate for various samples, leading to inaccurate and varying delta spectra.

Ideally, novel method for purity assessment of DNA should be able to determine DNA purity over a wide range of DNA concentration. To test the ability of KNN algorithm to recognize pure DNA of varying concentrations as pure, absorbance spectra of a DNA dilution series were recorded and classified with KNN algorithms. However, pure DNA samples applied as controls in enzyme activity assays, used to generate data for algorithm training, were applied in 45 or 30 ng/µL. Consequently, as expected, the KNN algorithm based on raw spectra was unable to recognize pure DNA with higher concentrations, since the shape of absorbance spectra of high DNA concentrations varied from those used as clean DNA controls in trainings dataset. To overcome this issue without recording more or different

trainings data, A_{260} normalized and delta spectra were calculated from raw spectra and used as input data for KNN algorithm.

The absorbance at A_{260} is used to calculate DNA concentration from an absorbance measurement, since DNA has an absorbance peak at this wavelength. A_{260} normalized spectra had an absorbance of 1 OD at 260 nm wavelength, independent of DNA concentration. Therefore, KNN algorithm based on A_{260} normalized spectra was expected to recognize pure DNA of varying concentrations. Interestingly, results showed that samples with higher DNA concentrations were assigned to lower purity classes by KNN algorithm. The similarity of A_{260} normalized spectra from DNA dilution series to DNA contaminated with dNTPs, used for enzyme activity assays, and resulting lower enzyme activities obtained for those samples could explain these results.

Delta spectra, as described above, were obtained by subtracting absorbance spectra of nucleic acids, detected in a contaminated DNA sample by SCP, from raw spectra. Thus, delta spectra represented absorbance of impurities only and KNN algorithm trained with delta spectra should be independent of DNA concentration. Since control DNA samples and DNA dilution series consisted of pure DNA samples, SCP results of DNA QIAsymphony should be accurate. As expected, classification of pure DNA samples with increasing concentrations using KNN based on delta spectra, resulted in recognition of all but 4 DNA samples as class 5, representing pure DNA.

The aim of this study was to develop a novel method for purity assessment of DNA that correlates with outcome of downstream applications. To investigate whether this goal was achieved and to test the usefulness of the KNN algorithms, absorbance spectra of DNA samples containing impurities were recorded and DNA samples were applied on qPCR. Subsequently, the absorbance spectra were classified with KNN algorithms to predict DNA purity and compared to delta Cq values obtained from qPCR. The results showed no correlation between KNN and qPCR outcome, and failed to prove usefulness of novel method for purity assessment of DNA samples developed in this thesis. DNA samples extracted from human saliva in class 5 predicted by KNN were expected to lead to low delta Cq values, whereas samples in class 1 were expected to inhibit qPCR and lead to high delta Cq values. Samples with high delta Cq values and high class predicted by KNN could have high delta Cq values due to overestimation of DNA concentration in these samples by DNA QIAsymphony application. Overestimation of DNA concentration could occur due to sample impurities unknown to SCP algorithm and would lead to less DNA template input in qPCR reactions and consequently higher delta Cq values, even if impurities had no inhibiting effect on qPCR. Samples with low delta Cq values and low class predicted by KNN,

could contain impurities that led to inhibition of enzyme activity assays, but had no inhibiting effect on qPCR, due to higher polymerase concentration or stabilizing additives in qPCR reaction mix such as BSA. These results suggest that an algorithm trained and tested with enzyme activity data might be unable to correctly determine purity for qPCR assays. Therefore, novel methods for DNA purity assessment might have to be developed for specific downstream applications, by using data obtained from each downstream application for algorithm training and testing.

In conclusion, by application of mathematical data modelling for development of a novel method for DNA purity assessment, it was found that KNN algorithm can be used to predict enzyme activities based on absorbance spectra with higher accuracy than current method using absorbance ratios. The comparison of three different input spectra for KNN algorithms, using three different test datasets led to inconclusive results. A_{260} normalized spectra led to best KNN performance for enzyme activity test dataset, while KNN based on delta spectra was able to most accurately classify pure DNA with varying concentrations. Interestingly, none of the KNN algorithms trained with enzyme activity data correlated with qPCR outcome.

3.3 Outlook and future perspectives

Using high quality nucleic acids is critical to obtain reliable and reproducible results for modern molecular biological methods, like qPCR and NGS [1]–[4], and a recent nature survey has shown that only 50% of biological research is considered reproducible [107]. Standardization and guidelines for quality control can improve reproducibility [3]. Three key elements of nucleic acids quality control are concentration, integrity and purity. Various methods such as UV/Vis absorbance measurement [62], fluorescence measurement [102], or gel electrophoresis [108] are available for nucleic acid concentration estimation. Nucleic acid integrity used to be determined by visual inspection of slap gel electrophoresis, until the RIN was introduced in 2004, combining digital gel electrophoresis and an algorithm analyzing various regions of the electropherogram to describe RNA integrity [73]. The RIN is an example for successful application of mathematical data modelling to analyze biological data and set new standards.

Nucleic acid purity is currently determined based on UV/Vis absorbance measurements, using the A_{260}/A_{280} and A_{260}/A_{230} ratios. While the validity of absorbance ratios has been controversially discussed over decades [12], [62], [65], [68], [69] and a survey taken at the 2005 London qPCR meeting revealed that only 4% of researchers rely on absorbance ratios

for nucleic acid purity assessment [47], no alternative has been introduced to replace this method. In this study, a novel method for nucleic acid purity assessment was developed, using mathematical data modelling to predict DNA purity based on absorbance spectra between 230 and 410 nm wavelength. Therefore, DNA purity was defined by using measured enzyme activities under the influence of possible nucleic acid contaminants. It was successfully demonstrated that novel method for nucleic acid purity assessment could predict enzyme activities with a higher accuracy compared to currently applied absorbance ratios. Although the algorithm trained and tested in this study failed to correlate with qPCR outcome, this study shows that using mathematical data modelling to analyze absorbance spectra is a promising approach to develop a novel method for nucleic acid purity assessment.

Being able to reliably assess nucleic acid purity would lead to new standards for nucleic acid quality control and could contribute to improving reproducibility in biological science.

4 Materials & Methods

4.1 Materials

4.1.1 Chemicals and reagents

Table 5: Substances collected to be measured as possible contaminants of nucleic acid samples.

Contaminant	CAS No.	Supplier	Catalog / Material No.
Betaine	107-43-7	Sigma-Aldrich	B2754
Cells /Debris	Pellet of frozen Jurkat cells from cell culture		
Chloroform	67-66-3	Merck	102431
Dithiothreitol (DTT)	3483-12-3	Sigma-Aldrich	DTT-RO Roche
DMEM, cell culture medium	-	ThermoFischer	
DNase I	9003-98-9	QIAGEN	79254
dNTP Mix (10 mM each)	-	ThermoFisher	R0191
Ethanol	64-17-5		
Glycerol	56-81-5	Merck	356352
Glycogen	9005-79-2	Roche	10901393001
Guanidine Hydrochloride (GuHCl)	50-01-1	Sigma Aldrich	G 4505 Sigma
HEPES	7365-45-9	Sigma-Aldrich	H4034
Isopropanol	67-63-0		
MOPS free acid	1132-61-2	AppliChem	A2947
Proteinase K	39450-01-6	QIAGEN	19131
RNase A	9001-99-4	QIAGEN	19101
RPMI, cell culture medium	-	ThermoFischer	
Tri-Sodiumcitrate-dihydrate (SC)	6132-04-3	VWR	567446-5
guanidine thiocyanate (GITC)	540-72-7	Merck	106627
ß-Mercaptoethanol	60-24-2	Calbiochem	444203
Sucrose	57-50-1	Sigma Aldrich	S 7903
Trizma® base	77-86-1	Sigma Aldrich	T8524
Urea	57-13-6	Merck	108487
Human Serum Albumin (HSA)	70024-90-7	Sigma-Aldrich	A9511
IgG from bovine serum (IgG)	-	Sigma-Aldrich	I9640
Human Hemoglobin (Hb)	9008-02-0	Sigma-Aldrich	H7379-1G
EDTA	6381-92-6	Sigma-Aldrich	E5134-5KG
Sodium Azide (SA)	26628-22-8	Merck	1066880100
Phenol	108-95-2	Sigma-Aldrich	P1037-25G

Table 6: Reagents applied in enzyme activity assays and to record additional test datasets for algorithm testing. DNA and cells marked with * were obtained as donations from other groups at QIAGEN.

Reagents	Manufacturer	Catalog No.
PhiX174 Virion DNA	NEB	N3023 L
PCR Buffer, 10x	QIAGEN	1005481
Tween 20	QIAGEN	1006170
NP40	QIAGEN	1004973
BSA 20mg/ml	NEB	B9000S
dNTP-Mix, 10 mM	QIAGEN	1005631
EvaGreen 20.000x in DMSO	Biotium	31002
pCMVbeta plasmid*	n/a	n/a
HotStarTaq Master Mix	QIAGEN	1010023
CutSmart® Buffer	NEB	B7204S
10x Ligase reaction buffer	enzymatics	B6030L
Alignment Marker 15bp / 10kb	QIAGEN	929523
Alignment Marker 15bp / 3kb	QIAGEN	929522
Size Marker FX 174/HaeIII	QIAGEN	929551
Jurkat cells*	n/a	n/a
ATP Solution (100 mM)	Sigma	R0441
ATP, [γ-^{32}P]- 6000Ci/mmol	PerkinElmer	NEG502Z250UC
10x T4 Polynucleotide Kinase Buffer	enzymatics	B9040
UltraPure™ Calf Thymus DNA Solution	Thermo Fisher Scientific	15633-019
DNA from human saliva*	n/a	n/a
Buffer TE	QIAGEN	1044246
Buffer EB	QIAGEN	1014612
RNase-free water	QIAGEN	1018017
Ice		

Table 7: Kits applied for preparation of or enzyme activity assays.

Kit	Manufacturer	Catalog No.
QIAxcel DNA High Resolution Kit	QIAGEN	929002
DyeEx 2.0 Spin Kit	QIAGEN	63206
MinElute PCR Purification Kit	QIAGEN	28004
QIAamp DNA Mini QIAcube Kit	QIAGEN	51326
QuantiNova Multiplex PCR Kit	QIAGEN	208452

4.1.2 Enzymes

Table 8: Enzymes applied for preparation of or enzyme activity assays.

Enzyme	Manufacturer	Catalog No.
DNA Taq Polymerase	NEB	M0320
XhoI	NEB	R0146S
T4 DNA Ligase	enzymatics	L6030-LC-L
T4 PNK	enzymatics	Y9040L

4.1.3 Oligonucleotides

Table 9: Primers and probes applied for enzyme activity assays and qPCR.

Name	Manufacturer	Sequence (5'-3')
Primer PhiX174	IDT or Biolegio	ACGACGTTTGGTCAGTTCCATCAACATCATAGC
Fwd pCMVbeta	Biolegio	CGGTTTGACTCACGGGGATT
Rev pCMVbeta	Biolegio	GACCGGCAACGAAAATCACG
Fwd T4 PNK assay	IDT	TGGAGGTGGTAAGGTGAT
Rev T4 PNK assay	IDT	CCAACTTTCTTTCCCTCACAT
Fwd *β-actin* qPCR	IDT	TCACCCACACTGTGCCCATCTACGA
Rev *β-actin* qPCR	IDT	CAGCGGAACCGCTCATTGCCAATGG
TaqMan probe *β-actin* qPCR	IDT	FAM-ATGCCCTCCCCCATGCCATCCTGCG-BHQ1

4.1.4 Consumables

Table 10: Consumables used for enzyme activity assays and additional experiments to record test datasets for algorithm testing.

