technische universität
dortmund

Technische Universität Dortmund

Fakultät Statistik

Dissertation

Thema: **Statistische Analyse von hochdimensionalen toxikologischen Expressionsdaten**

Autor: **Eugen Rempel**

Vorgelegt: Dortmund, 30.05.2016
Tag der mündlichen Prüfung: 23.09.2016

Erstgutachter: Prof. Dr. Jörg Rahnenführer
Zweitgutachter: Prof. Dr. Claus Weihs
Drittgutachter: Prof. Dr. Jan Hengstler
Kommisionsvorsitz: Prof. Dr. Katja Ickstadt

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listingverzeichnis

Glossar

DNA

Desoxyribonukleinsäure (englisch DNA) ist ein in allen Lebewesen vorkommendes Biomolekül. Es überträgt die Erbinformation in Form von Genen auf die Nachkommen 19

Epigenetik

Der Begriff Epigenetik umschreibt Mechanismen und Konsequenzen vererbbarer Chromosomen-Modifikationen, die nicht auf Veränderungen der DNA-Sequenz beruhen 97

Gene Ontology

Eine biomedizinische Ontologie, die einen der drei Bereiche abdeckt: „Zelluläre Komponente“, „Biologischer Prozess“ und „Molekulare Funktion“. Jeder Terminus besteht aus einem Namen, einer Nummer und assoziierten Daten. Jede Ontologie hat die Topologie eines gerichteten azyklischen Graphen 25

Genom

die Gesamtheit aller Träger vererbbarer Information einer Zelle 19

Histondeacetylase-Inhibitor

Als Histondeacetylase-Inhibitoren (**HDACi**) werden Substanzen bezeichnet, die in der Lage sind Histondeacetylasen zu hemmen und somit eine Hyperacetylierung der Histone zu bewirken 30, 32

Hybridisierung

Spezifische Bindung zwischen zwei komplementären Polymeren, z.B. zwei Nukleotidsträngen 24

Klassifikationsstudie

Untersuchung der Wirkung von verschiedenen Substanzen auf ein Testsystem, insbesondere der Möglichkeit ihrer Klassifikation 14

Konzentrationsstudie

Untersuchung der Wirkung von verschiedenen Konzentrationen einer Substanz auf ein Testsystem 14

Pluripotenz

Fähigkeit der Stammzellen sich zu jedem Zellentyp des Organismus zu differenzieren 22

Probe

Eine Nukleotidensequenz der Länge von 25 Basen. Sie wird komplementär zu einem bekannten Genabschnitt auf dem Chip synthetisiert 24

Probeset

Eine Gruppe von zu einem bestimmten Gen korrespondierenden Proben 24

RNA

Ribonukleinsäure (englisch RNA) ist eine Nukleinsäure, die sich als Polynukleotid aus mehreren Nukleotiden zusammensetzt. Eine wesentliche Funktion der RNA ist der Informationsübertragung 19

Testsystem

Repräsentiert einfache oder komplexe physiologische oder biochemische Prozesse bzw. Reaktion des Gesamtorganismus. Mit ihnen können Aspekte der Wirkung, Nebenwirkung und Toxizität von Substanzen aufgeklärt werden 22

Transkription

Transkription bezeichnet in der Genetik die Synthese von RNA anhand einer DNA als Vorlage 20

Transkriptionsfaktor

Transkriptionsfaktor bezeichnet in der Genetik ein Protein, welches spezifische DNA-Sequenzen (Transkriptionsfaktor-Bindestellen) bindet und somit die Transkription der genetischen Information kontrolliert und steuert 22

Transkriptomik

Transkriptomik bezeichnet die Erforschung aller Gene die als mRNA (Boten-RNA) vorliegen. Die mRNA ist eine Abschrift der Gene. Sie wird bei der Transkription produziert. Transkription ist der erste Schritt der Proteinbiosynthese, bei der anhand der Baupläne der Erbinformation Eiweiße aus entsprechenden Aminosäurebausteinen entstehen. Die Boten-RNA dient dabei als Indikator für die Aktivität von Genen 20

UKK

An der Universität zu Köln entwickeltes in-vitro Testsystem zu Untersuchung von embryonalen Stammzellen. Übersicht siehe Abbildung 4 22

UKN1

An der Universität zu Konstanz entwickeltes in-vitro Testsystem zu Untersuchung von embryonalen Stammzellen. Übersicht siehe Abbildung 4 22

Zeitfensterstudie

Untersuchung der Wirkung bestimmter Substanz auf ein Testsystem, wobei die Einwirkperiode unterschiedlich gewählt wird 14

1 Einleitung

Der technische Fortschritt im 20. Jahrhundert brachte sowohl neue Produktionsmethoden und Erzeugnisse als auch damit verbundene Herausforderungen mit sich. Die Umweltbelastung stellt dabei eines der wichtigsten Probleme unserer Zeit dar. Zu den bekannten umweltverschmutzenden Stoffen wie Schwermetallen, Kohlendioxid oder Schwefeldioxid kommen neue, vom Menschen erzeugte chemischen Substanzen dazu. Momentan sind über 112 Millionen organischer und anorganischen Substanzen bekannt (entnommen www.cas.org, Stand: 29. September 2016) und es kommen täglich ca. 4 Tausend neue hinzu [Binetti u. a. (2008)]. Gemäß der EU-Chemikalienverordnung Nr. 1907/2006 (REACH: Registration, Evaluation, Authorisation and Restriction of Chemicals) müssen die neuen Stoffe geprüft und registriert sein, ehe sie in Verkehr gebracht werden. Diese Prüfung beinhaltet auch die **Toxizitätsbestimmung** der neuen Substanz. Darunter versteht man die Feststellung der Giftigkeit oder Schädlichkeit eines Stoffes. Dabei ist es unerlässlich die durchgeführten Prüfungen mit der benötigten wissenschaftlichen Rigorosität und unter Einhaltung anerkannter Prinzipien durchzuführen [Harris u. a. (2014)]. Im Rahmen dieser Untersuchung werden unter anderem die Karzinogenität und die Reproduktions- bzw. Entwicklungstoxizität der Substanz analysiert. Das Letzte schließt Beeinträchtigungen der Entwicklung des Kindes/Embryos vor und nach der Geburt auf Grund einer Exposition eines Elternteils oder des Kindes selbst ein. Das sich entwickelnde Nervensystem ist dabei besonders empfänglich gegenüber äußeren Belastungen [Watkins u. a. (2010), van Thriel u. a. (2012)].

Es ist wünschenswert diese Untersuchungen in-vitro durchzuführen, um auf aufwendige und ethisch fragwürdige Tierexperimente zu verzichten. Deshalb unterstützte die Europäische Kommission im Rahmen der Förderung FP7 das Projekt **ESNATS** (embryonic stem cells derived novel alternative test systems). In der vorliegender Arbeit werden Daten analysiert, welche während dieses Projektes in den Universitäten zu Köln bzw. Konstanz erhoben wurden. Es handelt sich um 6 Studien, welche entweder den Einfluss von steigender Konzentration bestimmter toxischer Substanz unter die Lupe nahmen (Konzentrationsstudie), die Reaktion der Stammzellen auf Substanzen mit verschiedenen Einwirkzeiten analysierten (Zeitfensterstudie) oder die Möglichkeit einer Klassifizierung verschiedener Typen von Substanzen untersuchten (Klassifikationsstudie). Dazu wurde die „Aktivität“ mehrerer Tausend Gene unter Anwendung des Chips des Herstellers **Affymetrix** gemessen und analysiert.

Bei diesen Analysen spielen die statistischen Methoden eine bedeutende Rolle. Sowohl bei der Bildanalyse, der Normalisierung der Genexpression als auch der Modellierung werden statistischen Verfahren vermehrt angewendet. Ferner basieren verschiedene Methoden der Visualisierung, Bestimmung von differentiellen Genen und angereicherten biologischen Signaturen auf statistischen Modellen. So setzt auch diese Arbeit sich zum Ziel, ausgehend

von hochdimensionalen toxikologischen Daten unter Verwendung von verschiedenen statistischen Methoden eine Abfolge von Verfahren zur sequenziellen Analyse bereitzustellen und anzuwenden. Damit wird eine Methode vorgestellt, welche dem Anwender ermöglicht, eine transparente und reproduzierbare Datenanalyse durchzuführen. In Abbildung 1 ist der schematische Ablauf der Analyse (auch als „Pipeline“ bezeichnet) dargestellt. Eine spezielle Aufgabe der Statistik stellt die Zuordnung von neuen Substanzen in die vordefinierten Klassen dar. Dies wird als Klassifikation oder Diskriminanzanalyse bezeichnet. In dieser Arbeit werden vom Verfasser zwei Verfahren hinsichtlich der Klassifikationsgüte untersucht, welche sich in der täglichen Praxis großer Popularität erfreuen: Support Vector Machines und Random Forests. Dabei wird sowohl der Einfluss von Variablenauswahl als auch die Abhängigkeit der Klassifikationsgüte von der Replikatenanzahl unter die Lupe genommen. Das Erstere ist insofern von Bedeutung, als es nach wissenschaftlichen Erkenntnissen sinnvoll sein kann, die zugrundeliegende Variablenmenge zu reduzieren. Der zweite Punkt kann zur Lösung der Frage beitragen, wie viele Replikate man innerhalb einer Studie braucht. Ferner wird die Robustheit beider Verfahren gegenüber dem Rauschen analysiert (Sensitivitätsanalyse). Diese Untersuchung soll beleuchten, welchen Einfluss ein zu den erklärenden Variablen hinzugefügtes Rauschen auf die Klassifikationsgüte ausübt. Auf Grund der Tatsache, dass die biologischen hochdimensionalen Daten fehlerbehaftet sind, würde ein robusteres Klassifikationsverfahren verlässlichere Ergebnisse liefern.

Der Verfasser möchte an dieser Stelle hinweisen, dass die in dieser Arbeit erläuterte Ergebnisse und Verfahren zum Teil bereits in verschiedenen Publikationen präsentiert wurden. So erschienen die Analysen der Klassifikationsstudie UKN1 teilweise in Rempel u. a. (2015). Namentlich die Ergebnisse des Variablenauswahls, der Kreuzvalidierung und der Auswertung an einem externen Datensatz wurden erörtert. Neu dagegen sind in der vorliegenden Arbeit die vorgestellten Sensitivitätsanalysen und die Untersuchung, inwieweit die Anzahl der Replikate einen Einfluss auf die Klassifikationsergebnisse ausübt. Die Ergebnisse der VPA Chronic Konzentrationsstudie wurden in Waldmann u. a. (2014) präsentiert. Dort ist auch eine biologische Interpretation zu finden. Ferner kam die vorgestellte Pipeline teilweise in Balmer u. a. (2014) und Krug u. a. (2013) zur Anwendung.

Nun folgt eine kurze Übersicht über die Aufbau der vorliegenden Arbeit. Kapitel 2 beginnt mit einer kurzen Einführung in die biologischen Hintergründe der Genexpressionsmessung. Das zentrale Dogma der Molekularbiologie wird erläutert und die Bedeutung der Stammzellen für die Toxikologie verdeutlicht. Des Weiteren wird die Microarray-Technologie beschrieben. Anschließend werden die zur Verfügung stehenden Daten vorgestellt. Zum Ende des Kapitels werden die Ziele dieser Arbeit genauer beschrieben.

Kapitel 3 stellt die statistische Methoden und Verfahren vor, welche bei den späteren Analysen verwendet werden. Zunächst werden die Methoden der deskriptiven Analyse erläutert: Hauptkomponentenanalyse und Heatmap. In einem weiteren Unterkapitel wird der moderierte t-Test als Alternative zum Zweistichproben-t-Test beschrieben. Ferner wird

auf die Problematik des Batch-Effektes eingegangen und eine Methode zu dessen Modellierung und Verminderung vorgestellt. Anschließend werden zwei wichtige Verfahren der Diskriminanzanalyse beschrieben: Support Vector Machine (SVM) und Random Forests (RF). Zum Ende des Kapitels wird eine Methode zur Sensitivitätsanalyse vorgestellt und erläutert.

Im Kapitel 4 werden die vorgestellten Methoden auf die in dem Kapitel 2 erläuterten Daten angewendet und Ergebnisse präsentiert. Die ersten Abschnitte 4.1-4.4 dienen dabei der Vorstellung von Analysergebnissen von 4 Konzentrationsstudien. Die Daten werden sowohl deskriptiv als auch explorativ untersucht. In den Abschnitten 4.5 und 4.6 werden die Untersuchungsresultate der Klassifikationsstudien erörtert. Dabei werden die beiden Klassifikationsverfahren SVM und Random Forests parallel auf beide Studien angewendet. Für alle vier Paare aus Datensatz und Methode werden die Ergebnisse in eigenen Unterkapiteln präsentiert. Ein großer Teil der Resultate wird dabei im Anhang (ab Seite 99) gezeigt.

Im Kapitel 5 folgt dann eine Zusammenfassung und Diskussion.

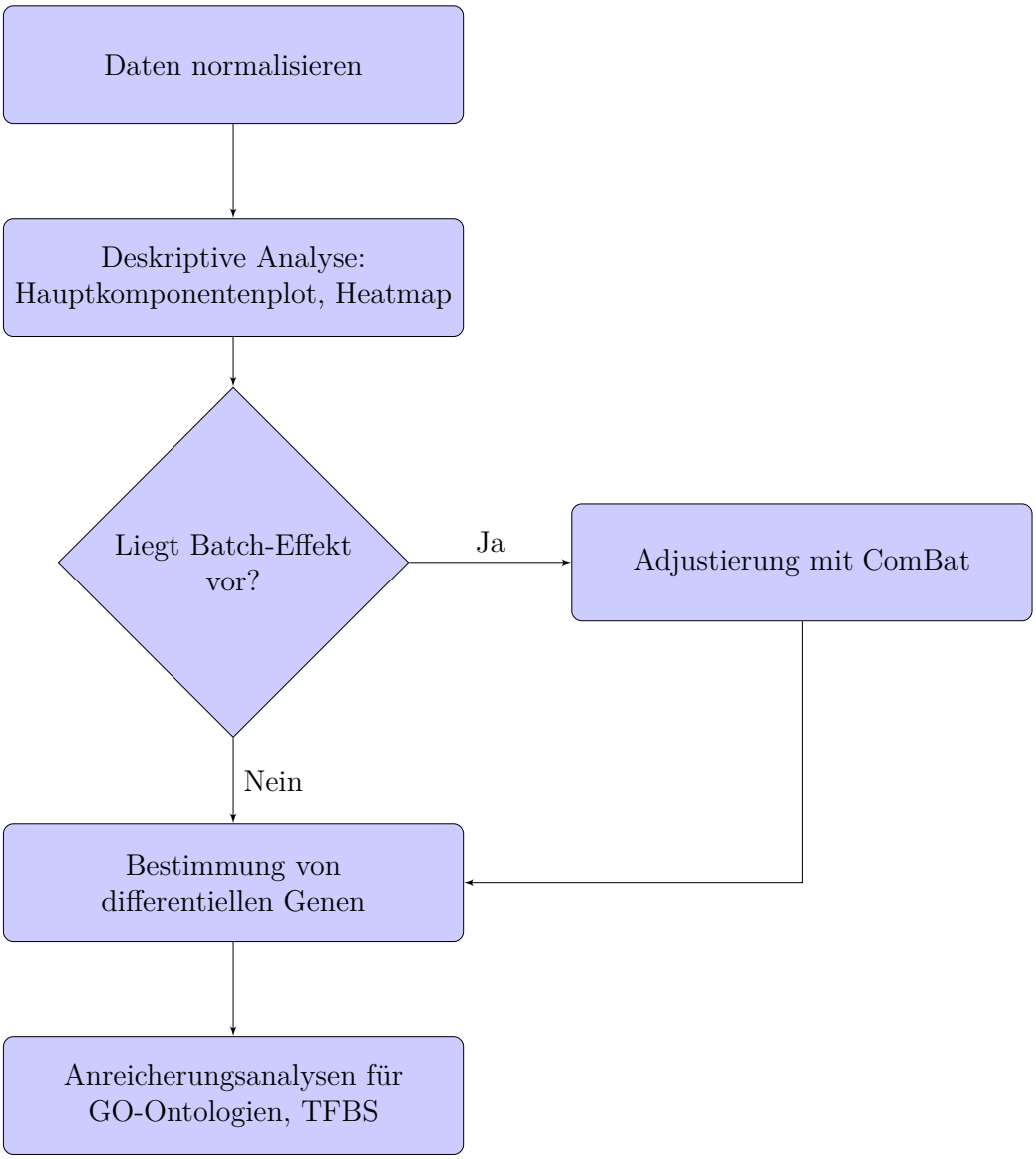

Abbildung 1: Eine schematische Abfolge von Schritten zur Analyse von hochdimensionalen Microarray-Daten (Pipeline). Nach dem Normalisieren werden die Daten zuerst deskriptiv analysiert. Mit Hilfe dieser Analysen wird entschieden, ob ein Batch-Effekt vorliegt. Im positiven Falle werden die Daten unter Verwendung des **ComBat**-Algorithums adjustiert. Im nächsten Schritt werden die differentiell exprimierten Probesets bestimmt und auf Anreicherung in biologischen Signaturen wie GO-Gruppen oder Bindestellen von Transkriptionsfaktoren (**TFBS**) untersucht.

2 Biologischer Hintergrund und Datensätze

Dieses Kapitel dient dem Zweck, dem Leser sowohl die biologischen Hintergründe als auch die Methoden der Messung von Genexpression nahe zu bringen. Im Rahmen dieser Arbeit wird die Genaktivität im Zuge der Differenzierung von humanen Stammzellen untersucht. Demzufolge wird in dem Unterkapitel 2.1 zuerst erläutert, was man unter „Genaktivität“ versteht und inwiefern dies zum Verständnis zellulärer Prozesse beiträgt. Ferner werden die Besonderheiten der Stammzellen vorgestellt. Zum Messen der Genaktivität bzw. der Transkription werden in der Wissenschaft verschiedene Methoden verwendet, z.B. Microarray-Technologie und Sequenzierungsverfahren. In dieser Arbeit verwendete Daten wurden ausschließlich mit Hilfe der Microarray-Technologie gewonnen. Im Unterkapitel 2.2 wird sie kurz dargestellt. Ferner werden in Unterkapitel 2.3 die verwendeten Daten vorgestellt, die in Kapitel 4 analysiert werden. Abschließend wird im Unterkapitel 2.4 auf die Ziele dieser Arbeit eingegangen.

2.1 Biologischer Hintergrund

Im Rahmen dieser Arbeit wird die „Genaktivität“ humaner Stammzellen untersucht. Eine besondere Rolle nimmt dabei die Idee des Informationsflusses ein: Die zur Bildung von Proteinen - wichtigsten Bausteinen der Zelle - benötigte Information ist auf spezialisierten Molekülen codiert. Diese Moleküle befinden sich im Zellkern. Deshalb muss diese Information auf irgendeine Weise in das Zellinnere transportiert und zur Proteinbildung „übersetzt“ werden. Die im Jahre 1958 von Francis Crick publizierte Hypothese beschreibt diese Übertragung von Information. Diese Hypothese ist von entscheidender Bedeutung und wird als das zentrale Dogma der Molekularbiologie bezeichnet. In Unterkapitel 2.1.1 geht der Verfasser detaillierter auf diese Problematik ein. Im weiteren Verlauf werden die biologischen Hintergründe der humanen Stammzellen verdeutlicht und ihre Wichtigkeit für die medizinische Forschung erläutert. Anschließend werden zwei Möglichkeiten vorgestellt, die Expression einzelner Gene zum Verständnis allgemeiner zellulärer Prozesse zu verwenden: das Konzept der Gene Ontology und die Analyse von Bindestellen der Transkriptionsfaktoren.

2.1.1 Zentrales Dogma der Molekularbiologie

Alle lebenden Organismen enthalten Information für deren Entwicklung und Funktion. Diese Information wird als genetischer Inhalt bezeichnet und ist in den Genen kodiert, welche somit als Informationseinheiten fungieren. Die Menge aller Gene eines Organismus wird als das **Genom** bezeichnet. Die funktionellen Einheiten des Genoms sind bei den meisten Organismen die Moleküle der *Desoxyribonukleinsäure* (**DNA**), während bei einigen Viren das Genom aus *Ribonukleinsäure* (**RNA**) gebaut wird. Ein einzelnes DNA-Molekül wird aus vier Nukleotiden zusammengesetzt: Adenin, Thymin, Cytosin und Guanin, oft abgekürzt mit **A**, **T**, **C** und **G**. Einzelne Nukleotide reihen sich aneinander an und bilden eine lange Kette (Polymer). Ein DNA-Molekül besteht nun aus zwei Nukleotidketten (-strängen), welche auf die Art einer Doppelhelix umeinander gewickelt sind. Die Nukleotidstränge werden dabei durch spezielle Kräfte zusammengehalten: die Wasserstoffbrücken. Diese anziehende Wechselwirkung trifft zwischen folgenden Nukleotidenpaaren auf: Adenin und Thymin bzw. Cytosin und Guanin, siehe Abbildung 2. Eine leicht geänderte Zusammensetzung präsentiert sich in dem Aufbau von RNA: Das Nukleotid Thymin wird durch Uracil ersetzt.

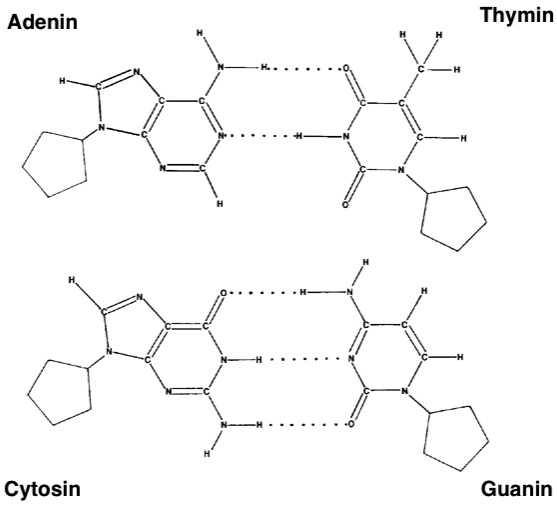

Abbildung 2: DNA Moleküle bestehen aus 4 Nukleotiden: Adenin, Thymin, Cytosin und Guanin. Die gestrichelten Linien stellen die Wasserstoffbrücken dar, welche die zwei einzelne Nukleotidenstränge zusammenhalten. Diese spezifische Zuordnung wird auch als Watson-Crick-Basenpaarung bezeichnet. Man beachte, dass sich zwischen Cytosin und Guanin drei Wasserstoffbrücken bilden und zwischen Adenin und Thymin nur zwei.

Die durch eine spezifische Sequenz der Nukleotide kodierte genetische Information wird mit Hilfe der sogenannten *Boten-RNA* (englisch: *messenger RNA* **mRNA**) in die Sequenz der Aminosäuren übersetzt. Dazu werden zuerst im Zellkern die mRNA-Moleküle komplementär zu dem DNA-Strang erstellt. Dieser Vorgang wird als **Transkription** bezeichnet. Ferner wird die Menge aller zu einem bestimmten Zeitpunkt hergestellten RNA-Moleküle als **Transkriptom** definiert. Die Abfolge der Nukleotide in der Boten-RNA muß im nächsten Schritt in die Abfolge der Aminosäuren übersetzt werden. Dazu wandern die mRNA-Moleküle in das Zellinnere und binden sich an das Ribosom. Das Ribosom ist eine besondere Struktur und katalysiert als **Translation** die Sequenz der Nukleotide in die Sequenz der Aminosäuren. Die Reihenfolge der Aminosäuren wiederum definiert die räumliche Struktur des Proteins und somit seine Funktion. Dieser Informationsfluss von DNA über mRNA zu Proteinen wird als das **zentrale Dogma** der Molekularbiologie bezeichnet (zur schematischen Darstellung betrachte Abbildung 3) und wurde von Francis Crick im Jahre 1958 als Hypothese vorgestellt. Es motiviert somit die Untersuchung von dynamischen Veränderungen des Transkriptoms - die Transkriptomik. Die Dynamik der Genexpression lässt Schlüsse auf die zellulären Veränderungen auf der Ebene der Proteine zu, was wiederum die Funktion der gesamten Zelle bzw. des gesamten Organismus bestimmt. Somit könnte das Wissen über die Änderung der Genexpression auf Grund von Behandlung, Krankheit oder anderen externen Stimuli zum Entwickeln besserer Methoden zur Diagnose und Vorhersage verwendet werden. Die in Kapitel 2.2 vorgestellte Microarray-Technologie wird zu der Untersuchung jener Dynamik verwendet: Sie stellt eine Möglichkeit dar, eine Momentaufname des Transkriptoms zu erstellen, welche dann mit Hilfe von verschiedenen statistischen Methoden aufgearbeitet und analysiert wird.

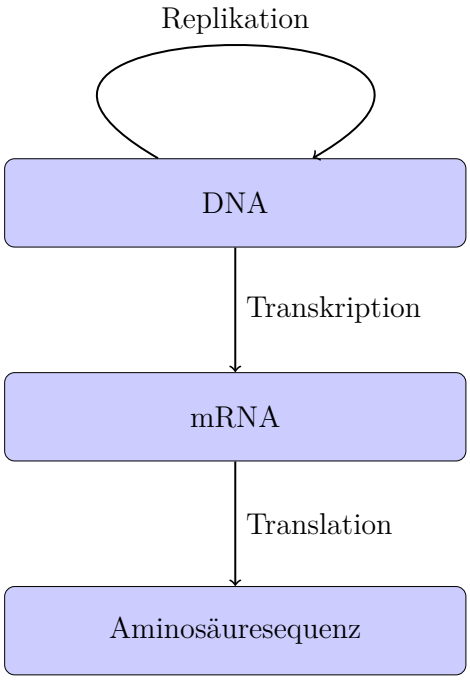

Abbildung 3: Das zentrale Dogma der Molekularbiologie: Die in DNA kodierte Information wird mit Hilfe von RNA-Molekülen für die Synthese von Proteinen verwendet. Dazu wird zuerst im Zellkern die DNA in RNA transkribiert. Die RNA-Moleküle wandern in das Zellinnere und werden mit Hilfe von Ribosomen in die Aminosäuresequenz übersetzt. Die Reihenfolge der Aminosäuren definiert die räumliche Struktur und somit die Funktion des Proteins. Im Laufe der Zellteilung werden die DNA-Moleküle repliziert und auf die Tochterzellen verteilt.