Consumables	Manufacturer
QIAxpert Slide-40	QIAGEN
MinElute Spin Columns	QIAGEN
QX 0.2 ml 12-Tube Strip Caps	QIAGEN
QX 0.2 ml 12-Tube Strip	QIAGEN
0.2 ml Thin-walled 12 Tube and Domed Cap Strips	Thermo Scientific
Pipette tips 10 µl	Sarstedt
Pipette tips 100 µl	Sarstedt
Pipette tips 1000 µl	GILSON
Pipette tips 20 µl	Eppendorf
Pipette tips 200 µl	GILSON

Consumables	Manufacturer
Rotor Adapters	QIAGEN
RotorGene 4-Strip tubes and caps	QIAGEN
SafeSeal micro tube 1,5 ml	Sarstedt
SafeSeal micro tube 2 ml	Sarstedt
SafeSeal micro tube 5 ml	Eppendorf
Gloves	Unigloves
Spatula	VWR
Scintillation tubes and caps, 18 mL	Beckman Coulter

4.1.5 Instruments

Table 11: Instruments used to record data for enzyme activity assays and additional test datasets for algorithm testing.

Instrument	Manufacturer
QIAxpert	QIAGEN
QIAxcel Advanced	QIAGEN
T100TM Thermal Cycler	BioRad
Centrifuge 5430	eppendorf
Centrifuge 5417C	Eppendorf
Centrifuge 5418R	Eppendorf
Galaxy Ministar	VWR
QIAcube	QIAGEN
QIAgility	QIAGEN
QIAxcel Advanced	QIAGEN
Rotilabo®-Block-Heater H250	Carl Roth
Rotor-Gene Q	QIAGEN
StripSpin 12	Benchmark Scientific
Thermomixer Comfort	Eppendorf
Vortex Genie 1 Touch-Mixer	Scientific Industries
Vortex Genie 2	Scientific Industries
Beckman LS 6500	Beckman Coulter
Clean Spot PCR Workstation	MIDSCI
Excellence Plus scale	Mettler Toledo
LCexv 4010 refrigerator	Liebherr
Premium NoFrost freezer	Liebherr

4.1.6 Software and online tools

Table 12: Software tools applied for this study.

Software	Version	Provider
QIAxpert Software	2.2.0.21 or higher	QIAGEN
DropQuant	1.5.0	Trinean N.V.
cDrop	1.3.0 or 3.1.0.89	Trinean N.V.
Rotor-Gene Q Series Software	2.3.1 (Build 49)	QIAGEN
Q-Rex	1.0.0 or higher	QIAGEN
QIAxcel ScreenGel	1.5	QIAGEN
MS Excel	2016	Microsoft
JMP	12.1.0 or higher	SAS Institute Inc.
Python	3.6.1 (Anaconda3 4.4.0)	Continuum Analytics, Inc.
MS Word	2016	Microsoft
Zotero	5.0.61	Corporation for Digital Scholarship

Online tools:

- NCBI Primer BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast/
- NCBI Nucleotide BLAST: https://blast.ncbi.nlm.nih.gov/Blast.cgi
- Multiple Primer Analyzer : https://www.thermofisher.com/de/de/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html
- Find unique restriction sites: www.addgene.org

4.1.7 Manuals

Table 13: Manuals and handboods applied for this study.

Manual	Edition	Publisher
DyeExTM Handbook	May 2002	QIAGEN
JMP® Software: ANOVA and Regression	04Apr2014	SAS Institute Inc.
JMP® Software: Data Exploration	04Feb2014	SAS Institute Inc.
MinElute Handbook	03/2008	QIAGEN
QIAxpert® User Manual	12/2015	QIAGEN
QIAxcel DNA Handbook	November 2014	QIAGEN
Scintillation System LS 6500	Manual 247971	Beckman Coulter

4.2 Methods

4.2.1 Preparation of possible contaminants

Substances were collected in powdered form and solved in RNase-free water to obtain stock concentrations, or obtained in liquid form. Stock concentrations of powdered substances were prepared in 1 mL final volumes by adding RNase-free water to the amount of substance listed in Table 14. To record absorbance spectra of possible contaminants, stock concentrations were either applied directly to measurement chip or diluted in RNase-free water.

Table 14: Concentrations of possible contaminants prepared for first or second measurement. All substances marked with a * were obtained in liquid form.

Contaminant	Added to 1 mL final volume	Stock concentration	Measurement concentration
Betaine	1000 mg	1 g/mL	100 mg/mL
Cells /Debris[1]	-	n/a	same as stock
Chloroform *	1 mL	unknown	same as stock
Dithiothreitol *	1 mL	1 M	100 mM
cell culture medium DMEM *	1 mL	n/a	same as stock
DNase I *	1 mL	unknown	same as stock
dNTPs*	1 mL	10 mM	2 mM
Ethano l*	0.1 mL	100%	10%
Glycerol *	0.1 mL	99%	9.9%
Glycogen *	1 mL	20 mg/mL	same as stock
Guanidine Hydrochloride (GuHCl)	190 mg	2 M	same as stock
HEPES	238 mg	1 M	same as stock
Isopropanol *	1 mL	100%	10%
MOPS free acid	210 mg	1 M	same as stock
Proteinase K *	1 mM	unknown	same as stock
RNase A *	1 mL	unknown	1 in 10 diluted
cell culture medium RPMI *	1 mL	n/a	same as stock
Sodium Citrate, Dihydrate (SC)	295 mg	1 M	same as stock
guanidine thiocyanate (GITC)	118 mg	1 M	same as stock
ß-Mercaptoethanol *	1 mL	99%	9.9%
Sucrose	324 mg	1 M	same as stock
Trizma® base	121 mg	1 M	same as stock
Urea	130 mg	2 M	same as stock
Human Serum Albumin (HSA)	20 mg	20 mg/mL	5 mg/mL

Contaminant	Added to 1 mL final volume	Stock concentration	Measurement concentration
IgG from bovine serum (IgG)*	1 mL	12 mg/mL	5 mg/mL
human Hemoglobin (Hb)	13 mg	13 mg/mL	1 mg/mL
EDTA[2]	10 mM	500 mM	10 mM
Sodium Azide (SA)	100 mg	10%	0.1%
Phenol	70 mg	750 mM	7.5 mM

[1] For Cells / Debris, 1x107 thawed Jurkat cells in a pellet were solved in RNase-free water and vortexed for 30 s.

[2] EDTA was solved in 3 M NaOH since it was insoluble in RNase-free water.

4.2.2 Preparation of contaminant pre-dilutions for DNA samples

To investigate the influence of possible contaminants on DNA absorbance spectrum and enzyme activities, contaminant dilution series in buffer EB were prepared. These pre-dilutions, summarized in Table 15, had 10fold concentration of target concentration in DNA sample.

Table 15: Concentrations of pre-dilutions and dilution factor for contaminant dilution series.

Contaminant	Dilution factor	Concen-tration 1	Concen-tration 2	Concen-tration 3	Concen-tration 4
Sodiumcitrate [mM]	1.6	400	250	156	97.7
Betaine [mg/µL]	2	1	0.5	0.25	0.13
EDTA [mM]	1.7	20	11.8	6.9	4.1
Sodiumazide [% (w/v)]	2	1	0.5	0.25	0.13
Hemoglobin [mg/mL]	3.3	10	3	0.9	0.27
Phenol [mM]	3	75	25	8.33	2.78
GITC [mM]	4	200	50	12.5	3.13
Glycogen [mg/mL]	1.25	15	12	9.6	7.68
DTT [mM]	2	100	50	25	12.5
dNTPs [mM]	4	1.2	0.31	0.08	0.02
HSA [mg/mL]	2	11	5.5	2.75	1.38
IgG [mg/mL]	2	3.6	1.8	0.9	0.45

4.2.3 DNA sample preparation with contaminants

Pure DNA was spiked with possible contaminants to record absorbance spectra and investigate the influence of contaminants on enzyme activity assays. A sample series was prepared for each contaminant, consisting of four contaminated DNA samples with

decreasing contaminant concentrations and one clean DNA control, containing buffer EB instead of contaminant. All samples were vortexed and shortly spun down before pipetting.

PhiX DNA for Phi-Inhibition-Assay or dsDNA restriction fragments for gel electrophorese based ligase assay were adjusted to 33.3 ng/µL in buffer EB. Samples with 108 µL pre-diluted DNA were combined with 12 µL pre-diluted contaminant (chapter 4.2.2).

4.2.4 Recording absorbance spectra for data modelling

UV/Vis absorbance spectra of pre-selected possible contaminants were measured on DropSense96 (Trinean) with General UV/Vis application or QIAxpert (QIAGEN) with UV/Vis application, using RNase-free water as blank. For each sample, 4 replicates of 2 µL were applied according to manufacturer's instructions.

For enzyme activity assays and data modelling, absorbance spectra of contaminated and clean DNA controls were recorded on QIAxpert (QIAGEN), using DNA QIAsymphony and UV/Vis application, following manufacturer's instructions. Duplicates of each sample were measured with each application. DNA QIAsymphony and UV/Vis application measurements were used to extract spectra for data modelling.

Raw spectra were extracted with Troubleshoot export application (version 3.2.0.1) of cDrop software, while nucleic acid spectra from SCP of DNA QIAsymphony application were obtained with QIAxpert Binary Reader. Both were QIAGEN internal software tools.

4.2.5 Determination of DNA concentration

DNA concentration at several working steps was determined on QIAxpert, applying 2 µL sample, varying applications and blanks, summarized in Table 16. All samples were vortexed and shortly spun down before pipetting.

Table 16: QIAxpert application and blank used for determination of DNA concentration for different working steps. DNA marked with * was single stranded; conversion of dsDNA concentration to ssDNA concentration was done by application of equation (1).

DNA	Use / working step	QIAxpert application	Blank
PhiX174 DNA	Phi-Inhibition-Assay	DNA QIAsymphony	None (autoblank)
pMCVbeta plasmid	template preparation for ligase activity assay	A260 dsDNA	buffer TE
PCR product	template preparation for ligase activity assay	PCR QIAquick	None (autoblank)
PhiX174 DNA*	Phi-Inhibition-Assay	DNA QIAsymphony	None (autoblank)
Restriction fragments	ligase activity assay	DNA QIAsymphony	None (autoblank)
Jurkat DNA	template preparation for kinase assay	DNA QIAamp	None (autoblank)
PCR product	radiometric kinase assay	DNA QIAsymphony	None (autoblank)
calf thymus DNA	DNA dilution series	DNA QIAsymphony	None (autoblank)
saliva DNA	qPCR for algorithm testing	DNA QIAsymphony	None (autoblank)

For quantification of single stranded PhiX DNA, concentration results of DNA QIAsymphony application for double stranded DNA were converted to ssDNA concentration using equation (1).

$$concentration\ ssDNA = \left(\frac{concentration\ dsDNA}{50}\right) \times 33 \qquad (1)$$

with *concentration dsDNA* being the concentration obtained from QIAxpert measurement.

4.2.6 Dilution buffer for Taq polymerase and T4 DNA ligase

The polymerase dilution buffer used to dilute Taq polymerase was prepared in a large batch (Table 17), aliquoted and stored a -20°C. Before use it was taken from the freezer and equilibrated to room temperature.

Table 17: Composition of polymerase dilution buffer for Phi-Inhibition-Assay.

component	stock concentration	final concentration	volume
PCR buffer	10 x	1 x	5 mL
BSA	10 mg/mL	10 µg/mL	25 µL
Tween 20	100 %	0.5 %	250 µL
Nonidet P 40	100 %	0.5 %	250 µL
RNase-free water	-	-	44.475 mL
Total volume			**50 mL**

To dilute T4 DNA Ligase, 100 µg/mL BSA were freshly added to 1x ligation buffer, just before enzyme dilutions were prepared.

4.2.7 Master mix preparation for Phi-Inhibition-Assay

For master mix preparation, EvaGreen was pre-diluted from 20,000x to 20x, and primer from 100 mM to 10 mM in RNase-free water. Subsequently, master mix was prepared containing reaction buffer, primer, dNTPs, and fluorescence dye (Table 18).