2.1.2 Biologie und Verwendung der embryonalen Stammzellen

In diesem Unterkapitel werden die biologischen Hintergründe der embryonalen humanen Stammzellen (englisch: *embryonic human stem cells* **eHSC**) erläutert.

Embryonale Stammzellen werden aus den Embryoblasten gewonnen, welche im Rahmen einer In-Vitro-Fertilisation entstehen. Dies stellt in einigen Fällen die einzige Möglichkeit dar, einem (Ehe-)Paar ihren Kinderwunsch zu erfüllen. Ist eine Reimplantation der befruchteten Eizelle (Zygote) in die zukünftige Mutter erfolgreich, so werden die übriggebliebene Embryonen entweder entsorgt, einem anderen (Ehe-)Paar gespendet oder für wissenschaftliche Zwecke verwendet. Im letzten Falle wird die äußere Wand der Blastozyste entfernt, um das Innere, welches als Embryoblast bezeichnet wird, freizugeben. Zu diesem Zeitpunkt sind die Embryonen 4-5 Tage alt und bestehen aus 50-150 Zellen. Da die Gewinnung der eHSC mit der Zerstörung der Blastozyste einhergeht, ist die Stammzellenforschung mit ethischen Fragen verbunden und wird kontrovers diskutiert (siehe u.a. McLaren (2001), Beeson u. Lippman (2006)).

Die embryonalen Stammzellen sind vor allem auf Grund ihrer Eigenschaft, sich in fast alle anderen Zelltypen differenzieren zu können (Pluripotenz), interessant. So kann eine Stammzelle je nach äußeren Stimuli sich in eine Muskel-, Leber- oder Nervenzelle (Neuron) entwickeln. Umweltbelastungen wie Toxine können diesen Prozess erheblich stören. Vor allem ist die Differenzierung zu einem Neuronen chemischen Einflüssen gegenüber empfindlich [Tamm u. a. (2006)]. Tierexperimente bieten eine verbreitete Möglichkeit auf Neurotoxizität zu analysieren. Diese Untersuchungen sind allerdings mit folgenden Problemen verbunden: Kosten, Länge des Experiments, Übertragbarkeit auf Menschen und ethische Bedenken. Ein auf menschlichen Zellen basierendes in-vitro Testsystem erlaubt nun den Wissenschaftlern auf die Tierexperimente zu verzichten.

An dem in-vitro rekapitulierten Differenzierungsprozess von einer Stammzelle bis zu einem Neuron lassen sich viele Vorgänge, wie z.B. Proliferation, Migration und Apoptose beobachten, beeinflussen und eventuell steuern. Dies erlaubt Experimentatoren auf aufwendige und ethisch fragwürdige Tierexperimente zu verzichten. Deshalb wurden an den Universitäten zu Konstanz und Köln zwei verschiedene Testsysteme entwickelt, welche im weiteren Verlauf der Arbeit als UKN1 und UKK bezeichnet werden. Diese rekapitulieren verschiedene Phasen früher Gewebespezialisierung und neuraler Entwicklung (vergleiche Abbildung 4).

In dem Testsystem UKK wird vor allem die multipotente Differenzierung von hESC in Ekto-, Meso- und Endoderm untersucht [Jagtap u. a. (2011), Meganathan u. a. (2012)]. Dazu werden die undifferenzierten Stammzellen durch Hinzugabe von verschiedenen Transkriptionsfaktoren zum Differenzieren veranlasst. Transkriptionsfaktoren sind Proteine, welche sich im Zellkern befinden und an den DNA-Strang haften können. Ein gebundener Transkriptionsfaktor bildet mit der RNA-Polymerase einen Komplex und initiiert somit

die Transkription des DNA-Moleküls in die RNA-Nukleotidensequenz. Die Biologie der Transkriptionsfaktoren wird in Unterkapitel 2.2.2 ausführlicher erläutert.

Nach dem Veranlassen der Differenzierung werden die Zellen auf eine Unterlage gebracht. Der Versorgung der Zellen mit den lebenswichtigen Nährstoffen dient das Medium. In regelmäßigen Abständen müssen das Medium und die Unterlage gewechselt werden. Je nach Studiendesign werden die Zellen für verschiedene Zeitintervalle der Wirkung einer toxischen Substanz ausgesetzt. Nach einer vorgeschriebenen Zeit werden die Zellen am Tag der Analyse „geerntet“: das Medium wird abgewaschen, die Zellwände gelöst und das mRNA-Material extrahiert. Im UKK Testsystem werden die Zellen nach 14 Tagen geerntet.

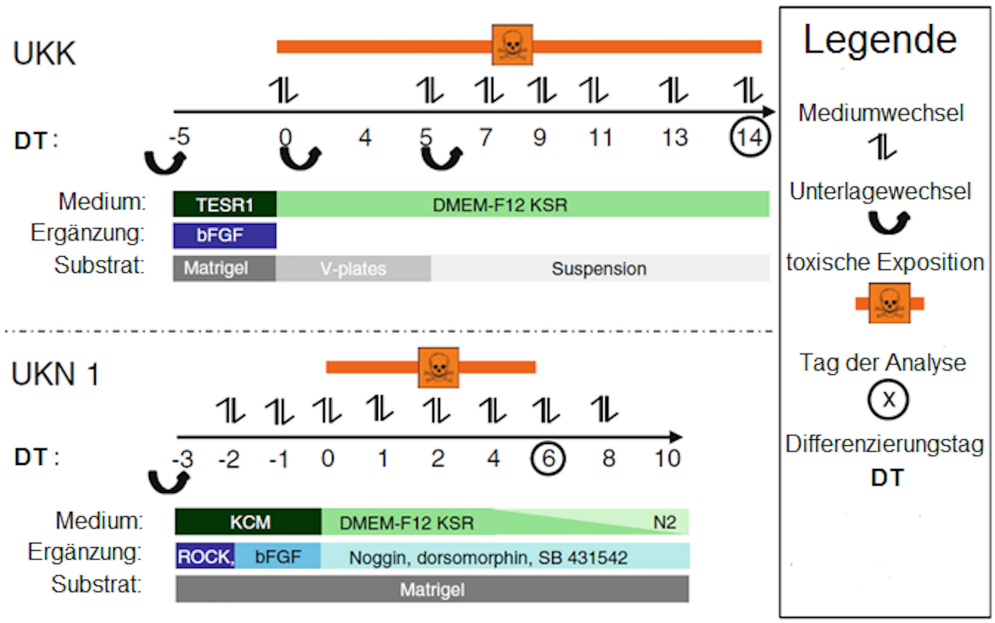

Abbildung 4: Übersicht über die UKK und UKN1 Testsysteme (adaptiert aus Krug u. a. (2013)). Die Zeitachse gibt den Zeitpunkt des Austausches von Medium oder Unterlage, Hinzugabe der toxischen Substanz und Analyse an. Zusätzlich ist die Information zu dem Typ der Beschichtung und Wahl des Mediums zu verschiedenen Phasen des Experiments dargestellt.

Das Testsystem UKN1 modelliert den Zustand der neuroektodermalen Induktion, woraus neuroektodermalen Vorläuferzellen entstehen [Balmer u. a. (2012), Chambers u. a. (2009)]. Dazu werden im Vergleich zum Testsystem UKK andere Transkriptionsfaktoren hinzugefügt und das Medium öfter gewechselt. Die Einwirkperiode der toxischen Substanz ist auf 6 Tage begrenzt. Eine detaillierte Beschreibung der beiden Testsysteme ist in Krug u. a. (2013) gegeben.

Auf Grund der Tatsache, dass die Testsysteme UKK und UKN1 verschiedene biologische Aspekte der Differenzierung von Stammzellen zu Neuronen rekapitulieren, werden sie in dieser Arbeit separat analysiert.

2.2 Microarray Technologie

Die Microarray Technologie stellt ein Verfahren dar, welches zur Messung der Genexpression eingesetzt wird. Dabei unterscheidet man zwei grundsätzliche Ansätze: Ablagerung von DNA Fragmenten und *in-situ* Synthese. Bei der ersten Methode werden die DNA-Abschnitte beliebiger Länge auf eine Kunststoff-Unterlage aufgebracht. Die *in-situ* Herstellung lässt sich in Photolithographie-, Tintenstrahldruck- und elektrochemische Verfahren aufteilen. Die in dieser Arbeit verwendeten Daten wurden mittels Chips des Herstellers **Affymetrix** gewonnen, deshalb konzentrieren wir uns in diesem Abschnitt auf eine Darstellung des entsprechenden photolithographischen Verfahrens. Für eine systematische Übersicht der Microarray-Technologie sei der Leser auf das Buch von Drăghici (2010) verwiesen.

Das Verfahren basiert auf der hochspezifischen Bindung (Hybridisierung) zwischen Fänger- (in unserem Fall kurze Oligonukleotidstücke) und Zielmolekülen (in unserem Fall mRNA, deren Gesamtmenge Rückschlüsse auf das Expressionsniveau der Gene zuläßt). Die Bindung erfolgt dabei nach dem Schema der Watson-Crick-Basenpaarung. Die Fängermoleküle werden auf der Trägerfläche des Chips unter genauer Notation des Ortes photolithographisch erstellt. Dabei wird die Nukleotidsequenz komplementär zu dem Zielmolekül Base für Base einzeln aufgebaut. Um eine hohe Genauigkeit der Reihenfolge zu gewährleisten, beschränkt man die Oligonukleotidsequenz auf eine Länge von 25-30 Basen. Dieses Stück wird als **Probe** bezeichnet. Um die unspezifischen Hybridisierungen zu erkennen, wird ein sogenanntes **Probenpaar** synthetisiert. Es besteht aus einem *perfect match* und einem *mismatch*. Während das perfect match zu der zu bestimmenden Sequenz komplett komplementär ist, unterscheidet sich das mismatch von perfect match genau in einer Base in der Mitte des Strangs. Das Probenpaar sollte somit unspezifische Hybridisierungen erkennen und die Genauigkeit des Experiments erhöhen. Diese Vermutung hat sich in der Praxis [Irizarry u. a. (2003b)] allerdings nicht bestätigt, so dass die Signalintensitäten vom mismatch nicht verwendet werden. Für ein einzelnes Gen werden mehrere Probenpaare synthetisiert: So werden in modernen Affymetrix-Chips 11 Probenpaare zu einem so genannten **Probeset** zusammengefasst. Ein einzelnes Probeset korrespondiert zu einem bestimmten Gen oder Genabschnitt. Ein Gen kann durch mehrere Probesets repräsentiert werden. So korrespondieren Probesets mit den Kennzeichen 200801_x_at, 213867_x_at und 224594_x_at des Chips **hgu133plus2** zu dem Gen *Beta Cytoskeletal Actin*.

Zum Erstellen der Zielmoleküle wird zuerst die mRNA aus dem zu untersuchenden Gewebe extrahiert und aufgereinigt. Ferner wird die mRNA mittels reverser Transkriptase in cDNA umgeschrieben. Anhand dieser cDNA Vorlage werden cRNA-Moleküle hergestellt, fragmentiert und an den Träger hydridisiert. Das gebundene Material wird eingefärbt und ungebundene Moleküle abgewaschen. Ein Laser liest den Träger ab und ordnet jedem ein-

zelnen Probeset seine Intensität zu, sodass von dem gesamten Träger ein Bild (Image) entsteht. Ein typisches Image ist auf Abbildung 5 präsentiert.

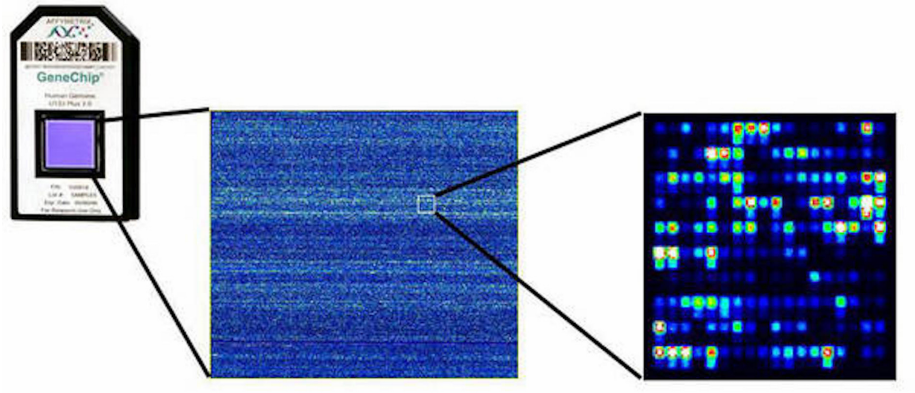

Abbildung 5: Affymetrix Chip (links), typisches Fluoreszenz-Image (mitte) und vergrößerter Abschnitt (rechts). Quelle der Abbildung ist `https://www.dkfz.de/gpcf/affymetrix_genechips.html` (Geprüft 29. September 2016).