Table 18: Reaction mix composition of Phi-Inhibition-Assay. Components marked with a * were not added to master mix.

component	stock concentration	final concentration	volume per reaction
PCR buffer	10 x	1 x	2 µL
Primer PhiX174	10 µM	0.1 µM	0.2 µL
dNTPs	10 mM	150 µM	0.3 µL
EvaGreen	20 x	1 x	1 µL
PhiX DNA*	30 ng/µL	15 ng/µL	varying
Taq dilution*	-	-	1 µL
RNase-free water*	-	-	varying
Total volume			**20 µL**

PhiX DNA, different polymerase dilutions, and RNase-free water were added directly into reaction mix.

4.2.8 Master mix preparation for ligase activity assay

Master mix for gel electrophoresis based ligase assay was ligation buffer only. RNase-free water, T4 DNA Ligase and dsDNA fragments were added directly to final reaction mix, described in Table 19.

Table 19: Reaction mix composition of gel electrophoresis based ligase assay. Components marked with a * were not added to master mix, but directly to reaction mix.

Component	Stock concentration	Final concentration	Volume / reaction
Ligation Buffer	10 x	1 x	2 µL
dsDNA fragments*	30 ng/µL	15 ng/µL	10 µL
T4 DNA Ligase*			1 µL
RNase-free water*			7 µL
Total volume			**20 µL**

4.2.9 Volume of DNA and water applied in enzyme activity assay

The volume of DNA and RNase-free water added to reaction mix, varied depending on measured DNA concentration (chapter 4.2.5). Therefore, the DNA volume was calculated using the equation (2).

$$c_1 \times v_1 = c_2 \times v_2 \quad (2)$$

where c_1 was DNA concentration of contaminated sample measured on QIAxpert, c_2 was the final DNA concentration in reaction mix, and v_2 the final reaction volume.

The amount of RNase-free water added was determined with the equation (3)

$$v_{RNase-free\ water} = v_{total} - v_{MasterMix} - v_{DNA} \quad (3)$$

where v_{total} was the total reaction volume.

4.2.10 Enzyme activity assay reaction mix setup

All standard, samples and controls on enzyme activity assays were applied in four replicates, if not noted otherwise in the result section. The reaction mix for all four replicates was prepared in one tube by adding master mix, RNase-free water, DNA and enzyme in this order, mixed, spun down and divided into four single reaction tubes.

For Phi-Inhibition-Assay, four standards with same amount of pure DNA and varying concentrations of enzyme, as well as five samples for two contaminants with varying amount of contaminated DNA and highest enzyme concentration were applied on each run. In addition, a negative control with DNA but without enzyme was added.

For gel electrophorese based ligase assay, five standards and five samples for three contaminant series, as well as a ligase negative control were applied on each run.

4.2.11 Temperature profiles and detection of enzyme activity assays

Phi-Inhibition-Assay was carried out in 4-Strip tubes and 72-Well Rotor Disk on RotorGene Q, using Q-Rex Software. The cycler was set to run a 2 min hold at 95°C for initial denaturation, followed by 30 cycles of 5 s primer annealing at 55°C, and 32 s amplification at 72°C. Fluorescence signal was acquired on green channel after amplification.

The reactions of gel electrophoresis based ligase assay were placed in a BioRad thermal cycler set to run 60 min ligation at 16°C and 5 min ligase inactivation at 70°C, before reactions were cooled to 4°C until further processing. For detection of ligated and un-ligated dsDNA fragments, 80 µL RNase-free water were added to each ligation reaction and tubes were transferred to QIAxcel Advanced. Gel image and electropherograms were recorded using 15 bp – 5 kb alignment marker, 5 ng/µL FX174 size marker, and method OM 500. Analysis parameter in QIAxcel ScreenGel Software were set to standard settings, except:

Minimum Distance:	3.00 s	
Threshold:	Start: 0.00 min	Value: 10.00 S/N
	Start: 4.50 min	Value: 2.50 S/N
Alignment Marker Threshold:	10 S/N	

To determine the amount of ligated vs. non-ligated fragments after ligation reaction, the "NAPercentage" (percent normalized area, % NA) value was chosen. This value gives the percentage of the area under the curve of each peak in an electropherogram, where the sum of the areas under the curve of all detected peaks between, but without alignment marker peaks equals 100%.

4.2.12 Data collection for enzyme activity assays

For Phi-Inhibition-Assay, fluorescence signal of intercalating dye was recorded and slope of measured relative fluorescence intensity (RFI) was calculated using equation (4).

$$slope_{RFI} = \frac{\Delta_{RFI}}{\Delta_{cycle}} \quad (4)$$

The $slope_{RFI}$ was calculated for each measurement replicate using RFI values of cycle 10 – 25.

For gel electrophoresis based ligase assay, normalized area percentage (NA %) values were read out from ScreenGel Software for each measurement replicate.

To detect and exclude maximum one outlier of four replicate measurements, Nalimov outlier test (5) was applied. Remaining values, after deletion of outliers, were called valid $slope_{RFI}$ values. The outlier test was applied twice in a row and if two outliers were detected, all replicates of the sample were excluded from further evaluation and corresponding value was treated as missing value.

$$q = \left|\frac{x_1 - \bar{x}}{s_n}\right| \times \sqrt{\frac{n}{n-1}} \tag{5}$$

where s_n is standard deviation, n the number of replicates, x_1 is the tested value, and $\bar{x}$ the arithmetical mean.

The valid $slope_{RFI}$ values of standards for Phi-Inhibition-Assay were plotted as y against applied enzyme concentration of standards as x to obtain standard curve described by

$$y = mx + b \tag{6}$$

where

$$m = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \tag{7}$$

and

$$b = \bar{y} - m\bar{x} \tag{8}$$

$\bar{x}$ and $\bar{y}$ are arithmetical means of four replicates for corresponding values. Furthermore, the coefficient of determination was determined as

$$R^2 = \frac{\sum(x - \bar{x})\,(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{9}$$

The standard curve or each run was then used to calculate the enzyme activity of contaminated DNA samples, using mean values of valid $slope_{RFI}$.

For gel electrophoresis based ligase assay, valid % NA values of standards were used instead of valid $slope_{RFI}$ values to create standard curve as described in equation (6), (7), and (8). Subsequently, valid % NA of contaminated DNA samples were applied to determine enzyme activity in presence of defined contaminant concentration.

Each run with same enzyme and samples was carried our six times to obtain six biological replicates for each enzyme and contaminant concentration combination. If R^2 of standard curve was < 0.975 (9), the complete run was excluded from further evaluation and all values

from this run were treated as missing values, or run was repeated to obtain at least four valid values of six biological replicates.

To calculate enzyme activity in percent, a second standard curve, as described in equation (6), (7), and (8), was created for each contaminant series and enzyme, where measured enzyme activities in U/µL were x and enzyme activity in percent were y. Thereby, enzyme activities of clean DNA controls were set to 100% and measured activities of no enzyme controls were set to 0%. Percent enzyme activity values were used for data modelling.

4.2.13 Statistical comparison of measured enzyme activity means

One-way Analysis of Variance (ANOVA) and Tukey-Kramer Honest Significant Difference (Tukey-Kramer HSD) test were applied to compare means of enzyme activities. A p-value < 0.05 was considered a statistically significant difference between the samples with different contaminant concentrations.

4.2.14 Primer design for plasmid PCR for ligase activity assay

The NCBI Primer BLAST was used to find Primers flanking the XhoI restriction site, using default setting with following alterations:

- PCR Template: pMCVbeta FASTA sequence [109]
- Range Forward Primer From “300” To “400”
- PCR Product size Min “400” Max “600”
- Specificity check Database: “Custom” with pMCVbeta FASTA sequence

Of 10 suggested results, a primer pair for a PCR product of 549 bp, flanking the XhoI restriction site, and ranging from position 370 to 918 on the plasmid template, was selected and the NCBI Nucleotide BLAST was used to assure specificity, using default settings entering the following search criteria:

- Enter query sequence: Sequence of selected forward and corresponding reverse primer
- Organism: “Cloning vector pCMVbeta (taxid: 31798)”

The selected primer pair resulted to be specific with binding sites of maximum 8 bp on off-target sequences, without leading to any product. Furthermore, no primer dimer formation was found, using the Multiple Primer Analyzer form Thermo Fisher Sientific.

4.2.15 Plasmid PCR for ligase activity assay

A standard PCR setup was used to generate a PCR Product, flanking XhoI restriction site. The PCR setup described in Table 20 was used to prepare three replicates.

Primer were obtained in lyophilized form, diluted according to manufacturer's instructions to obtain 100 µM concentration in Buffer TE, and further diluted in Buffer TE to obtain 20 µM stock concentrations. The pMCVbeta plasmid was diluted from 600 ng/µL to 5 ng/µL in Buffer EB.

Table 20: Setup for plasmid PCR.

Component	Stock concentration	Final concentration	Volume / reaction
HotStarTaq MM	2 x	1 x	25 µL
Fwd primer	20 µM	0.25 µM	1.25 µL
Rev primer	20 µM	0.25 µM	1.25 µL
pMCVbeta DNA	5 ng/µL	10 ng	2 µL
RNase-free water			20.5 µL
Total volume			**50 µL**

PCR reactions were placed in BioRad Cycler, set to perform a 15 min polymerase activation at 95°C, followed by 35 cycles of 30 s denaturation at 94°C, 30 s primer annealing at 56°C, and 60 s extension at 72°C. The run ended with a 10 min final extension at 72°C and a cool down to 4°C.

4.2.16 Restriction digest of dsDNA template for ligase activity assay

Restriction enzyme digest was setup according to NEB recommendations and as described in Table 21 for XhoI digest of PCR product.

Table 21: Reaction setup for XhoI restriction digest using Time-Saver™ Protocol.

Component	Stock concentration	Final concentration	Volume / reaction
CutSmart buffer	10 x	1 x	5 µL
XhoI	20000 Units/mL	20 Units	1 µL
PCR product	420 ng/µL	1 µg	2.4 µL
RNase-free water			41.6 µL
Total volume			**50 µL**

Reactions were prepared and transferred to BioRad cycler, set to run 15 min at 37°C and cool down to 12°C.

4.2.17 DNA purification and PCR for radiometric kinase assay

DNA was extracted from Jurkat cells, cultured in RPMI medium with 10% FCS, 1% Pen/Strep and 1% L-Glutamine, with the DNA QIAamp Kit on QIAcube according to manufacturer's protocol and eluted in buffer EB. DNA stock concentrations and purity were measured with RNA RNeasy application on QIAxpert.

In order to generate a PCR product without 5'-phosphorylation as dsDNA substrate for radiometric kinase assay, a standard PCR setup was applied as described in Table 20 using Jurkat DNA as template, primer for T4 PNK assay, and annealing temperature set to 58°C instead of 56°C.

4.2.18 PCR product and restriction fragments purification

PCR products and digested dsDNA fragments were purified using MinElute PCR Purification Kit as described in the MinElute Handbook 03/2008 page 19-20, to remove PCR template DNA, enzymes and buffer components. Purified PCR product and dsDNA fragments were eluted in 10 µL Buffer EB.

4.2.19 Gel electrophoresis of PCR products and restriction fragments

After PCR or restriction digest dsDNA fragments were diluted 1 in 10 with QIAxcel Dilution Buffer and applied on QIAxcel Advanced, using a High Resolution cartridge, 15 bp – 3 kb alignment marker, and FX174 or *Hae*III size marker in combination with method OM 500.

4.2.20 Radiometric kinase assay for T4 PNK activity measurement

Before reaction mix setup for kinase assay, T4 PNK was diluted in 1x Polynucleotide Kinase buffer with freshly added 100 µg/mL BSA, and 0.5 mM ATP and 240 µCi/mL radioactive labeled [γ-^{32}P]ATP were added to 10x PNK buffer. Therefore, volume of [γ-^{32}P]ATP was adjusted daily based on radioactive activity obtained from PerkinElmer's Radioactive Decay Calculator.

The 10x reaction buffer with ATP, RNase-free water, T4 PNK dilutions and dsDNA fragments were then combined as described in Table 22 to obtain 25 µL reaction mix in single reaction tubes. Per run, four standards and one no enzyme control were applied in single replicates.

Table 22: Reaction mix composition of gel electrophoresis based ligase assay. Components marked with a * were not added to master mix, but directly to reaction mix.

Component	Stock concentration	Final concentration	Volume / reaction
PNK buffer with ATP	10 x	1 x	2.5 µL
dsDNA fragments	30 ng/µL	15 ng/µL	12.5 µL
T4 DNA Ligase			2.5 µL
RNase-free water			7.5 µL
Total volume			**25 µL**

Prepared reaction mix was transferred to BioRad thermal cycler and incubated for 30 min at 37°C before cool down to 4°C. Subsequently dsDNA fragments from reaction mix were purified using DyeEx 2.0 Spin columns and 20 µL reaction volume, according to manufacturer's protocol. After centrifugation, spin columns containing free [γ-^{32}P]ATP were discarded, whereas collection tubes with flow through containing unlabeled and radioactive labeled dsDNA fragments were transferred to scintillation tubes. Counts per minute (CPM) were recorded on Beckman LS 6500 scintillation counter and measured CPM were used as x to to create standard curve using equations (6), (7), and (8).

4.2.21 Preparation of DNA dilution series for algorithm testing

Calf thymus DNA at a stock concentration of 1000 ng/µL was diluted to 750, 500, 250, 100, 50, 25, 10, 5, and 2.5 ng/µL in Buffer EB. Absorbance spectra were recorded as described in 4.2.4 and DNA concentration was determined as described in chapter 4.2.5, Table 16.

4.2.22 qPCR for algorithm testing

DNA extracted from human saliva samples were obtained as donation from the PAX group at QIAGEN, and applied on a probe based qPCR targeting the human *β-actin* gene. Therefore, absorbance spectra of 38 samples were recorded (chapter 4.2.4), DNA concentrations were determined (chapter 4.2.5, Table 16) and all samples were diluted in Buffer EB to obtain 5 ng/µL DNA. Subsequently, DNA samples were applied in triplicates on qPCR with reaction mix described in Table 23 prepared at room temperature in RotorGene 4-strip tubes. In addition to saliva DNA samples, a standard curve with 4 standards containing 100, 10, 1, or 0.1 ng clean Jurkat DNA as final concentration, as well as a no template control (NTC) were applied in triplicates.

Table 23: Reaction mix for probe based qPCR.

Component	Stock concentration	Final concentration	Volume / reaction
QN Multiplex MM	4 x	1 x	5 µL
Fwd primer	10 µM	0.4 µM	0.8 µL
Rev primer	10 µM	0.4 µM	0.8 µL
TaqMan probe	10 µM	0.25 µM	0.5 µL
saliva DNA	5 ng/µL	10 ng	2 µL
RNase-free water			11 µL
Total volume			**20 µL**

PCR reactions were placed in 72-well RotorDisk and transferred to RotorGene Q. The Q-Rex software was set to run an initial activation step at 95°C for 2 min, followed by 35 cycles of denaturation for 5 s at 95°C and combined annealing and extension for 30 s at 60°C. Data acquisition on green channel was performed after annealing and extension step.

To obtain Cq values of saliva DNA samples, standard curve was evaluated using the Absolute Quantification plug-in and auto threshold function in Q-Rex software. Delta Cq values of saliva DNA samples were calculated by subtracting mean Cq value of standard with 10 ng DNA final concentration from mean Cq value of each sample.

4.2.23 Feature selection using near zero variance

The near zero variance (nzv) algorithm removes all features or variables, here wavelengths of UV/vis absorbance spectra, that had a variance lower than a selected threshold within the trainings dataset. Feature selection with nzv was performed using the sklearn.feature_selection.VarianceThreshold for python [110], [111], with threshold set to 0.01, 0.02, 0.03, 0,06, 0.1, 0.2, 0.3, 0.6, 1, 2, 3, 6, 10 or 20 OD. All other parameters were set to default settings.

The variance is the average of squared differences from the mean and could be determined for each wavelength using equation (10), where x is a specific wavelength of UV/Vis absorbance spectrum and n is the total number of observations for each wavelength.

$$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (10)$$

Subsequently, all wavelengths with $Var(x) \leq threshold$ would be eliminated from input features.

4.2.24 Feature selection with principal component analysis

The goal of principal component analysis (PCA) is to reduce the number of input variables consisting of many correlated variables such as the wavelengths of UV/vis absorbance spectra, while retaining as much of the variation present in the data set as possible. This is accomplished by an orthogonal linear transformation of the original input variables to a new set of uncorrelated and ordered variables in a new coordinate system, called the principal components. The order is such, that the first few principal components retain most of the variation present in all the original input variables [112].

For this thesis, the principal components were computed using the sklearn.decomposition.PCA for python [110], [113]. All parameters were set to default settings, except *n_components*, which was set to 0.9, 0.95, 0.975 or 0.99 to retain 90, 95, 97.5 or 99% of variance from original input data.

The mathematical steps behind the applied code can be described as follows: To obtain the principal components from UV/vis absorbance spectra, they were arranged into a ***d*** dimensional design matrix **A**, where each column ***x*** was a wavelength between 230 and 410 nm and each row ***n*** was a measurement instance corresponding to an enzyme activity or label.

$$A_{nx} = \begin{bmatrix} a_{11} & \dots & a_{1x} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nx} \end{bmatrix} \tag{11}$$

First, the mean of each column ***x*** was determined, giving the mean of matrix **A**:

$$\bar{A} = \begin{bmatrix} \frac{a_{11} + a_{21} + \cdots + a_{n1}}{n} & \frac{a_{12} + a_{22} + \cdots + a_{n2}}{n} & \dots & \frac{a_{1x} + a_{2x} + \cdots + a_{nx}}{n} \end{bmatrix} \tag{12}$$

The mean of matrix **A** was then used to calculate the variance-covariance matrix, which contains the variance for each ***n*** in the diagonal and the covariance between different ***n***s in the off-diagonal elements. The variance-covariance matrix of matrix **A** was computed using the following formula:

$$B = cov(A, A') = \frac{1}{n-1} \sum_{i=1}^{n} (A_i - \bar{A})\,(A_i - \bar{A})' \tag{13}$$

were $\boldsymbol{A}'$ was the *transpose* of the matrix ***A***. The result was a square matrix **B**, of which in a next step, the eigenvectors and corresponding eigenvalues were determined. The

eigenvector $\boldsymbol{v}$ is a non-zero vector of the square matrix **B**, such that for some scalar **λ** the equation

$$Bv = \lambda v \tag{14}$$

is satisfied. The scalar **λ** is called an eigenvalue of matrix **B**. Each eigenvalue has its set of eigenvectors. Therefore, the eigenvalues were determined in order to find their eigenvectors, by stating the equation (14) as

$$(B - \lambda I)v = 0 \tag{15}$$

where $\boldsymbol{I}$ was the $\boldsymbol{n}$ by $\boldsymbol{n}$ identity matrix an 0 was the zero vector. The equation (15) had a non-zero solution $\boldsymbol{v}$ and the eigenvalues **λ** of **B** could be computed, if:

$$\det(B - \lambda I) = 0 \tag{16}$$

The result of computing the determinant was an equation that was used to obtain the eigenvalues by solving it for **λ.** The eigenvalues were then used to determine the eigenvectors, which defined the directions of the axis in the new coordinate system with the unit length 1. The eigenvalues contained the information about the variation of the data, with higher eigenvalues containing higher variance of the original input data. Thus, the eigenvectors were sorted by decreasing eigenvalues, and the $\boldsymbol{k}$ number of eigenvectors with corresponding eigenvalues containing 90, 95, 97.5 or 99% of variance from original input data represented the wanted principal components and were kept. All remaining eigenvectors and corresponding eigenvalues were dropped. The saved principal components were added to a new $\boldsymbol{d \times k}$ dimensional eigenvector matrix **C**, which was used to transform the measurements onto a new coordinate system via the equation

$$y = C^{'} \times x \tag{17}$$

were $\boldsymbol{C}'$ was the *transpose* of the matrix $\boldsymbol{C}$, and served as new input data for classification algorithms.

4.2.25 Multiclass logistic regression for DNA purity estimation

Multiclass logistic regression (MLR) is a classification method derived from logistic regression, which is commonly used to predict the probabilities of a binary output. In MLR

the output has more than two possible categorical outcomes that are predicted using a set of independent variables [114]. In this study, 5 classes based on measured enzyme activities in percentage: c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5, also called "actual class", were used as categorical outcome, and independent variables used to predict actual classes were different versions of measured UV/Vis absorbance spectra.

The sklearn.linear_model.LogisticRegression for python with *solver* "lbfgs" was applied to compute the MLR algorithm [110], [115]. All parameters were set to default settings, except *class_weight* set to "None" or "balanced" and the inverse regularization strength C set to 0.01, 0.05, 0.1, 0.5, 1.0, 5, 10, 50, or 100 for optimization of MLR performance.

Mathematically, multiclass logistic regression uses a softmax function (18) to determine the probability for an observation $\boldsymbol{x}$ to belong to each of $\boldsymbol{k}$ actual classes. It takes the input vector $\boldsymbol{z}$, containing all outputs of a linear regression, and converts all values to be positive values between 0 and 1, and add up to 1. Thus, they can be interpreted as probabilities.

$$softmax(z) = \frac{e^{z}}{\sum_{i=1}^{k} e^{z_i}} \tag{18}$$

Observation $\boldsymbol{x}$ is then assigned to the class with the highest probability [116], [117].

4.2.26 *K*-nearest-neighbor for DNA purity estimation

The *K*-nearest-neighbor method is a supervised machine learning algorithm for classification. To learn a function for prediction of unknown data, it needs labeled input data [110], [116]. Here 5 classes based on measured enzyme activities in percentage: c1 ≤ 20% < c2 ≤ 40% < c3 ≤ 60 < c4 ≤ 80% < c5 were used as labels or target values, also called "actual class". The data used to predict these classes were different versions of measured UV/Vis absorbance spectra.

The KNN algorithm was computed using the sklearn.neighbors.KNeighborsClassifier for python [110], [118]. All parameters were set to default settings, except *n_neigbors* and *weights*.

To apply the KNN algorithm, it was assumed that all observations or measured UV/Vis absorbance spectra, assigned to an "actual class", were lying within a multi-dimensional Euclidean space. Therefore, the distance between each observation and all other observations could be determined using the Euclidean distance function:

$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2 + \cdots + (x_n - x'_n)^2} \quad (19)$$

The KNN algorithm calculates the distance $\boldsymbol{d}$ between an unknown observation $\boldsymbol{x}$ and all trainings observations $\boldsymbol{N}$. To classify the unknown observation $\boldsymbol{x}$, a sphere is drawn centered on $\boldsymbol{x}$ containing $\boldsymbol{k}$ neighbors from the trainings set independent of their class. $\boldsymbol{A}$ is the subset of $\boldsymbol{k}$ nearest neighbors lying within this sphere and the algorithm determines the probability for $\boldsymbol{x}$ falling within any of the classes, in which are k nearest neighbors in, using equation (20)

$$P(y = j|N = x) = \frac{1}{k}\sum_{i \epsilon A} I(y^{(i)} = j) \quad (20)$$

where $\boldsymbol{I(x)}$ is the indicator function, evaluating membership of an observation in a subset $\boldsymbol{A}$ of $\boldsymbol{N}$ and resulting in 1 for all observations of $\boldsymbol{A}$ and 0 for all observations of $\boldsymbol{N}$ not in $\boldsymbol{A}$. Finally, the unknown observation $\boldsymbol{x}$ is assigned to the class with the highest probability [116], [119].

To optimize the performance of KNN algorithm, the number of k nearest neighbors *n_neigbors* was set to 1 through 15, and *weights* were set to *'uniform'* or *'distance'*, with *'uniform'* being the Euclidean distance described in equation (19) and *'distance'* being the inverse of the Euclidean distance in equation (19).