Ein Bildverarbeitungssoftware konvertiert den Image in Messungen der Intensitätsstärke. Diese Messungen werden anschließend aufbereitet, um die Expressionsmatrix zu gewinnen. In dieser Arbeit wird die **Robust Multi-array Average (RMA)** Methode verwendet, welche gegenüber dem MAS5-Verfahren zu verbessertem Auffinden von differentiell exprimierten Probesets führt, siehe Irizarry u. a. (2003a). Im Rahmen dieser Vorverarbeitung werden nacheinander eine Hintergrundkorrektur, eine Quantilnormalisierung und eine Zusammenfassung durchgeführt. Eine detaillierte Einführung in die RMA-Methode findet sich in der Arbeit von Irizarry u. a. (2003b).

2.2.1 Gene Ontology

Um die Ansammlung von biologischem Wissen zu strukturieren und zu vereinheitlichen, werden Ontologien verwendet. Ontologien dienen zum formalen Zusammenfassen vom Erkenntnissen („Konzepten“) und der zwischen ihnen bestehenden Beziehungen. Die **Gene Ontology**(GO)[Ashburner u. a. (2006)] ist eine biomedizinische Ontologie und beschreibt das Wissen über biologische Prozesse, molekulare Funktion und die Lokalisierung der Genprodukten in drei unabhängigen Topologien: „Zelluläre Komponente“, „Molekulare Funktion“ und „Biologischer Prozess“. Jede Topologie ist ein gerichteter azyklischer Graph. Der Graph kann dabei bis zu 28 Tausend Knoten bestehen (Topologie „Biologischer Prozess“ enthält laut dem R-Paket `GO.db` [Carlson] 28007 Knoten). Jeder Knoten besteht aus mehreren Einträgen und einem Namen, einer eindeutigen alphanumerischen Identifizierung, einer Definition und einer Zuordnung zu einem von drei Topologien. Das Beispiel des GO-Eintrags „GO:0048863“ in Tabelle 1 dient der Veranschaulichung.

Der Einträge *id*, *name*, *namespace* und *def* enthalten entsprechend die Identifizierung, den Namen, den Bereich und die Definition des Knotens. Der Eintrag *xref* enthält Verwei-

id:	GO:0048863
name:	stem cell differentiation
namespace:	biological process
def:	„The process in which a relatively unspecialized cell acquires specialized features of a stem cell. A stem cell is a cell that retains the ability to divide and proliferate throughout life to provide progenitor cells that can differentiate into specialized cells“ [CL:000034, GOC: isa_complete]
xref:	Wikipedia: stem_cell_differentiation
is_a:	GO:0030154 ! cell differentiation

Tabelle 1: Beispiel eines GO-Eintrags

se auf das gleiche oder sehr ähnliche Objekte in anderen Datenbanken, wie z.B. Enzyme Commission (EC), Database of metabolyc pathways and enzymes (MetaCyc) oder Reactome. Der Eintrag *is_a* gibt den Vaterknoten an und verweist darauf, dass „stem cell differentiation“ ein Unterknoten von „cell differentiation“ ist. Die Gene Ontology unterscheidet neben *is_a* auch andere Beziehungen zwischen den Knoten: *part of*, *has part*, *regulates*, *negatively regulates* und *positively regulates*.

Während der Aufbau des GO-Graphen unabhängig vom jeweiligen Organismus erfolgt, werden die Gene bzw. Genprodukte Organismus-spezifisch einem GO-Knoten annotiert. Basierend auf experimentellen Befunden oder strukturellen Ähnlichkeiten werden einzelne Gene der jeweiligen GO-Gruppe zugeordnet. Dies stellt eine Möglichkeit dar, eine Liste von Genprodukten auf gemeinsame Eigenschaften hin zu untersuchen. Diese gemeinsamen Eigenschaften sind durch diejenigen GO-Gruppen repräsentiert, welche mit dem größeren Teil der Liste von Genprodukten assoziiert sind. Ferner lassen sich zwei Listen vergleichen, z.B. die Liste der differentiell deregulierten Genprodukte und die Liste aller bekannten Gene. Sind bestimmte GO-Gruppen in der ersten Liste überrepräsentiert, könnte dies ein Hinweis auf die zu Grunde liegenden biologischen Prozesse sein. Man würde z.B. erwarten, dass beim Vergleich einer sich differenzierenden Stammzelle mit einer Muskelzelle diejenigen Gene größere Aktivität aufweisen, welche mit dem GO-Knoten „stem cell differentiation“ assoziiert sind. In der Praxis würde man die Expression der Gene bzw. Genprodukte auf Unterschiede zwischen zwei Populationen mit Hilfe von moderiertem t-Test testen (zur Definition und Erläuterung dieses Tests siehe das Kapitel 3.3). Diejenigen Genprodukte, deren Test auf Gleichheit der Expression abgelehnt wird, werden als dereguliert bezeichnet. Ferner wird für jede GO-Gruppe eine Vierfeldertafel erstellt und die beiden Merkmale „Genprodukt dereguliert“ und „Genprodukt enthalten in der GO-Gruppe“ mittels des exakten Tests nach Fisher auf Unabhängigkeit getestet. In Rahmen dieser Arbeit wurde das Paket `topGO` [Alexa u. Rahnenführer (2010)] verwendet. Die entsprechenden **R**-Skripte sind im Anhang auf Seite 230 angegeben.

2.2.2 Biologie der Transkriptionsfaktoren und Analyse der Bindungsstellen

Eine Messung der Genexpression stellt eine Momentaufnahme dar. Eine Möglichkeit die dynamische Regulation der Transkription zu entschlüsseln, stellt die Analyse der Bindungsstellen von Transkriptionsfaktoren (englisch: *transcription factor binding site* **TFBS**) dar. Transkriptionsfaktoren sind regulatorische Proteine, die bestimmte Sequenzabschnitte (Bindungsstellen) der DNA erkennen und durch Bindung an diese Abschnitte die Transkription bestimmter Gene beeinflussen. Die Bindungsstellen sind generell vor dem RNA-codierenden Bereich zu finden. Dieser Abschnitt wird auch als Promotor bezeichnet. Die Expression eines bestimmten Gens kann somit die Expression anderer Gene positiv oder negativ regulieren. Weisen mehrere Genprodukte in ihrem Promotor die gleiche Nukleotidensequenz auf, kann man vermuten, dass deren Regulation über denselben Transkriptionsfaktor erfolgt. In der Praxis greift man auf die Datenbanken zurück, welche sowohl Transkriptionsfaktoren als auch Bindungsstellen enthalten, wie z.B. TRANSFAC [Wingender (2008)] oder JASPAR [Bryne u. a. (2008)]. Die Liste der deregulierten Genprodukte wird für jeden Transkriptionsfaktor auf das Vorhandensein von Bindestellen in den Promotorbereich untersucht. Enthalten überproportional viele Genprodukte eine gemeinsame Bindungstelle, so liefert dies einen Hinweis darauf, dass ein gemeinsamer Transkriptionsfaktor ihre Regulation bestimmt. Aus biologischer Sicht könnte man somit vermuten, dass diese Genprodukte Prozesse regulieren, die eng miteinander verbunden sind.

In Rahmen der vorgestellten Pipeline wird die Analyse der deregulierten Genprodukte auf gemeinsame Transkriptionsfaktoren mit Hilfe der webbasierten Anwendung **oPOSSUM** [Sui u. a. (2007)] durchgeführt. Die Version 3 der Anwendung ist unter dem Link `http://opossum.cisreg.ca/oPOSSUM3/` zugänglich.

2.3 Verwendete Datensätze

In diesem Abschnitt stellen wir die Datensätze vor, welche im Rahmen dieser Arbeit analysiert werden. Sie wurden unter Verwendung derselben Affymetrix Chips **hgu133plus2** in Köln bzw. Konstanz für verschiedene Aufgaben erstellt. Eine Übersicht der Datensätze ist in Tabelle 2 angegeben.

Im Anhang ab Seite 99 sind detaillierte Übersichten zu den verwendeten Daten zu finden.

2.3.1 VPA Chronic Konzentrationsstudie

In dieser Studie setzte man sich zum Ziel, die Reaktion von Stammzellen auf Valproinsäure (**VPA**) zu untersuchen. Die Substanz wurde dabei für eine Zeitperiode von 6 Tagen in 8 verschiedenen Konzentrationen zu den Zellen hinzugegeben. 6 Proben wurden ausschließlich dem Lösungsmittel ausgesetzt und bildeten deshalb die Kontrolle. Jeweils

Bezeichnung	Testsystem	Anzahl der Proben
VPA Chronic Konzentrationsstudie	UKN1	30
VPA Acute Konzentrationsstudie	UKN1	18
MeHg Chronic Konzentrationsstudie	UKN1	24
MeHg Acute Konzentrationsstudie	UKN1	17
VPA Zeitfensterstudie	UKK	57
Klassifikationsstudie UKN1	UKN1	85
Klassifikationsstudie UKK	UKK	100

Tabelle 2: Übersicht der Datensätze

drei Proben wurden mit der entsprechenden Konzentration der Valproinsäure behandelt. Insgesamt enthält die Studie 30 Proben, siehe Tabelle 3 und auch Übersicht 19 im Anhang.

Konzentration in µM	0	25	150	350	450	550	650	800	1000
Anzahl der Proben	6	3	3	3	3	3	3	3	3

Tabelle 3: Übersicht über die VPA Chronic Konzentrationsstudie

2.3.2 VPA Acute Konzentrationsstudie

Ähnlich zur VPA Konzentrationsstudie 2.3.1 wurde in dieser Versuchssreihe die Reaktion von Stammzellen auf Valproinsäure quantifiziert. Anders als bei der chronischen Untersuchung wurde die Einwirkperiode auf 2 Tage reduziert und die Konzentrationen wurden erhöht. Insgesamt wurden 18 Experimente in drei Abfertigunsschüben durchgeführt (siehe Tabelle 4 und eine vollständige Übersichtstabelle 20 im Anhang).

Konzentration in mM	0	0.6	1	2	4	5
Anzahl der Proben	3	3	3	3	3	3

Tabelle 4: Übersicht über die VPA Acute Konzentrationsstudie

2.3.3 MeHg Chronic Konzentrationsstudie

In dieser Studie untersuchte man analog zur VPA Chronic Konzentrationsstudie (siehe Kapitel 2.3.1) die Reaktion von Stammzellen auf die Wirkung von Methylquecksilber (**MeHg**). Die Einwirkzeit betrug ebenfalls 6 Tage. Zu jeder experimentellen Kondition (unbehandelt, Einwirkung von Lösungsmittel Ethanol (EtOH), Inkubation mit 6 verschiedenen Konzentrationen von MeHg) wurden jeweils 3 Replikate erstellt. Somit beinhaltet diese Studie insgesamt 24 Beobachtungen, siehe Tabelle 5 und eine vollständige Übersicht im Anhang in Tabelle 21.

Konzentration in µM	0	EtOH	0.25	1.4	1.6	1.8	2	3
Anzahl der Proben	3	3	3	3	3	3	3	3

Tabelle 5: Übersicht über die MeHg Chronic Konzentrationsstudie

2.3.4 MeHg Acute Konzentrationsstudie

Analog zur VPA Acute Konzentrationsstudie (Abschnitt 2.3.2) wurde in dieser Studie die Reaktion von Stammzellen auf die im Vergleich zu 2.3.3 erhöhte Konzentration von Methylquecksilber. Die Einwirkzeit wurde auf 2 Tage reduziert. In drei Fertigungslosen wurden die Proben mit 5 verschiedenen Konzentrationen inkubiert. Als Kontrolle wurden 3 Replikate von unbehandelten Beobachtungen gemessen. Da eine mit der maximalen Dosis (40 µM MeHg) versetzte Probe unbrauchbar geworden war, wurde sie aus der Studie entfernt. Somit wurden insgesamt 17 Beobachtungen analysiert, siehe Tabelle 6. Eine vollständige Übersichtstabelle 22 befindet sich im Anhang.

Konzentration in µM	0	1.5	10	15	20	40
Anzahl der Proben	3	3	3	3	3	2

Tabelle 6: Übersicht über die MeHg Acute Konzentrationsstudie

2.3.5 VPA Zeitfensterstudie

Zusätzlich zu der Untersuchung des Einflusses von verschiedenen Konzentrationen auf die Zellentwicklung ist ferner die Frage interessant, ob die Einwirkperiode eine Rolle spielt. Für die Beantwortung dieser Frage wurde in der Universität zu Köln eine Zeitfensterstudie durchgeführt. Dabei wurden die Zellen der Wirkung von Valproinsäure ausgesetzt, wobei die Einwirkperiode 2, 3, 5, 6, 8, 9, 12 und schließlich 14 Tage betrug. Sowohl der Beginn bzw. das Ende der Einwirkung als auch der Tag der Analyse wurden variiert. In Tabelle 7 sind die Einzelheiten über die behandelten Stammzellen zu finden.

Begin der Einwirkperiode	0	0	0	0	0	0	0	0	0	6	9	12
Ende der Einwirkperiode	3	6	9	12	3	6	9	12	14	14	14	14
Tag der Analyse	3	6	9	12	14	14	14	14	14	14	14	14
Anzahl der Proben	3	3	3	3	3	3	3	3	3	3	3	3

Tabelle 7: Übersicht über behandelte Zellen in der VPA Zeitfensterstudie

Parallel wurden Zellen für die entsprechenden Zeitfenster dem Lösungsmittel ausgesetzt. Sie bildeten somit die Kontrollen.