Bibliography

[1] D. Klein, “Quantification using real-time PCR technology: applications and limitations,” *Trends Mol Med*, vol. 8, no. 6, pp. 257–260, Jun. 2002.

[2] S. Imbeaud *et al.*, “Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces,” *Nucleic Acids Res.*, vol. 33, no. 6, p. e56, Mar. 2005.

[3] S. A. Bustin *et al.*, “The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments,” *Clin. Chem.*, vol. 55, no. 4, pp. 611–622, Apr. 2009.

[4] C. Unger, O. Kofanova, K. Sokolowska, D. Lehmann, and F. Betsou, “Ultraviolet C radiation influences the robustness of RNA integrity measurement,” *Electrophoresis*, vol. 36, no. 17, pp. 2072–2081, Sep. 2015.

[5] G. Matthijs *et al.*, “Guidelines for diagnostic next-generation sequencing,” *Eur. J. Hum. Genet.*, vol. 24, no. 1, pp. 2–5, Jan. 2016.

[6] I. G. Wilson, “Inhibition and facilitation of nucleic acid amplification,” *Appl. Environ. Microbiol.*, vol. 63, no. 10, pp. 3741–3751, Oct. 1997.

[7] L. Garibyan and N. Avashia, “Research Techniques Made Simple: Polymerase Chain Reaction (PCR),” *J Invest Dermatol*, vol. 133, no. 3, p. e6, Mar. 2013.

[8] C. Schrader, A. Schielke, L. Ellerbroek, and R. Johne, “PCR inhibitors - occurrence, properties and removal,” *J. Appl. Microbiol.*, vol. 113, no. 5, pp. 1014–1026, Nov. 2012.

[9] T. Nolan, R. E. Hands, W. Ogunkolade, and S. A. Bustin, “SPUD: a quantitative PCR assay for the detection of inhibitors in nucleic acid preparations,” *Anal. Biochem.*, vol. 351, no. 2, pp. 308–310, Apr. 2006.

[10] L. A. Pikor, K. S. S. Enfield, H. Cameron, and W. L. Lam, “DNA extraction from paraffin embedded material for genetic and epigenetic analyses,” *J Vis Exp*, no. 49, Mar. 2011.

[11] C. Endrullat, J. Glökler, P. Franke, and M. Frohme, “Standardization and quality management in next-generation sequencing,” *Applied & Translational Genomics*, vol. 10, pp. 2–9, Sep. 2016.

[12] C. Unger, N. Lokmer, D. Lehmann, and I. M. Axmann, “Detection of phenol contamination in RNA samples and its impact on qRT-PCR results,” *Analytical Biochemistry*, vol. 571, pp. 49–52, Apr. 2019.

[13] L. Rossen, P. Nørskov, K. Holmstrøm, and O. F. Rasmussen, "Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions," *Int. J. Food Microbiol.*, vol. 17, no. 1, pp. 37–45, Sep. 1992.

[14] M. Sidstedt *et al.*, "Inhibition mechanisms of hemoglobin, immunoglobulin G, and whole blood in digital and real-time PCR," *Anal Bioanal Chem*, vol. 410, no. 10, pp. 2569–2583, 2018.

[15] H. L. Katcher and I. Schwartz, "A distinctive property of Tth DNA polymerase: enzymatic amplification in the presence of phenol," *BioTechniques*, vol. 16, no. 1, pp. 84–92, Jan. 1994.

[16] D. Loffert, S. Stump, N. Schaffrath, M. Berkenkopf, and J. Kang, "PCR: Effects of template quality," *Qiagen News*, vol. 1, pp. 8–10, 1997.

[17] M. Simbolo *et al.*, "DNA Qualification Workflow for Next Generation Sequencing of Histopathological Samples," *PLoS One*, vol. 8, no. 6, Jun. 2013.

[18] J. F. Huggett, J. O'Grady, and S. Bustin, "qPCR, dPCR, NGS – A journey," *Biomol Detect Quantif*, vol. 3, pp. A1–A5, Jan. 2015.

[19] M. Kubista *et al.*, "The real-time polymerase chain reaction," *Mol. Aspects Med.*, vol. 27, no. 2–3, pp. 95–125, Jun. 2006.

[20] A. Kornberg, I. R. Lehman, M. J. Bessman, and E. S. Simms, "Enzymic synthesis of deoxyribonucleic acid," *Biochimica et Biophysica Acta*, vol. 21, no. 1, pp. 197–198, Jul. 1956.

[21] I. R. Lehman, S. B. Zimmerman, J. Adler, M. J. Bessman, E. S. Simms, and A. Kornberg, "Enzymatic Synthesis of Deoxyribonucleic Acid. V. Chemical Composition of Enzymatically Synthesized Deoxyribonucleic Acid," *PNAS*, vol. 44, no. 12, pp. 1191–1196, Dec. 1958.

[22] A. Chien, D. B. Edgar, and J. M. Trela, "Deoxyribonucleic acid polymerase from the extreme thermophile Thermus aquaticus.," *J Bacteriol*, vol. 127, no. 3, pp. 1550–1557, Sep. 1976.

[23] K. B. Mullis, "The unusual origin of the polymerase chain reaction," *Sci. Am.*, vol. 262, no. 4, pp. 56–61, 64–65, Apr. 1990.

[24] M. A. A. Valones, R. L. Guimarães, L. A. C. Brandão, P. R. E. de Souza, A. de Albuquerque Tavares Carvalho, and S. Crovela, "Principles and applications of polymerase chain

reaction in medical diagnostic fields: a review," *Braz J Microbiol*, vol. 40, no. 1, pp. 1–11, 2009.

[25] A. E. Tomkinson, S. Vijayakumar, J. M. Pascal, and T. Ellenberger, "DNA Ligases: Structure, Reaction Mechanism, and Function," *Chemical Reviews*, vol. 106, no. 2, pp. 687–699, Feb. 2006.

[26] S. Shuman and M. S. Glickman, "Bacterial DNA repair by non-homologous end joining," *Nat. Rev. Microbiol.*, vol. 5, no. 11, pp. 852–861, Nov. 2007.

[27] T. Ellenberger and A. E. Tomkinson, "Eukaryotic DNA Ligases: Structural and Functional Insights," *Annu Rev Biochem*, vol. 77, pp. 313–338, 2008.

[28] S. Shuman, "DNA Ligases: Progress and Prospects," *J Biol Chem*, vol. 284, no. 26, pp. 17365–17369, Jun. 2009.

[29] M. Gellert, "Formation of covalent circles of lambda DNA by E. coli extracts.," *Proc Natl Acad Sci U S A*, vol. 57, no. 1, pp. 148–155, Jan. 1967.

[30] B. Weiss and C. C. Richardson, "Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage.," *Proc Natl Acad Sci U S A*, vol. 57, no. 4, pp. 1021–1028, Apr. 1967.

[31] B. M. Olivera, Z. W. Hall, and I. R. Lehman, "Enzymatic Joining of Polynucleotides, V. A DNA-Adenylate Intermediate in the Polynucleotide-Joining Reaction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 61, no. 1, pp. 237–244, 1968.

[32] M. L. Gefter, A. Becker, and J. Hurwitz, "The enzymatic repair of DNA. I. Formation of circular lambda-DNA.," *Proc Natl Acad Sci U S A*, vol. 58, no. 1, pp. 240–247, Jul. 1967.

[33] N. R. Cozzarelli, N. E. Melechen, T. M. Jovin, and A. Kornberg, "Polynucleotide cellulose as a substrate for a polynucleotide ligase induced by phage T4," *Biochemical and Biophysical Research Communications*, vol. 28, no. 4, pp. 578–586, Aug. 1967.

[34] I. R. Lehman, "DNA ligase: structure, mechanism, and function," *Science*, vol. 186, no. 4166, pp. 790–797, Nov. 1974.

[35] J. E. Mertz and R. W. Davis, "Cleavage of DNA by R1 Restriction Endonuclease Generates Cohesive Ends," *Proc Natl Acad Sci U S A*, vol. 69, no. 11, pp. 3370–3374, Nov. 1972.

[36] U. Landegren, R. Kaiser, J. Sanders, and L. Hood, *A Ligase-Mediated Gene Detection Technique*, vol. 241. 1988.

[37] M. Zirvi, F. Barany, T. Nakayama, G. Newman, T. McCaffrey, and P. Paty, "Ligase-based detection of mononucleotide repeat sequences," *Nucleic Acids Res*, vol. 27, no. 24, pp. e40–e47, Dec. 1999.

[38] J. P. Schouten, C. J. McElgunn, R. Waaijer, D. Zwijnenburg, F. Diepvens, and G. Pals, "Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification," *Nucleic Acids Res*, vol. 30, no. 12, p. e57, Jun. 2002.

[39] D. Voet, J. G. Voet, and C. W. Pratt, *Principles of biochemistry, 4th Edition International Student Versio*, 4th ed. Singapore: John Wiley & Sons, 2013.

[40] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry: International Edition*, 7th edition. Houndmmills: W. H. Freeman and Company, 2012.

[41] A. Novogrodsky and J. Hurwitz, "The Enzymatic Phosphorylation of Ribonucleic Acid and Deoxyribonucleic Acid I. PHOSPHORYLATION AT 5'-HYDROXYL TERMINI," *J. Biol. Chem.*, vol. 241, no. 12, pp. 2923–2932, Jun. 1966.

[42] M. Amitsur, R. Levitz, and G. Kaufmann, "Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA.," *EMBO J*, vol. 6, no. 8, pp. 2499–2503, Aug. 1987.

[43] L. K. Wang, C. D. Lima, and S. Shuman, "Structure and mechanism of T4 polynucleotide kinase: an RNA repair enzyme," *EMBO J*, vol. 21, no. 14, pp. 3873–3880, Jul. 2002.

[44] Y. Cen, W.-J. Deng, R.-Q. Yu, and X. Chu, "Sensitive fluorescence sensing of T4 polynucleotide kinase activity and inhibition based on DNA/polydopamine nanospheres platform," *Talanta*, vol. 180, pp. 271–276, Apr. 2018.

[45] J. Sambrook, T. Maniatis, E. F. Fritsch, and C. S. H. Laboratory, *Molecular cloning : a laboratory manual*, 2nd ed. Cold Spring Harbor, N.Y. : Cold Spring Harbor Laboratory Press, 1987.

[46] J. T. Y. Lee, K. M. C. Cheung, and V. Y. L. Leung, "Correction for concentration overestimation of nucleic acids with phenol," *Anal. Biochem.*, vol. 465, pp. 179–186, Nov. 2014.

[47] C. C. Richardson, C. L. Schildkraut, H. V. Aposhian, and A. Kornberg, "Enzymatic Synthesis of Deoxyribonucleic Acid XIV. FURTHER PURIFICATION AND PROPERTIES

OF DEOXYRIBONUCLEIC ACID POLYMERASE OF ESCHERICHIA COLI," *J. Biol. Chem.*, vol. 239, no. 1, pp. 222–232, Jan. 1964.

[48] M. Seville, A. B. West, M. G. Cull, and C. S. McHenry, "Fluorometric Assay for DNA Polymerases and Reverse Transcriptase," *BioTechniques*, vol. 21, no. 4, pp. 664–672, Oct. 1996.

[49] D. R. Zweitzig, N. M. Riccardello, B. I. Sodowich, and S. M. O'Hara, "Characterization of a novel DNA polymerase activity assay enabling sensitive, quantitative and universal detection of viable microbes," *Nucleic Acids Res*, vol. 40, no. 14, p. e109, Aug. 2012.

[50] H. Tveit and T. Kristensen, "Fluorescence-Based DNA Polymerase Assay," *Analytical Biochemistry*, vol. 289, no. 1, pp. 96–98, Feb. 2001.

[51] C. Ma *et al.*, "Real-time monitoring of DNA polymerase activity using molecular beacon," *Analytical Biochemistry*, vol. 353, no. 1, pp. 141–143, Jun. 2006.

[52] J. Tong, F. Barany, and W. Cao, "Ligation reaction specificities of an NAD+-dependent DNA ligase from the hyperthermophile Aquifex aeolicus," *Nucleic Acids Res*, vol. 28, no. 6, pp. 1447–1454, Mar. 2000.

[53] S. Franke, T. Kreisig, K. Buettner, and T. Zuchner, "One-step assay for the quantification of T4 DNA ligase," *Anal Bioanal Chem*, vol. 407, no. 4, pp. 1267–1271, Feb. 2015.

[54] Z. Tang *et al.*, "Real-time monitoring of nucleic acid ligation in homogenous solutions using molecular beacons," *Nucleic Acids Res*, vol. 31, no. 23, p. e148, Dec. 2003.

[55] Z. L. Wu, "Phosphatase-Coupled Universal Kinase Assay and Kinetics for First-Order-Rate Coupling Reaction," *PLoS One*, vol. 6, no. 8, Aug. 2011.

[56] K. M. Kleman-Leyer *et al.*, "Characterization and Optimization of a Red-Shifted Fluorescence Polarization ADP Detection Assay," *Assay Drug Dev Technol*, vol. 7, no. 1, pp. 56–67, Feb. 2009.

[57] C. Jiang, C. Yan, J. Jiang, and R. Yu, "Colorimetric assay for T4 polynucleotide kinase activity based on the horseradish peroxidase-mimicking DNAzyme combined with λ exonuclease cleavage," *Anal. Chim. Acta*, vol. 766, pp. 88–93, Mar. 2013.

[58] S. Liu *et al.*, "Amplified detection of T4 polynucleotide kinase activity based on a λ-exonuclease cleavage-induced DNAzyme releasing strategy," *Sensors and Actuators B: Chemical*, vol. 192, pp. 157–163, Mar. 2014.

[59] T. Hou, X. Wang, X. Liu, T. Lu, S. Liu, and F. Li, "Amplified Detection of T4 Polynucleotide Kinase Activity by the Coupled λ Exonuclease Cleavage Reaction and Catalytic Assembly of Bimolecular Beacons," *Anal. Chem.*, vol. 86, no. 1, pp. 884–890, Jan. 2014.

[60] A. Psifidi *et al.*, "Comparison of Eleven Methods for Genomic DNA Extraction Suitable for Large-Scale Whole-Genome Genotyping and Long-Term DNA Banking Using Blood Samples," *PLoS One*, vol. 10, no. 1, Jan. 2015.

[61] C. Foley, C. O'Farrelly, and K. G. Meade, "Technical note: Comparative analyses of the quality and yield of genomic DNA from invasive and noninvasive, automated and manual extraction methods," *Journal of Dairy Science*, vol. 94, no. 6, pp. 3159–3165, Jun. 2011.

[62] S. R. Gallagher and P. R. Desjardins, "Quantitation of DNA and RNA with absorption and fluorescence spectroscopy," *Curr Protoc Protein Sci*, vol. Appendix 3, p. Appendix 4K, May 2008.

[63] T. M. Stulnig and A. Amberger, "Exposing contaminating phenol in nucleic acid preparations," *BioTechniques*, vol. 16, no. 3, pp. 402–404, Mar. 1994.

[64] P. Desjardins and D. Conklin, "NanoDrop microvolume quantitation of nucleic acids," *J Vis Exp*, no. 45, Nov. 2010.

[65] K. L. Manchester, "Value of A260/A280 ratios for measurement of purity of nucleic acids," *BioTechniques*, vol. 19, no. 2, pp. 208–210, Aug. 1995.

[66] J. M. Teare, R. Islam, R. Flanagan, S. Gallagher, M. G. Davies, and C. Grabau, "Measurement of nucleic acid concentrations using the DyNA Quant and the GeneQuant," *BioTechniques*, vol. 22, no. 6, pp. 1170–1174, Jun. 1997.

[67] O. Warburg and W. Christian, "Isolierung und Kristallisation des Gärungsferments Enolase," *Naturwissenschaften*, vol. 29, pp. 589–590, Sep. 1941.

[68] J. A. Glasel, "Validity of nucleic acid purities monitored by 260nm/280nm absorbance ratios," *BioTechniques*, vol. 18, no. 1, pp. 62–63, Jan. 1995.

[69] W. W. Wilfinger, K. Mackey, and P. Chomczynski, "Effect of pH and ionic strength on the spectrophotometric assessment of nucleic acid purity," *BioTechniques*, vol. 22, no. 3, pp. 474–476, 478–481, Mar. 1997.

[70] J. R. Warner, "The economics of ribosome biosynthesis in yeast," *Trends Biochem. Sci.*, vol. 24, no. 11, pp. 437–440, Nov. 1999.

[71] H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, uido Marcucci, and K. Kornacker, "Chipping away at the chip bias: RNA degradation in microarray analysis," *Nature Genetics*, vol. 35, no. 4, pp. 292–293, Dec. 2003.

[72] D. Nadano and T.-A. Sato, "Caspase-3-dependent and -independent Degradation of 28 S Ribosomal RNA May Be Involved in the Inhibition of Protein Synthesis during Apoptosis Initiated by Death Receptor Engagement," *J. Biol. Chem.*, vol. 275, no. 18, pp. 13967–13973, May 2000.

[73] O. Mueller, S. Lightfoot, and A. Schroeder, "RNA Integrity Number (RIN) - Standarization of RNA Quality Control. An Application Note of Agilent Technologies." Publication Number 5989-1165EN, 01-May-2004.

[74] A. Schroeder *et al.*, "The RIN: an RNA integrity number for assigning integrity values to RNA measurements," *BMC Mol Biol*, vol. 7, p. 3, Jan. 2006.

[75] S. Fleige and M. W. Pfaffl, "RNA integrity and the effect on the real-time qRT-PCR performance," *Molecular Aspects of Medicine*, vol. 27, no. 2–3, pp. 126–139, Apr. 2006.

[76] "Methods to Check RNA Integrity - DE." [Online]. Available: https://www.thermofisher.com/de/de/home/references/ambion-tech-support/rna-isolation/tech-notes/is-your-rna-intact.html. [Accessed: 15-Mar-2019].

[77] J. Saurina, S. Hernández-Cassou, R. Tauler, and A. Izquierdo-Ridorsa, "Procedure for the Quantitative Determination of Mixtures of Nucleic Acid Components Based on Multivariate Spectrophotometric Acid−Base Titrations," *Analytical Chemistry*, vol. 71, no. 1, pp. 126–134, Jan. 1999.

[78] T. Boonefaes, "Dna and/or rna determination from uv-vis spectrophotometer data," EP2681532B1, 08-Apr-2015.

[79] T. Boonefaes and B. Luyssaert, "Deconvolution of spectra," EP2495546A1, 05-Sep-2012.

[80] "(PDF) Ins and outs of nucleic acid quantification and the DropSense solution towards standardization," *ResearchGate*. [Online]. Available: https://www.researchgate.net/publication/311717557_Ins_and_outs_of_nucleic_acid_quantification_and_the_DropSense_solution_towards_standardization. [Accessed: 12-Feb-2019].

[81] P. Chomczynski and N. Sacchi, "Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction," *Analytical Biochemistry*, vol. 162, no. 1, pp. 156–159, Apr. 1987.

[82] M. R. Green and J. Sambrook, “Precipitation of DNA with Ethanol,” *Cold Spring Harbor Protocols*, vol. 2016, no. 12, p. pdb.prot093377, Dec. 2016.

[83] N. E. Good, G. D. Winget, W. Winter, T. N. Connolly, S. Izawa, and R. M. Singh, “Hydrogen ion buffers for biological research,” *Biochemistry*, vol. 5, no. 2, pp. 467–477, Feb. 1966.

[84] J. M. Thomas and M. E. Hodes, “A new discontinuous buffer system for the electrophoresis of cationic proteins at near-neutral pH,” *Anal. Biochem.*, vol. 118, no. 1, pp. 194–196, Nov. 1981.

[85] B. Saha, D. Saha, S. Niyogi, and M. Bal, “A new method of plasmid DNA preparation by sucrose-mediated detergent lysis from Escherichia coli (gram-negative) and Staphylococcus aureus (gram-positive),” *Anal. Biochem.*, vol. 176, no. 2, pp. 344–349, Feb. 1989.

[86] R. Kansal, K. Kuhar, I. Verma, R. N. Gupta, V. K. Gupta, and K. R. Koundal, “Improved and convenient method of RNA isolation from polyphenols and polysaccharide rich plant tissues,” p. 4.

[87] M. Ishizawa, Y. Kobayashi, T. Miyamura, and S. Matsura, “Simple procedure of DNA isolation from human serum,” *Nucleic Acids Research*, vol. 19, no. 20, pp. 5792–5792, 1991.

[88] G. Gomori, “Preparation of Buffers for Use in Enzyme Studies,” *Methods Enzymology*, vol. 1, pp. 138–146, 1955.

[89] D. R. Nelson, A. L. Lehninger, and M. Cox, “Lehninger principles of biochemistry,” New York: W.H. Freeman, 2005, p. 148.

[90] N. Bonturi, V. S. C. O. Radke, S. M. A. Bueno, S. Freitas, A. R. Azzoni, and E. A. Miranda, “Sodium citrate and potassium phosphate as alternative adsorption buffers in hydrophobic and aromatic thiophilic chromatographic purification of plasmid DNA from neutralized lysate,” *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, vol. 919–920, pp. 67–74, Mar. 2013.

[91] “Why Is Sodium Used in DNA Extraction?,” *Sciencing*. [Online]. Available: https://sciencing.com/sodium-used-dna-extraction-6504902.html. [Accessed: 10-May-2019].

[92] W. Henke, K. Herdel, K. Jung, D. Schnorr, and S. A. Loening, “Betaine Improves the PCR Amplification of GC-Rich DNA Sequences,” *Nucleic Acids Res*, vol. 25, no. 19, pp. 3957–3958, Oct. 1997.

[93] H. C. Lichstein and M. H. Soule, “Studies of the Effect of Sodium Azide on Microbic Growth and Respiration,” *J Bacteriol*, vol. 47, no. 3, pp. 221–230, Mar. 1944.

[94] D. Nadano, T. Yasuda, and K. Kishi, “Measurement of deoxyribonuclease I activity in human tissues and body fluids by a single radial enzyme-diffusion method.,” *Clinical Chemistry*, vol. 39, no. 3, pp. 448–452, Mar. 1993.

[95] B. Chen, H. R. Costantino, J. Liu, C. C. Hsu, and S. J. Shire, “Influence of calcium ions on the structure and stability of recombinant human deoxyribonuclease I in the aqueous and lyophilized states,” *Journal of Pharmaceutical Sciences*, vol. 88, no. 4, pp. 477–482, Apr. 1999.

[96] S. Fjelstrup *et al.*, “The Effects of Dithiothreitol on DNA,” *Sensors (Basel)*, vol. 17, no. 6, May 2017.

[97] J. E. Nelson and S. A. Krawetz, “Purification of cloned and genomic DNA by guanidine thiocyanate/isobutyl alcohol fractionation,” *Analytical Biochemistry*, vol. 207, no. 1, pp. 197–201, Nov. 1992.

[98] S. Tracy, “Improved rapid methodology for the isolation of nucleic acids from agarose gels,” *Prep. Biochem.*, vol. 11, no. 3, pp. 251–268, 1981.

[99] K. S. Kirby, “[98] Isolation of nucleic acids with phenolic solvents,” in *Methods in Enzymology*, vol. 12, Academic Press, 1968, pp. 87–99.

[100] M. Sunasra, “Performance Metrics for Classification problems in Machine Learning,” *Medium*, 11-Nov-2017. .

[101] V. L. Singer, L. J. Jones, S. T. Yue, and R. P. Haugland, “Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation,” *Anal. Biochem.*, vol. 249, no. 2, pp. 228–238, Jul. 1997.

[102] L. J. Jones, S. T. Yue, C. Y. Cheung, and V. L. Singer, “RNA quantitation by fluorescence-based solution assay: RiboGreen reagent characterization,” *Anal. Biochem.*, vol. 265, no. 2, pp. 368–374, Dec. 1998.