2.3.6 Klassifikationsstudie UKN1

Die Frage nach dem genetischen Fingerabdruck bei bestimmten Substanzklassen bildet eine der wichtigsten Herausforderungen in der Toxikologie. Zu dieser Fragestellung wurde

eine Studie durchgeführt, bei der die Stammzellen mit Stoffen behandelt wurden, die in ihrer Wirkung entweder Methylquecksilber oder Valproinsäure ähnlich sind. Insgesamt wurden 12 Substanzen in ihrer Wirkung untersucht: 6 aus der Klasse von quecksilberhaltigen Substanzen (**Hg+**) und 6 aus der Klasse der Histondeacetylase-Inhibitoren (**HDACi**). Ursprüngliche Intention der Studie war die Einteilung von 6 ursprünglich verblindeten Substanzen (auch als Testsubstanzen bezeichnet) auf der Grundlage eines Klassifikators, zu dessen Konstruktion die Expressionsmessungen von 6 unverblindeten Substanzen (Trainingssubstanzen) herangezogen wurden. Dieses Vorhaben wurde allerdings fallengelassen, da die quecksilberhaltigen Substanzen in ihrer Wirkung auf die Neuronenentwicklung inhomogen waren. Eine mögliche Ursache dafür sind die unterschiedlichen Zeiten der Analysedurchführung: Auf Grund von technischen Restriktionen war es unmöglich, alle Proben zusammen zu erstellen und zu analysieren. Man hat sie in 4 verschiedenen Gruppen, auch Batches genannt, gemessen. Dies führte zu einer Vergrößerung von nicht-biologischer Varianz - dem Batch-Effekt. Zu einer ausführlichen Definition und Analyse davon wird an dieser Stelle auf ein späteres Kapitel 3.4 verwiesen. Eine kurze Übersicht über die Proben findet man in Tabelle 8, während eine ausführlichere Darstellung sich im Anhang in Tabelle 23 befindet.

2.3.7 Klassifikationsstudie UKK

Analog zur in Unterkapitel 2.3.6 vorgestellten Klassifikationsstudie UKN1 wurde in dieser Studie die Reaktion von eHSC des Testsystems UKK auf die quecksilberhaltigen Substanzen bzw. HDAC-Inhibitoren untersucht. Dazu hat man die gleichen zwölf Substanzen genommen, welche in der UKN1 Klassifikationsstudie verwendet wurden (vergleiche Tabelle 8). Im Unterschied zur UKN1 Studie wurden allerdings die Substanzen MeHg und VPA in zwei verschiedenen Konzentrationen verwendet. In Tabelle 8 werden sie zusammengefasst, in den späteren Analysen werden jedoch die jeweiligen Konzentrationen separat behandelt. Einen weiteren Unterschied zur UKN1 Studie stellt die Hinzunahme der Proben von nichtdifferenzierten Stammzellen (ESC) dar. Sie wurden nicht dem Lösungsmittel ausgesetzt und frisch präpariert. Ihre Expression unterscheidet sich sehr stark von anderen Zellen, deshalb werden sie von den späteren Analysen entfernt. Eine ausführliche Übersicht ist im Anhang in Tabelle 26 zu finden. Auch in dieser Studie war es unmöglich, alle Proben gleichzeitig zu hydridisieren. So wurden die RNA-Proben in 4 zeitlich getrennten Schüben (Batches) hydridisiert.

2.4 Ziele der Arbeit

Es wird zum Ziel gesetzt, ein Verfahren bereitzustellen, welches Biologen ermöglicht, eine sequentielle Analyse von Microarray-Daten durchzuführen, welche im Rahmen einer Studie zur Toxizitätsbestimmung erhoben wurden. Als Ausgangslage dienen zwei bestimmte

Arten von Studien: Konzentrations- und Klassifikationsstudien. Die biologische Fragestellung bei Konzentrationsstudien lautet: Wie hängt der Einfluss der zu untersuchenden Substanz von der Konzentration ab? Hierzu wird eine Abfolge von Methoden erstellt, welche der deskriptiven Analyse, einer eventuellen (Batch-)Korrektur und der Bestimmung von differentiellen Genen dienen. Während deskriptiven Analysen zur Veranschaulichung der Daten herangezogen werden und Hinweise zu Quellen nichtbiologischer Varianz liefern können, kann man mit Methoden zur Batch-Korrektur deren Effekte vermindern. Die Frage nach den durch die Hinzugabe der toxischen Substanz hervorgerufenen zellulären Prozessen wird mit Hilfe des moderierten t-Tests (vorgestellt im Unterkapitel 3.3) adressiert. Dazu wird jedes Gen auf Unterschiede zwischen den Konditionen untersucht. Die signifikant deregulierten Gene geben einen Hinweis auf die zugrundeliegenden Veränderungen des Transkriptoms und folglich der Zelle selbst.

Im Falle von Klassifikationsstudien liegen Expressionsmessungen von zwei oder mehr Gruppen vor, welche in sich homogen sind. Der Fokus der Untersuchung liegt in der Bestimmung von Unterschieden auf der Ebene des Transkriptoms. Dazu dienen die Methoden der Klassifikation. Im Rahmen dieser Arbeit konzentriert man sich auf zwei Verfahren, die sich großer Beliebtheit erfreuen: Support Vector Machines (3.5.1) und Random Forests (3.5.2). Zu jeder dieser Methoden werden in Rahmen dieser Arbeit folgende Fragestellungen adressiert:

- Wie wählt man die Prädiktoren?
- Wie übertragbar (generalisierbar) sind die Ergebnisse der Klassifikation?
- Wie hängt die Klassifikationsgüte von der Anzahl der verwendeter Replikate ab?
- Ist die Klassifikationsgüte empfindlich einem zusätzlichen Verrauschen der Daten gegenüber?

Substanz	Klasse	# Replikate in UKN1	Batch in UKN1	# Replikate in UKK	Batch in UKK
Belinostat	HDACi	4	IV	8	IV
Dimethylsulfoxid (**DMSO**)	Kontrolle	9	I,II	4	II
Entinostat	HDACi	4	IV	7	IV
Ethanol (**EtOH**)	Kontrolle	5	I	-	-
Quecksilber(II)-bromid (**$HgBr_2$**)	Hg+	4	IV	4	III
Quecksilber(II)-chlorid (**$HgCl_2$**)	Hg+	4	II	4	II
Methylquecksilber	Hg+	5	I	10	I
Panobinostat	HDACi	4	IV	4	III
4-Chlormercuribenzoesäure (**PCMB**)	Hg+	4	IV	4	III
Phenylquecksilberacetat (**PMA**)	Hg+	4	IV	4	III
suberoylanilide hydroxamische Säure (**SAHA**)	HDACi	4	II	4	II
Thimerosal	Hg+	4	II	4	II
Trichostatin A (**TSA**)	HDACi	5	I	4	II
Unbehandelt	Kontrolle	21	I,II,III,IV	25	I, II, III, IV
Valproinsäure	HDACi	4	III	10	I

Tabelle 8: Übersicht über die Proben der Klassifikationsstudien UKN1 und UKK. Die Spalte **Substanz** gibt den vollständigen Namen und ggfs. die Abkürzung der hinzugegebenen Substanz an. Die Spalte **Klasse** zeigt an, ob die betroffene Substanz quecksilberhaltig (Hg+) ist, als Histondeacetylase-Inhibitor (HDACi) fungiert oder eine Kontrolle bildet.

3 Statistische Methoden

Dieses Kapitel dient der Vorstellung der statistischen Methoden, welche in der vorliegenden Arbeit angewandt wurden. Dabei handelt es sich sowohl um die deskriptive Analysen, wie die Hauptkomponentenanalyse (siehe 3.1) und die Heatmap (3.2), als auch um Verfahren der schließenden Statistik, wie der moderierte t-Test (3.3) und die Analyse des Batch-Effektes (3.4). Eine Einführung in die Theorie der Klassifikationsmethoden (3.5), insbesondere die Verfahren Support-Vector-Machine (**SVM**) und Random Forests (**RF**), rundet das Kapitel ab.

3.1 Hauptkomponentenanalyse

Die *Hauptkomponentenanalyse* (**HKA**) ist ein multivariates Verfahren, welches eine Menge von Variablen in eine (evtl. kleinere) Menge unkorrelierter Variablen transformiert. Die transformierten Variablen werden als *Hauptkomponenten* bezeichnet und werden absteigend nach dem Anteil der erklärten Varianz sortiert. Aus mathematischer Sicht stellt das Verfahren eine lineare orthogonale Koordinatentransformation dar, wobei die Eigenvektoren der entsprechenden Kovarianzmatrix als die Basisvektoren fungieren.

Seien die Datenpunkte $x_1, x_2, ..., x_n \in \mathbb{R}^p$ in Form einer $(p \times n)$ Matrix X mit Rang $r \leq n$ gegeben. Als Erstes werden die Messungen zentriert, indem ein p-diminsionaler Mittelwert von jeder Spalte abgezogen wird. Die Kovarianzmatrix Z ist dann durch

$$Z = \frac{1}{n} X\, X^\top$$

gegeben. Die Matrix Z ist symmetrisch und positiv semidefinit, somit lässt sie sich wie folgt zerlegen

$$Z = U \Sigma U^\top, \tag{3.1}$$

wobei die Diagonalmatrix Σ die nichtnegativen Eigenwerte und die orthonormale Matrix U die Eigenvektoren von Z enthält. Die Matrix U wird in der Literatur auch als Rotations- oder Ladungsmatrix bezeichnet. Dabei sind die Eigenwerte $\sigma_1, \sigma_1, ..., \sigma_r$ absteigend sortiert. Die ersten beiden Eigenvektoren u_1 und u_2 eignen sich nun am besten, um die Datenpunkte $x_1, x_2, ..., x_n$ in $\mathbb{R}^2$ graphisch darzustellen.

In Abbildung 6 (siehe auch Abbildung 1C in Waldmann u. a. (2014)) sind die ersten beiden Hauptkomponenten der transformierten Daten der in Unterkapitel 2.3.1 vorgestellten VPA Konzentrationsstudie dargestellt (erstellt mit Sprache **R**, Code im Anhang auf Seite 230). Die erste Hauptkomponente erfasst bereits knapp 89 % der gesamten Varianz, die zweite 7,16 %. Somit wird auf dem Hauptkomponentenplot unter Verwendung von zwei Hauptkomponenten insgesamt über 95 % der Varianz wiedergegeben. Die Abbildung reproduziert die Daten sehr gut, denn auf die restlichen Hauptkomponenten entfallen lediglich weniger als 5 % der Gesamtvarianz. Beachtenswert ist hierbei vor allem die erste

Hauptkomponente: Bis auf eine Ausnahme (Behandlung mit 650 µM) geht die Steigerung der Konzentration mit der Verschiebung der Punkte nach rechts einher.

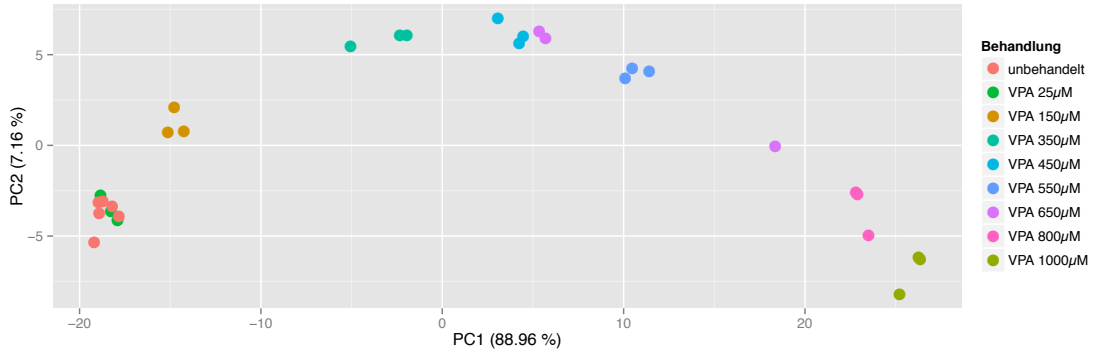

Abbildung 6: Darstellung der ersten beiden Hauptkomponenten der HKA der VPA chronischen Konzentrationsstudie. Der Anteil der erfassten Varianz ist in Prozent angegeben. Die verschiedenen Behandlungskonzentrationen sind durch Farbe gekennzeichnet. Bis auf eine Ausnahme (Behandlung mit 650 µM) weisen die mit höheren Konzentrationen behandelte Proben einen größeren Wert auf der ersten Hauptkomponente auf.

Diese Korrelation lässt folgende biologische Interpretation zu: Geht man von einem additiven Behandlungseffekt aus, ruft die steigende Dosis eine stärkere Reaktion hervor. Eine schrittweise Erhöhung der Konzentration löst eine kontinuierliche Antwort auf der Ebene der transkribierten Gene aus. Diese kumulative Reaktion ist verantwortlich für den größten Anteil der Varianz und lässt sich an der ersten Hauptkomponente auf Abbildung 6 beobachten.

Im Rahmen der Untersuchung war es nicht möglich, die exakten Gründe für das abweichende Verhalten der Behandlung mit 650 µM zu bestimmen. Man ging davon aus, dass hier ein experimentelles Artefakt vorlag und nahm die entsprechenden Proben aus der Untersuchung. Die Hauptkomponentenanalyse eignet sich hiermit auch zur Datenkontrolle und Feststellung eines Batch-Effektes (siehe dazu auch das Unterkapitel 3.4).

3.2 Heatmap

Ein weiteres multivariates Verfahren stellt das *Clusterverfahren* dar. Dabei werden die zu untersuchenden Objekte auf Grund eines Ähnlichkeitsmaßes hierarchisch zu Gruppen zusammengefasst, die als *Cluster* bezeichnet werden. Eine in der Molekularbiologie weit verbreitete Darstellung der Ergebnisse eines hierarchischen Clusterverfahrens stellt die *Heatmap* dar: Von ihrer Einführung in Weinstein u. a. (1997) im Jahr 1997 bis zum Jahr 2008 wurde sie in mehr als 4000 Publikationen verwendet [Weinstein (2008)]. Die in einer zweidimensionaler Matrix gegebenen Daten, z.B. Expressionsmessungen von mehreren Genen in verschiedenen Proben, werden in einem Rechteck farblich präsentiert. Die Farbe

gibt wieder, wie stark ein bestimmtes Genprodukt exprimiert ist. Die Proben und die Probesets werden dabei mit Hilfe des Clusterverfahrens hierarchisch gruppiert, und die Ergebnisse in Form des Dendrogramms präsentiert. Beispielhaft ist die auf Abbildung 7 (siehe auch Abbildung 1D in Waldmann u. a. (2014)) abgebildete Heatmap der VPA Konzentrationsstudie. Für diese Darstellung werden alle 30 Proben (in den Spalten) und 1000 Probesets (in den Zeilen) mit der höchsten Varianz über alle Proben verwendet. Oberhalb des farbigen Rechteckes ist das Dendrogramm angebracht, welches die hierarchische Einteilung der Proben in die Gruppen wiedergibt. Die farbige Leiste unterhalb des Dendrogramms symbolisiert verschiedene Konzentrationen. Die Zuweisung der Konzentrationen zu den Farben ist in der Legende unten links angegeben. Auf die Gruppierung der Probesets wurde für diese Darstellung verzichtet. Zu der ersten Interpretation dieser Heatmap lässt sich sagen, dass übereinstimmend mit den Ergebnissen der Hauptkomponentenanalyse die Abstände zwischen den Proben kleiner sind, je näher die entsprechenden Konzentrationen liegen. So bilden die niedrigen Konzentrationen (0, 25 und 150 µM), die mittleren (350, 450, 550 und 650 µM) und die hohen (800 und 1000 µM) jeweils ein kompaktes Cluster. Eine Ausnahme stellt allerdings die Konzentration 650 µM dar: Während zwei Proben sich zu den mittleren Konzentration gruppieren, bildet das dritte Replikat ein Cluster mit den hohen Konzentrationen ausgesetzten Proben. Somit liefert diese Analyse einen weiteren Hinweis darauf, dass die mit 650 µM behandelten Proben ein heterogenes Expressionsprofil aufweisen.

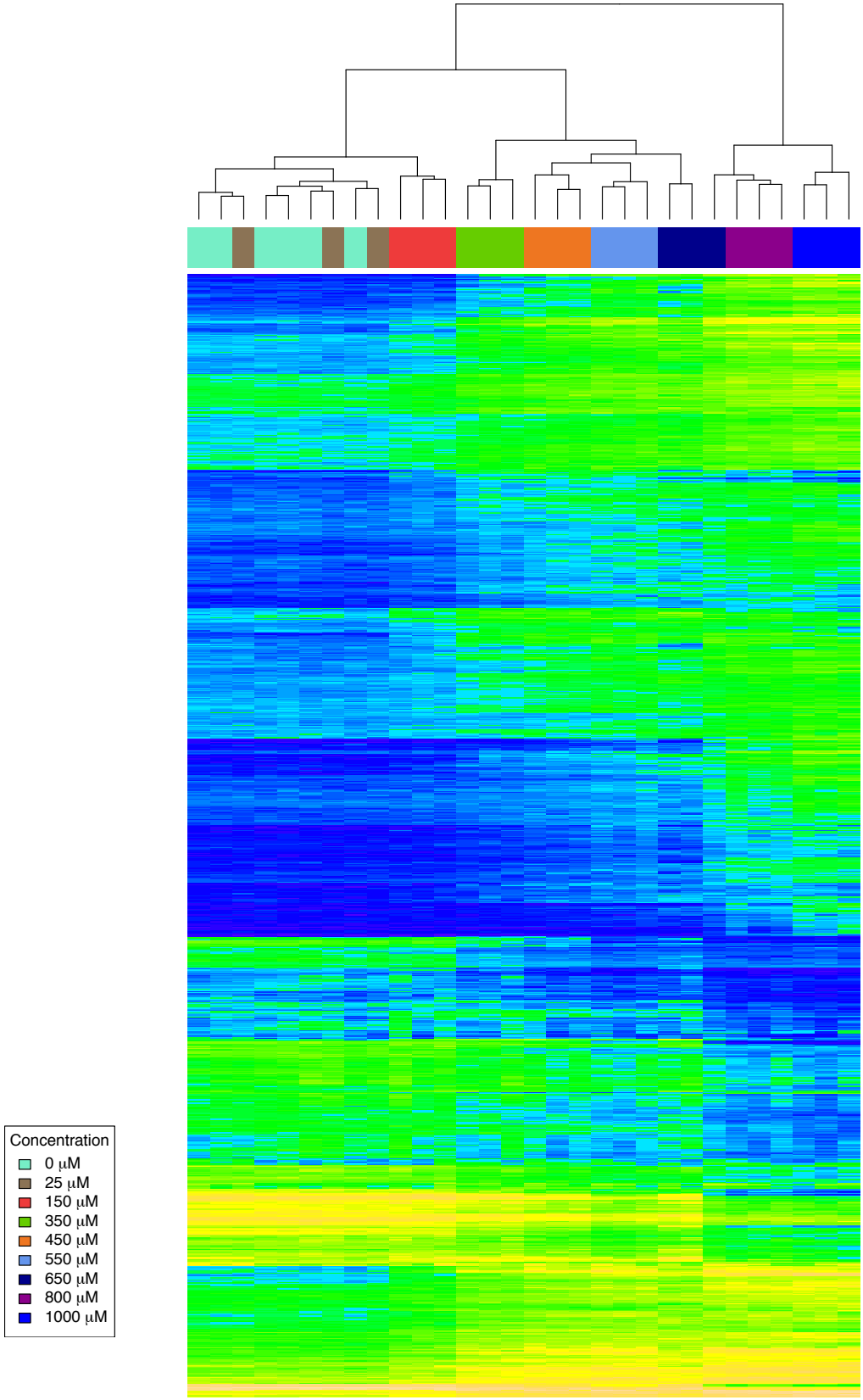

Abbildung 7: Heatmap der VPA Konzentrationsstudie. Das Dendrogramm am oberen Rand der Abbildung spiegelt die Entfernungen der Beobachtungen zueinander schematisch wieder. Die farbige Leiste unterhalb des Dendrogramms teilt die Proben je nach verwendeter Konzentration ein.

3.3 Moderierter *t*-Test

Das Bestimmen von differentiell exprimierten Genen/Probesets ist eine der Hauptanwendungen der Genexpressionsmessungen. Es ist zuerst naheliegend, den **Fold-Change** zu benutzen, d.h. die normalisierte Expressionsrate des Probesets in der Testprobe als Vielfaches der normalisierten Expressionsrate in der Kontrollprobe. Das Vorgehen, bei welchem die Probesets mit Fold-Change größer als 2 oder kleiner als 0.5 als dereguliert bezeichnet werden, entbehrt allerdings einer statistischen Begründung und ist solchen Methoden wie dem Zweistichproben-t-Test unterlegen, siehe Witten u. Tibshirani (2007). Der Zweistichproben-t-Test stellt wiederum ein bekanntes Verfahren dar, welches in der Praxis und Forschung oft angewendet wird. Dabei fasst man die Expressionswerte eines Gens in zwei verschiedenen Konditionen als zwei unabhängige Messreihen auf. Die Expressionmessungen $y_{j11}, y_{j12}, ..., y_{j1n}$ des Gens j in der ersten Messreihe werden als Beobachtungen aus einer $N(\mu_{j1}, \sigma^2_{j1})$-verteilten Grundgesamtheit und die Beobachtungen $y_{j21}, y_{j22}, ..., y_{j2m}$ der zweiten Messreihe als die aus einer $N(\mu_{j2}, \sigma^2_{j2})$-verteilten Grundgesamtheit stammenden Größen betrachtet. Es wird angenommen, dass die Varianzen beider Messreihen für jedes Gen j gleich sind, d.h

$$\sigma_{j1} = \sigma_{j2} = \sigma_j \quad \forall j \tag{3.2}$$

Für verschiedene Gene können Varianzen jedoch unterschiedlich sein. Wir möchten jetzt für Gen j testen, ob die Erwartungswerte in den beiden Messreihen gleich sind. Somit lauten unsere Null- bzw. Alternativhypothese:

$$H_{0j} : \mu_{j1} = \mu_{j2} \quad \text{gegen} \quad H_{1j} : \mu_{j1} \neq \mu_{j2}.$$

Für den Test konstruieren wir die Statistik

$$t_j = \frac{\bar{y}_{j1} - \bar{y}_{j2}}{S_j \sqrt{\frac{1}{n} + \frac{1}{m}}}, \tag{3.3}$$

wobei mit S_j^2 die Stichprobenvarianz

$$S_j^2 = \frac{\sum\limits_{i=1}^{n} y_{j1i}^2 - \frac{1}{n}\left(\sum\limits_{i=1}^{n} y_{j1i}\right)^2 + \sum\limits_{i=1}^{m} y_{j2i}^2 - \frac{1}{m}\left(\sum\limits_{i=1}^{m} y_{j2i}\right)^2}{n + m - 2}$$

bezeichnet wird. Gilt die Nullhypothese, so ist die Statistik (3.3) t_{n+m-2}-verteilt [Hartung u. a. (2005)].

Trotz der weiten Verbreitung des t-Tests gibt es wichtige Einschränkungen, die dem Anwender bewußt sein sollten. Erstens könnte die Stichprobenstandardabweichung S_j auf Grund einer kleinen Varianz verzerrt sein, siehe z.B. Tusher u. a. (2001). Dies resultiert in einem großen Wert der Teststatistik und demzufolge einer fälschlichen Ablehnung der

Nullhypothese. Ein weiterer Punkt, den man berücksichtigen sollte, ist die Anwendung im Falle von kleinen Stichprobenumfängen [Murie u. a. (2009)]. Eine Modifizierung des t-Tests, welche diese Probleme adressiert, stellt der **moderierte t-Test (Limma-Test)** dar [Smyth u. a. (2004)]. Ein Vergleich [Jeanmougin u. a. (2010)] mit Welch-Test, Wilcoxon-Vorzeichen-Rang-Test, ANOVA, VarMixt [Delmar u. a. (2005)] und anderen Methoden zeigte die Überlegenheit des Limma-Tests hinsichtlich des Umganges mit der kleinen Stichprobengröße. Die Verfasser empfehlen die Verwendung des R Pakets `limma` auch auf Grund der praktischen Benutzerfreundlichkeit. Auch ein neuerer Vergleich in Bandyopadhyay u. a. (2014) bestätigt folgende Vorteile des moderierten t-Tests: besserer Umgang mit den kleinen Stichproben, Verwendung von empirischem Bayes für die Varianzenkorrektur, usw. Als einzigen Minuspunkt sehen die Autoren die Annahme (3.2) des Tests, d.h Varianz des jeweiligen Gens sei für beide Messreihen gleich.

In weiteren Abschnitten wird nun der moderierte t-Test vorgestellt und erläutert. Die grundsätzliche Idee des Tests besteht darin, dass man annimmt, die unbekannten Varianzen σ_j^2 für $j = 1, \ldots, p$ folgen a priori einer inversen Gamma-Verteilung (3.9), welche von zwei Parametern (σ_0 und d_0) abhängt. Ferner gilt die Annahme, dass gegeben σ_j^2 die Stichprobenvarianzen S_j^2 - die man bestimmen kann - einer bedingten Verteilung, namentlich, einer skalierten χ^2 Verteilung, folgen (3.4). Bestimmt man den a posteriori Erwartungswert von σ_j^2 in (3.13), so hängt dieser nur von σ_0 und d_0 ab. Die beiden Parameter lassen sich bestimmen, indem man die ersten beiden Momente der logarithmierten Stichprobenvarianzen mit den theoretischen Größen vergleicht (siehe Unterkapitel 3.3.3). Nun lässt sich der a posteriori Erwartungswert von σ_j^2 als ein adjustierter Schätzer von S_j^2 interpretieren. Analog dem Ausdruck (3.3) wird eine Statistik (3.21) berechnet, welcher unter der Nullhypothese einer t-Verteilung folgt.