[103] M. Nagai, A. Yoshida, and N. Sato, “Additive effects of bovine serum albumin, dithiothreitol and glycerolon PCR,” *IUBMB Life*, vol. 44, no. 1, pp. 157–163, Jan. 1998.

[104] “What is the composition of elution buffers used in QIAsymphony DNA Investigator kits? - QIAGEN.” [Online]. Available: https://www.qiagen.com/jp/resources/faq?id=38b153d4-0cfc-4f99-bd7e-543ed66fe16f&lang=en&Print=1. [Accessed: 10-Mar-2019].

[105] T. C. Dingle, R. H. Sedlak, L. Cook, and K. R. Jerome, "Tolerance of droplet-digital PCR versus real-time quantitative PCR to inhibitory substances," *Clin Chem*, vol. 59, no. 11, pp. 1670–1672, Nov. 2013.

[106] R. H. Sedlak, J. Kuypers, and K. R. Jerome, "A multiplexed droplet digital PCR assay performs better than qPCR on inhibition prone samples," *Diagnostic Microbiology and Infectious Disease*, vol. 80, no. 4, pp. 285–286, Dec. 2014.

[107] M. Baker, "A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help.," p. 3.

[108] J. W. Tweedie and K. M. Stowell, "Quantification of DNA by agarose gel electrophoresis and analysis of the topoisomers of plasmid and M13 DNA following treatment with a restriction endonuclease or DNA topoisomerase I," *Biochemistry and Molecular Biology Education*, vol. 33, no. 1, pp. 28–33, 2005.

[109] "Cloning vector pCMVbeta, complete sequence," Mar. 1996.

[110] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, Oct. 2011.

[111] "sklearn.feature_selection.VarianceThreshold — scikit-learn 0.21.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html. [Accessed: 31-May-2019].

[112] I. T. Jolliffe, *Principal Component Analysis, Series: Springer Series in Statistics*, 2nd ed. New York: Springer, 2002.

[113] "sklearn.decomposition.PCA — scikit-learn 0.20.3 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html. [Accessed: 20-Mar-2019].

[114] W. H. Greene, "Econometric Analysis," Seventh ed., Boston: Pearson Education, pp. 803–806.

[115] "sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation." [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Accessed: 30-May-2019].

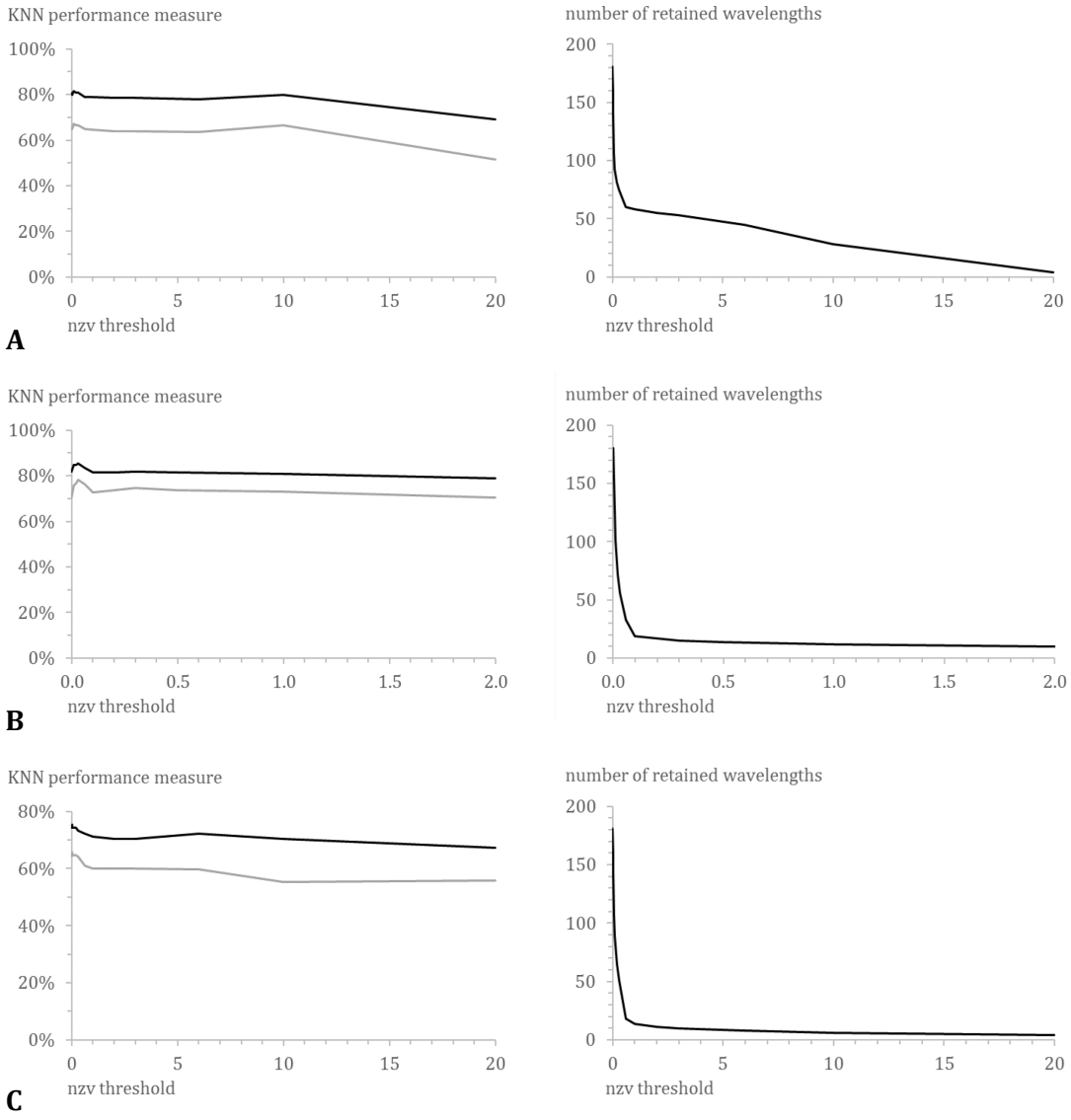
[116] C. M. Bishop, “Pattern Recognition and Machine Learning,” New York: Springer Science + Business Media, LLC, 2006.

[117] “Multiclass logistic regression from scratch — The Straight Dope 0.1 documentation.” [Online]. Available: https://gluon.mxnet.io/chapter02_supervised-learning/softmax-regression-scratch.html. [Accessed: 31-May-2019].

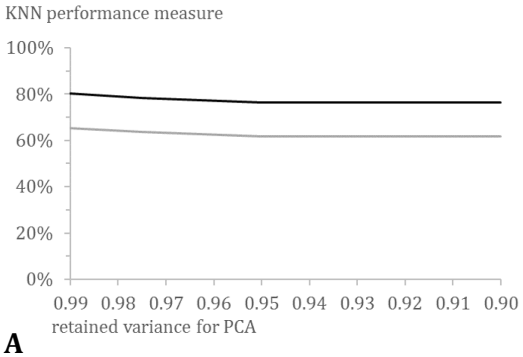
[118] “sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.21.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. [Accessed: 31-May-2019].

[119] K. Zakka, “A Complete Guide to K-Nearest-Neighbors with Applications in Python and R.” [Online]. Available: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/. [Accessed: 25-May-2019].

Appendix

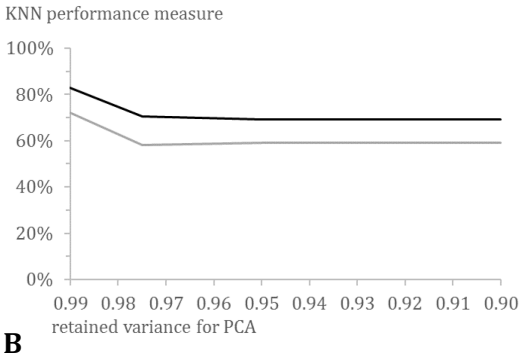


Supplementary Figure 1: (left) KNN performance and (right) number of retained wavelengths for development dataset after data pre-processing with increasing nzv thresholds using (A) raw, (B) A_{260} normalized, or (C) delta spectra. KNN algorithms were run using different thresholds for nvz for feature reduction, with development dataset and default parameter settings: k = 5 and non-weighted distances. On left, accuracy was presented in black and F-measure in grey.



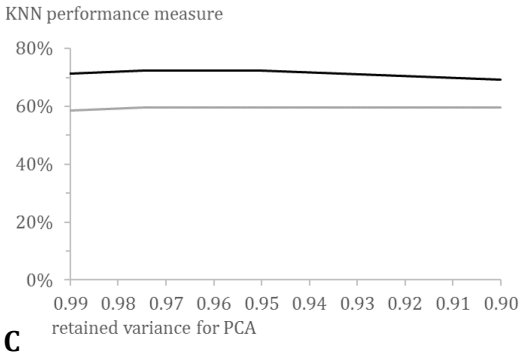
A

retained variance for PCA	number of principal components
0.99	4
0.975	3
0.95	2
0.90	2



B

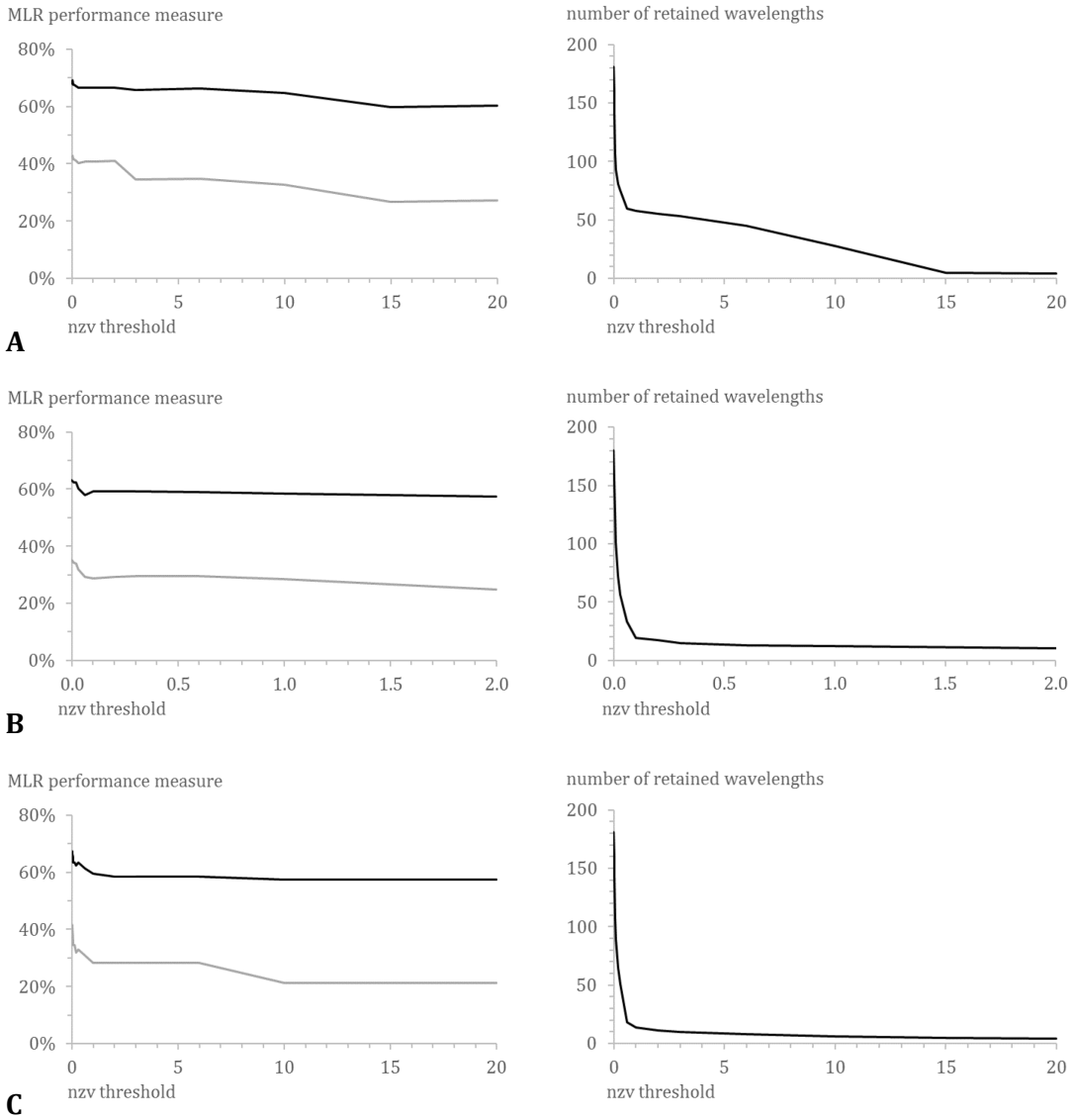
retained variance for PCA	number of principal components
0.99	3
0.975	2
0.95	1
0.90	1