3.3.1 Hierarchisches Modell

Auf Grund der großen Anzahl von linearen Modellen, welche an die Daten der Microarray-Versuche angepasst werden, erscheint es sinnvoll, eine parallele Struktur zu entwickeln, sodass das gleiche Modell für jedes Gen j angepasst wird. Dies wird durch ein hierarchisches bayesianisches Modell implementiert. In diesem Unterkapitel wird der Leser mit den Grundlagen vertraut gemacht. Für eine ausführlichere Einführung in die Theorie der hierarchischen Modelle sei auf Gelman u. a. (2003) verwiesen.

Zuerst nehmen wir an, dass für die Likelihood der Stichprobenvarianz S_j^2 der Expression des j-ten Gens gegeben der Varianz σ_j^2 gilt:

$$S_j^2 \mid \sigma_j^2 \sim \frac{\sigma_j^2}{d_j} \chi_{d_j}^2. \tag{3.4}$$

Hierbei ist d_j die Anzahl der residualen Freiheitsgrade, die sich als $n + m - 2$ berechnen lässt. Die Varianzen σ_j^2 folgen nun ihrerseits a priori einer inversen χ^2-Verteilung

$$\frac{1}{\sigma_j^2} \sim \frac{1}{\sigma_0^2 d_0} \chi^2_{d_0}. \tag{3.5}$$

Die Verteilungen der Hyperparameter σ_0^2 und d_0 werden nun nicht wie üblich in der Bayes-Analysis vom Anwender a priori festgelegt, sondern aus den Daten geschätzt. Dies ist die bestimmende Eigenschaft vom empirischen Bayes-Analyse.

3.3.2 A posteriori Verteilung von σ_j^2

In diesem Unterkapitel wird die a posteriori Verteilung von σ_j^2 bestimmt. Nach dem Satz von Bayes gilt

$$p\Big(\frac{1}{\sigma_j^2}\Big|\ d_0, \sigma_0^2, S_j^2\Big) = \frac{p(\,S_j^2 \mid \frac{1}{\sigma_j^2}, d_0, \sigma_0^2)\ p(\frac{1}{\sigma_j^2} \mid d_0, \sigma_0^2)}{p(S_j^2 \mid d_0, \sigma_0^2)}. \tag{3.6}$$

Als Erstes stellen wir fest, dass gegeben σ_j^2 die Stichprobenvarianz S_j^2 von den Hyperparametern unabhängig ist

$$p(\,S_j^2 \mid \frac{1}{\sigma_j^2}, d_0, \sigma_0^2) = p(\,S_j^2 \mid \frac{1}{\sigma_j^2}).$$

Für die Vereinfachung der Gleichung (3.6) empfiehlt es sich ferner eine andere Darstellung der Verteilungen im Zähler zu benutzen. Dazu wird folgende Beziehung verwendet:

$$X \sim \chi^2_\nu \text{ und } c \in \mathbb{R}, c > 0 \Rightarrow cX \sim \Gamma(k = \frac{\nu}{2}, \Theta = 2c), \tag{3.7}$$

wobei mit Γ die Gamma-Verteilung bezeichnet wird. Für den Nachweis wird der Leser auf den Anhang (Seite 103) verwiesen. Mit dieser Transformation lassen sich die Beziehungen (3.4) und (3.5) wie folgt präsentieren:

$$S_j^2 \mid \sigma_j^2 \sim \frac{\sigma_j^2}{d_j}\chi^2_{d_j} \Rightarrow \frac{d_j S_j^2}{\sigma_j^2}\Big|\sigma_j^2 \sim \chi^2_{d_j} \Rightarrow S_j^2 \mid \frac{1}{\sigma_j^2} \sim \Gamma\left(k = \frac{d_j}{2}, \Theta = \frac{2\sigma_j^2}{d_j}\right) \tag{3.8}$$

$$\frac{1}{\sigma_j^2} \sim \frac{1}{\sigma_0^2 d_0}\chi^2_{d_0} \Rightarrow \frac{d_0\sigma_0^2}{\sigma_j^2} \sim \chi^2_{d_0} \Rightarrow \frac{1}{\sigma_j^2} \sim \Gamma\left(k = \frac{d_0}{2}, \Theta = \frac{2}{d_0\sigma_0^2}\right). \tag{3.9}$$

Setzt man die Dichte der Γ-Verteilung in den Zähler von (3.6) ein, erhält man folgende Beziehung:

$$\begin{aligned} p(S_j^2 \,\big|\, \frac{1}{\sigma_j^2})\; p(\frac{1}{\sigma_j^2}\big|\; d_0, \sigma_0^2) &= \frac{1}{\Gamma(\frac{d_j}{2})\big(\frac{2\sigma_j^2}{d_j}\big)^{\frac{d_j}{2}}} S_j^{2(\frac{d_j}{2}-1)} \exp\big(\frac{-S_j^2 d_j}{2\sigma_j^2}\big) \\ &\quad \frac{1}{\Gamma(\frac{d_0}{2})\big(\frac{2}{d_0\sigma_0^2}\big)^{\frac{d_0}{2}}} \left(\frac{1}{\sigma_j^2}\right)^{\frac{d_0}{2}-1} \exp\big(-\frac{1}{\sigma_j^2}\frac{d_0\sigma_0^2}{2}\big) \\ &= C\big(\frac{1}{\sigma_j^2}\big)^{\frac{d_j}{2}+\frac{d_0}{2}-1} \exp\left(-\frac{1}{\sigma_j^2}\big(\frac{d_0\sigma_0^2 + d_j S_j^2}{2}\big)\right), \end{aligned} \tag{3.10}$$

wobei mit Γ die Gamma-Funktion (siehe auch Unterkapitel 6.1.1) und mit C die von σ_j^2 unabhängigen Faktoren bezeichnet werden. Betrachtet man die letzte Zeile von (3.10) als eine Funktion von $\frac{1}{\sigma_j^2}$, so stellt man fest, dass es sich hier um eine Gamma-Verteilung mit den Parametern $k = \frac{d_0+d_j}{2}$ und $\Theta = \frac{2}{d_0\sigma_0^2+d_j S_j^2}$ handelt. Dies bedeutet, dass für die a posteriori Verteilung von $\frac{1}{\sigma_j^2}$ gilt:

$$\left(\frac{1}{\sigma_j^2} \mid d_0, \sigma_0^2, S_j^2\right) \sim \Gamma(k = \frac{d_0 + d_j}{2}, \Theta = \frac{2}{d_0\sigma_0^2 + d_j S_j^2}). \tag{3.11}$$

Somit ist die Zufallsvariable σ_j^2 invers gammaverteilt mit

$$(\sigma_j^2 \mid d_0, \sigma_0^2, S_j^2) \sim \Gamma^{-1}(k = \frac{d_0 + d_j}{2}, \Theta = \frac{d_0\sigma_0^2 + d_j S_j^2}{2}). \tag{3.12}$$

Demzufolge erfüllt der a posteriori Erwartungswert von σ_j^2 die Gleichung:

$$E(\sigma_j^2 \mid d_0, \sigma_0^2, S_j^2) = \frac{\Theta}{k-1} = \frac{\frac{d_0\sigma_0^2+d_j S_j^2}{2}}{\frac{d_0+d_j}{2}-1} = \frac{d_0\sigma_0^2 + d_j S_j^2}{d_0 + d_j - 2}. \tag{3.13}$$

Für die Bestimmung des Erwartungswertes einer invers gammaverteilten Zufallsvariablen sei auf Kapitel 6.1.1 im Anhang auf Seite 99 verwiesen. Der Ausdruck auf der rechten Seite von (3.13) könnte nun als adjustierte Schätzung $\tilde{S}_j^2$ der Varianz σ_j^2 verwendet werden. In seinem Artikel [Smyth u. a. (2004)] schlägt der Verfasser (ohne Begründung) allerdings vor, dass

$$\tilde{S}_j^2 = \frac{d_0\sigma_0^2 + d_j S_j^2}{d_0 + d_j}. \tag{3.14}$$

verwendet wird. Diese Formel erlaubt eine anschauliche Interpretation dahingehend, dass die adjustierte Schätzung eine Kombination von der empirischen Stichprobenvarianz S_j^2 und dem Wert σ_0^2 darstellt. Fasst man σ_0^2 als eine „mittlere“ Varianz, so wird die individuelle Stichprobenvarianz S_j^2 zur mittleren hin geschrumpft. Für einen geringen Wert von d_0

und kleinere Stichprobengrößen n und m verhält sich Formel (3.14) aus praktischer Sicht besser als (3.13). Denn seien z.B. $n = m = 2$, dann berechnet sich der Nenner in (3.13) zu d_0. Sollte dieser Wert geringfügig größer als 0 sein, führen selbst kleinste Abweichungen zu einer wesentlichen Veränderung des Erwartungswertes.

3.3.3 Schätzung der Hyperparameter

Die Hyperparameter σ_0^2 und d_0 in (3.5) werden mit Hilfe der log-transformierten empirischen Varianz S_j^2 für $j = 1, \ldots, p$ geschätzt. Es lässt sich zeigen (siehe Seite 8 in Smyth u. a. (2004)), dass sie einer skalierten F-Verteilung folgen. Es gilt:

$$S_j^2 \sim \sigma_0^2 F_{d_0, d_j}. \tag{3.15}$$

Mit $z_j = \log S_j^2$ folgt z_j der Fisherschen Z-Verteilung plus Konstante:

$$z_j \sim \log F_{d_0, d_j} + \log \sigma_0^2.$$

Die Momente von z_j sind endlich und erfüllen folgende Gleichungen:

$$\mathrm{E}(z_j) = \log \sigma_0^2 + \psi(d_j/2) - \psi(d_0/2) + \log(d_0/d_j), \tag{3.16}$$

$$\mathrm{var}(z_j) = \psi'(d_j/2) + \psi'(d_0/2). \tag{3.17}$$

Dabei werden mit ψ und ψ' die Digamma- bzw. Trigamma-Funktion bezeichnet:

$$\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}$$

$$\psi'(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \log \Gamma(x).$$

Schätzen wir die ersten beiden Momente von z_j durch

$$\mathrm{E}(z_j) \approx \bar{z} = \frac{1}{p} \sum_{j=1}^{p} z_j$$

$$\mathrm{var}(z_j) \approx \frac{1}{p-1} \sum_{j=1}^{p} (z_j - \bar{z})^2,$$

so lässt sich d_0 als Lösung $\hat{d}_0$ von

$$\psi'(d_0/2) = \frac{1}{p-1} \sum_{j=1}^{p} (z_j - \bar{z})^2 - \psi'(d_j/2) \tag{3.18}$$

schätzen. Die Inverse einer Trigamma-Funktion besitzt keine geschlossene Form, deshalb muß (3.18) numerisch gelöst werden. Nachdem d_0 geschätzt wird, lässt sich ein Schätzer von σ_0^2 auf Grund von (3.16) wie folgt angeben:

$$\hat{\sigma}_0^2 = \exp(\bar{z} - \psi\Big(\frac{d_j}{2}\Big) + \psi\Big(\frac{\hat{d}_0}{2}\Big) - \log\Big(\frac{\hat{d}_0}{d_j}\Big)). \tag{3.19}$$

3.3.4 Schätzung der a posteriore Varianz und moderierte t-Statistik

Nachdem wir in Kapitel 3.3.3 die Schätzer für die Hyperparameter σ_0^2 und d_0 angegeben haben, können wir die a-posteriori Schätzer für die Varianzen bestimmen. Smyth schlägt als Erstes

$$\tilde{S}_j^2 = \frac{\hat{d}_0 \hat{s_0^2} + d_j S_j^2}{\hat{d}_0 + d_j}. \tag{3.20}$$

vor und konstruiert die moderierte t-Statistik analog (3.3):

$$\tilde{t}_j = \frac{\bar{y}_{j1} - \bar{y}_{j2}}{\tilde{S}_j \sqrt{\frac{1}{n} + \frac{1}{m}}}. \tag{3.21}$$

Es lässt sich zeigen (siehe Seite 9 in Smyth u. a. (2004)), dass unter der Nullhypothese $\mu_{j1} = \mu_{j2}$ die Statistik in (3.21) t-verteilt ist mit $d_j + \hat{d}_0$ Freiheitsgraden.

3.4 Bereinigung des Batch-Effektes

In diesem Unterkapitel wird in die Problematik des Batch-Effektes eingeführt und *ComBat* vorgestellt - ein Verfahren zur Adjustierung bezüglich Variablen, die für den Batch-Effekt verantwortlich vermutet werden.

3.4.1 Definition und Veranschaulichung

Die Durchführung eines Microarray-Experiments benötigt eine große Menge von chemischen Substanzen, eine ausgeklügelte Hardware und gut ausgebildete Spezialisten. Ändern sich die Rahmenbedingungen im Laufe des Experiments, werden die gemessenen Variablen simultan von biologischen und nichtbiologischen Faktoren beeinflusst. In diesem Kapitel wird der Batch-Effekt vorgestellt, eine weit verbreitete Quelle der nichtbiologischen Varianz bei Hochdurchsatz-Analysen. Eine mögliche Definition des Batch-Effektes lautet nach Lazar u. a. (2012):

> Der Batch-Effekt repräsentiert die systematischen technischen Unterschiede, falls die Proben in verschiedenen Gruppen (auch als Abfertigungsschübe, Fertigungslose oder vom Englischen eingedeutscht Batches) aufgearbeitet und gemessen wurden und welche in keinem Bezug zu der in dem Experiment gemessenen biologischen Varianz stehen.