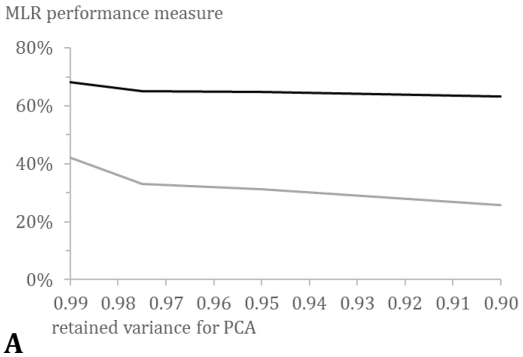
C

retained variance for PCA	number of principal components
0.99	4
0.975	3
0.95	3
0.90	2

Supplementary Figure 2: (left) KNN performance and (right) number of principal components for development dataset after data pre-processing with decreasing variances retained in PCA using (A) raw, (B) A_{260} normalized, or (C) delta spectra. KNN algorithms were run using PCA for feature reduction, with development dataset and default parameter settings: k = 5 and non-weighted distances. On left, accuracy was presented in black and F-measure in grey.

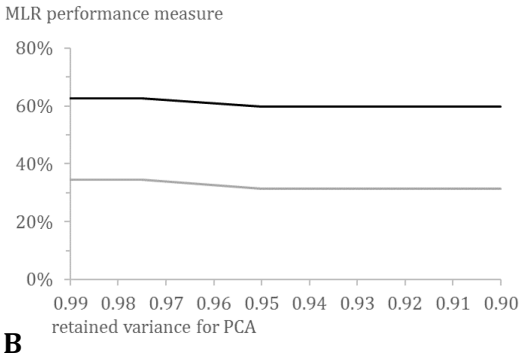


Supplementary Figure 3: (left) MLR performance and (right) number of retained wavelengths for development dataset after data pre-processing with increasing nzv thresholds using (A) raw, (B) A_{260} normalized, or (C) delta spectra. MLR algorithms were run using different thresholds for nvz for feature reduction, with development dataset and default parameter settings: C = 1 and non-weighted classes. On left, accuracy was presented in black and F-measure in grey.



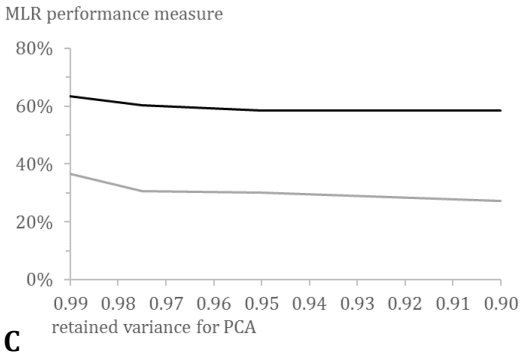
A

retained variance for PCA	number of principal components
0.99	6
0.975	4
0.95	3
0.90	2



B

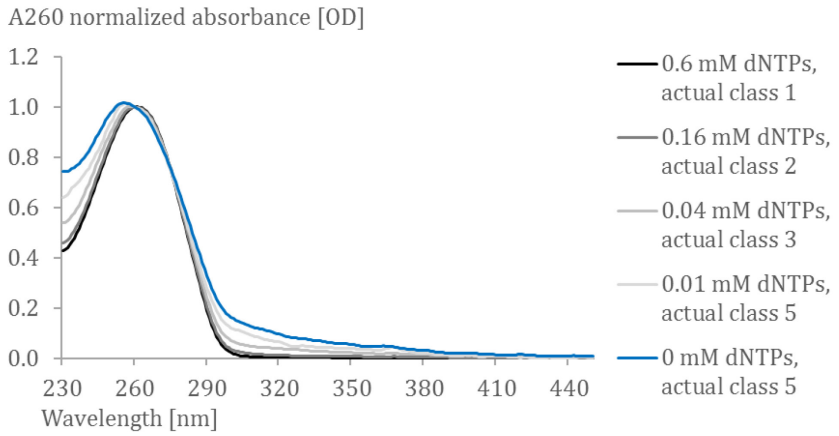
retained variance for PCA	number of principal components
0.99	6
0.975	5
0.95	4
0.90	4



C

retained variance for PCA	number of principal components
0.99	6
0.975	5
0.95	4
0.90	3

Supplementary Figure 4: (left) MLR performance and (right) number of principal components for development dataset after data pre-processing with decreasing variances retained in PCA using (A) raw, (B) A_{260} normalized, or (C) delta spectra. MLR algorithms were run using PCA for feature reduction, with development dataset and default parameter settings: C = 1 and non-weighted classes. On left, accuracy was presented in black and F-measure in grey.



Supplementary Figure 5: A_{260} normalized spectra of DNA samples contaminated with dNTPs. Presented were 1 measurement replicate of 5 DNA samples with or without dNTPs. Samples were applied on enzyme activity assay and actual classes were assigned based on measured enzyme activities [%], with $c1 \leq 20\% < c2 \leq 40\% < c3 \leq 60 < c4 \leq 80\% < c5$.

Supplementary Table 1: R^2 values of standard curves from 4 independent runs of Phi-Inhibition-Assay. On each run, 4 standards with decreasing polymerase concentration were applied in 4 measurement replicates.

Run	R^2
1	1.00
2	1.00
3	1.00
4	0.99

Supplementary Table 2: Mean polymerase and ligase activity measured after contamination of DNA samples, and corresponding standard deviations. N = 6 independent runs.

Contaminant	Concentration		polymerase activity [%]		ligase activity [%]	
			mean	StDev	mean	StDev
betaine	0.1	mg/µL	93.5	9.3	117.7	7.5
	0.05	mg/µL	96.4	7.1	108.3	4.3
	0.025	mg/µL	99.1	8.9	103.1	6.1
	0.013	mg/µL	102.4	6.8	95.6	7.9
	0	mg/µL	99.5	5.1	99.1	9.4
dNTPs	0.6	mM	11.7	4.0	-0.1	2.2
	0.16	mM	25.3	3.2	-0.1	2.2
	0.04	mM	56.7	3.9	51.6	4.3
	0.01	mM	96.4	10.8	86.4	7.9
	0	mM	98.6	11.1	102.9	4.3
DTT	10	mM	98.5	11.6	73.4	1.9
	5	mM	101.3	11.0	80.8	5.2
	2.5	mM	102.2	8.1	88.4	4.6
	1.25	mM	102.5	9.9	93.9	4.4
	0	mM	107.9	6.3	99.8	4.4
EDTA	2	mM	0.9	3.0	99.1	3.8
	1.18	mM	33.8	14.4	99.0	8.4
	0.69	mM	78.0	13.4	103.4	6.2
	0.41	mM	93.2	6.1	104.9	10.4
	0	mM	99.1	9.3	99.4	7.4
glycogen	1.5	mg/mL	64.7	5.8	71.7	10.1
	1.2	mg/mL	67.5	5.4	77.3	8.5
	0.96	mg/mL	69.7	7.1	74.9	8.6
	0.77	mg/mL	73.7	6.9	77.2	7.0
	0	mg/mL	99.0	9.5	98.2	13.2
GITC	20	mM	71.2	10.0	69.1	7.0
	5	mM	97.1	9.5	86.0	10.6
	1.25	mM	104.3	14.7	83.1	7.1
	0.31	mM	101.1	11.5	89.2	6.9
	0	mM	99.5	6.2	99.0	10.2
HB	1	mg/mL	3.3	3.8	123.3	10.4
	0.3	mg/mL	11.2	3.9	134.2	20.6
	0.09	mg/mL	50.3	10.5	140.2	11.4
	0.027	mg/mL	93.5	14.4	109.6	7.2
	0	mg/mL	99.1	8.9	99.7	4.9

Contaminant	Concentration		polymerase activity [%]		ligase activity [%]	
			mean	StDev	mean	StDev
HSA	5.5	mg/mL	42.1	6.2	110.7	9.4
	2.75	mg/mL	55.3	7.8	129.4	7.3
	1.38	mg/mL	63.2	5.4	147.4	12.4
	0.69	mg/mL	81.1	9.1	151.0	14.7
	0	mg/mL	99.5	6.8	94.6	6.8
IgG	1.8	mg/mL	-0.1	2.7	4.1	2.2
	0.9	mg/mL	-2.6	3.7	4.1	2.2
	0.45	mg/mL	36.5	2.8	2.5	2.2
	0.23	mg/mL	75.6	4.9	81.8	8.0
	0	mg/mL	99.4	6.8	101.6	12.9
NA	0.1	% (w/v)	73.2	10.6	82.4	6.7
	0.05	% (w/v)	86.1	6.0	88.0	4.5
	0.025	% (w/v)	90.3	6.2	93.1	8.4
	0.013	% (w/v)	90.7	8.9	94.6	6.3
	0	% (w/v)	99.1	8.7	92.7	5.5
Na-Citrat	40	mM	0.5	2.8	6.1	0.3
	25	mM	0.8	2.9	12.0	3.7
	15.6	mM	0.9	3.1	30.6	3.5
	9.7	mM	0.9	3.0	61.3	4.3
	0	mM	99.2	8.4	82.1	3.5
Phenol	7.5	mM	22.5	4.4	5.6	0.3
	2.5	mM	49.0	7.6	4.7	2.7
	0.83	mM	85.5	11.1	58.6	4.4
	0.28	mM	112.8	9.8	67.1	3.8
	0	mM	99.7	4.9	82.6	3.9

Acknowledgment

My gratitude goes to Dr. Daniel Lehmann, Jun.-Prof. Dr. Ilka Axmann and Prof. Dr. Markus Kollmann for giving me the opportunity to explore professional life in a biotech company while preparing my doctoral dissertation. I also greatly appreciate the opportunity to explore scripting, learning about, and working with mathematical data modelling in addition to conducting biological experimental research.

I want to express my gratitude to Dr. Daniel Lehmann, who gave me the freedom and resources to work independently, provided the chance to be part of a project team and acquire basic knowledge of project management, and always took time to help me solve difficulties I encountered or discuss results. Thank you, Prof. Dr. Ilka Axmann for your mentorship, support and scientific discussions, and thank you, Prof. Dr. Markus Kollmann for sharing your knowledge and support throughout preparation of the mathematical part of this thesis.

My thanks go to the whole Instrument Detection team of the Life Science R&D department at QIAGEN Hilden for receiving me as a team member again. Working with you was a great and educational experience and this thesis would not have been possible without the help of so many people who in one way or the other contributed their valuable assistance. Thank you for sharing your workspace, your knowledge and your coffee breaks that were always accompanied with laughter and lively discussions.

Special thanks go to Nicole Lokmer for supporting me with my laboratory work, experimental design and the discussion of results.

I greatly thank Katharina Pfeifer-Sancar and Pierre-Henri Ferdinand for sharing their knowledge and experience in scientific discussions and for proofreading my thesis.

My gratitude also goes to Nicolas Schmelling for patiently introducing me to the world of python scripting.

Last but not least, I want to thank Jaime, my friends and family for their support and company along the way.

Thank you!