Zur Veranschaulichung der Auswirkungen von einem Batch-Effekt dient ein Beispiel der Klassifikationsstudie UKN1, welche in 2.3.6 vorgestellt wurde. Auf Grund der technischen Begrenzung konnte man nicht alle 85 Proben zusammen erstellen. Somit mussten sie in 4 Gruppen aufgeteilt und separat analysiert werden. Die genaue Aufteilung ist in Tabelle 8 auf Seite 19 wiedergegeben. Es lag also eine Zeitspanne bis zu einigen Monaten zwischen den einzelnen Messungen, welche zu einem Batch-Effekt führen konnte. Für die deskriptive Analyse werden einzelne Hauptkomponentengraphiken erstellt, wobei die Anzahl in der Analyse verwendeten Variablen variiert wird. Auf Abbildung 8 sind zwei Hauptkomponentengraphiken dargestellt, wobei für das linke Bild 1000 und für das rechte 10000 Probesets mit der größten Varianz über alle Proben vorausgewählt werden. Auf der linken Grafik sieht man, dass sich die Proben zu zwei Clustern gruppieren. Es stellt sich heraus, dass diese Cluster die Proben genau nach der Batch-Zuordnung aus Tabelle 8 aufteilen.

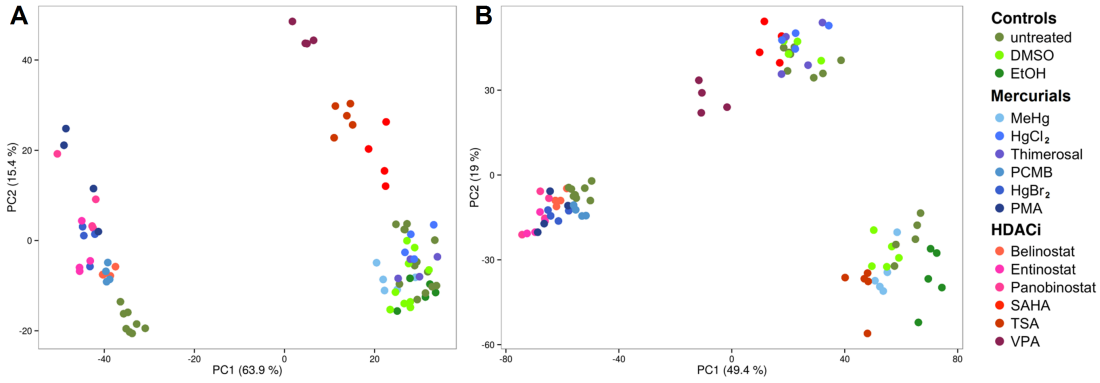

Abbildung 8: Darstellung der ersten beiden Hauptkomponenten der HKA der UKN1 Klassifikationsstudie. Für die Grafik (A) werden 1000 und die Grafik (B) 10000 Probesets mit der größten Varianz über alle Proben verwendet.

So befinden sich auf der linken Seite der Abbildung (A) die am spätesten erstellten Proben, welche zu dem Batch IV in Tabelle 8 zusammengefasst werden. Auf der rechten Seite befinden sich somit die Gruppen I bis III. Die erste Hauptkomponente, die über 60% der Gesamtvarianz erklärt, erfasst somit die Varianz zwischen zwei Clustern, deren Unterschiede kaum auf die alleinige Wirkung der Substanzen zurückzuführen wären. So befinden sich in beiden Clustern die unbehandelte Proben. Ein derart großer Unterschied zwischen den Genexpressionen von Zellen, welche lediglich mit einem Lösungsmittel behandelt wurden, ist am ehesten durch den Batch-Effekt zu erklären. Interessant ist an dieser Stelle die Tatsache, dass die zweite Hauptkomponente hauptsächlich die Varianz zwischen den Behandlungsklassen erfasst. So befinden sich jeweils am unteren Rand sowohl die unbehandelten Proben als auch die dem Lösungsmittel EtOH bzw. DMSO ausgesetzte Proben. Die mit quecksilberhaltigen Substanzen behandelten Proben haben jeweils einen

größeren Wert auf der zweiten Hauptkomponente, wobei mit HDACi inkubierte Zellen eine weitere Steigerung des Wertes auf der y-Achse aufweisen.

Erhöht man die Anzahl der in die Analyse aufgenommenen Probesets von 1000 auf 10000, erhält man ein anderes Bild (B). Nun sind 3 Cluster sichtbar, wobei die Aufteilung weiterhin nach den Fertigungslosen gemäß Tabelle 8 erfolgt. So bilden die Proben aus dem Batch I den Cluster rechts unten, aus den Schüben II und III den Cluster oben und die Proben aus dem Batch IV den Cluster am linken Rand. Auch innerhalb einzelner Cluster ist eine kontinuierliche Änderung der x-Koordinate je nach Behandlungsklasse festzustellen. Betrachtet man nacheinander unbehandelte Proben samt Lösungsmittel, quecksilberhaltige Substanzen und HDACi, so verkleinert sich innerhalb jedes Clusters der Wert der ersten Hauptkomponente.

Die Tatsache, dass der größte Anteil der Gesamtvarianz nicht durch die genotypische Wirkung der Substanzen, sondern durch die Abfertigungszeiten erklärt wird, ist nach Scherer (2009) ein deutliches Indiz für das Vorhandensein eines Batch-Effektes. Um die biologischen Einflüsse besser analysieren zu können, empfiehlt es sich alle möglichen Verzerrungen zu schätzen bzw. zu entfernen. Eine mögliche Vorgehensweise zum Entfernen des Batch-Effektes wurde in Leek u. a. (2010) beschrieben. Abbildung 9 zeigt einen schematischen Verlauf. Dabei schlagen die Autoren vor, zu Beginn eine deskriptive Analyse durchzuführen. Werden Artefakte, wie z.B. durch technische Einflüsse hervorgerufene Heterogenität in den Daten, (größtenteils) durch die für den Batch-Effekt verantworlichen Messungen (Batch-Variablen) wie Abfertigunsschübe oder -orte repräsentiert, so empfehlen die Autoren solche als Ersatzvariablen zu verwenden. Falls die Artefakte nicht mit den Batch-Variablen zu korrelieren scheinen, bietet die *Surrogat Variable Analysis* (**SVA**) [Leek u. Storey (2007)] einen möglichen Ausweg. Hierbei werden die für den Batch-Effekt verantwortlichen Faktoren aus den Daten geschätzt. Da sie keinen Messungen entsprechen, werden sie als Surrogat-Variablen bezeichnet. Die Autoren schlagen anschließend vor, je nach Fall die Batch- oder die Surrogat-Variablen in den nachfolgenden Analysen zu verwenden oder für diese Effekte mit dem nachfolgend vorgestelltem *ComBat* Algorithmus zu adjustieren.

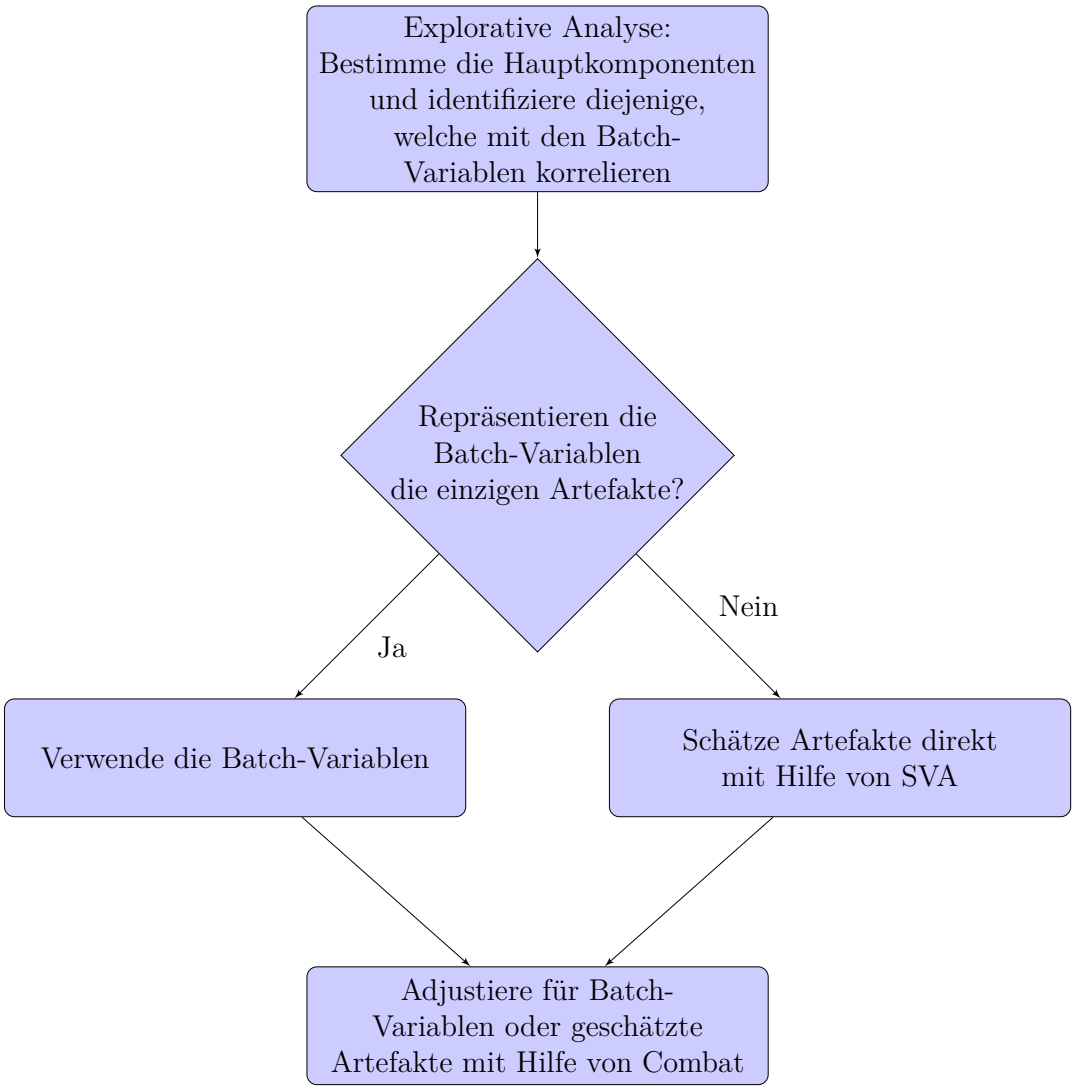

Abbildung 9: Eine schematische Darstellung des Verlaufes der Batch-Korrektur nach Leek u. a. (2010). Zu Beginn wird mit Hilfe von explorativen Methoden bestimmt, ob die Batch-Variablen wie z.B. Hybridisierungszeit die einzigen Artefakte repräsentieren. Dazu werden die Hauptkomponenten bestimmt und überprüft, inwiefern die Batch-Variablen damit korrelieren. Lassen sich Artefakte alleine auf Batch-Variablen zurückführen, werden sie für die ComBat-Adjustierung verwendet. Falls die bekannten Batch-Variablen nicht die Artefakte repräsentieren, wird mit Hilfe des **SVA**-Pakets die Surrogat-Variablen zuerst aus den Daten geschätzt und dann für die ComBat-Adjustierung benutzt.

3.4.2 ComBat Algorithmus

In diesem Abschnitt stellen wir das ComBat [Johnson u. a. (2007)] Verfahren vor, dessen Überlegenheit anderen Verfahren wie *distance-weighted discrimination*, *mean-centering PAMR* oder *Ratio_G* gegenüber in Chen u. a. (2011) nachgewiesen wurde. ComBat zeigte dabei im Vergleich bessere Genauigkeit und Präzision. Dabei wird angenommen, dass für die Expression Y_{jkl} des Gens j aus dem Batch k der Probe l das Modell

$$Y_{jkl} = \alpha_j + X\beta_j + \eta_{jk} + \delta_{jk}\epsilon_{jkl}, \qquad j = 1, \ldots, p, k = 1, \ldots, K, l = 1, \ldots, n \tag{3.22}$$

gilt. Dabei stellt α_j die mittlere Expression des Probesets dar, X die zentrierte Designmatrix und β_j die Regressionskoeffizienten. Der Fehlerterm ϵ_{jkl} folgt einer Normalverteilung mit Erwartungswert Null und Varianz σ_j^2. Die Parameter η_{jk} und δ_{jk} repräsentieren einen additiven bzw. multiplikativen Effekt des Batches k auf das Probeset j.

Als Erstes werden die Daten normalisiert. Dazu werden die Modellparameter α_j, β_j und η_{jk} mit Hilfe der Methode der kleinsten Quadrate unter der Nebenbedingung

$$\sum_{k=1}^{K} n_k \hat{\eta}_{jk} = 0$$

geschätzt, wobei mit n_k die Anzahl der Proben in Batch k bezeichnet wird. Dann wird die Varianz σ_j^2 wie folgt geschätzt:

$$\hat{\sigma}_j^2 = \frac{1}{nK} \sum_{kl} (Y_{jkl} - \hat{\alpha}_j - X\hat{\beta}_j - \hat{\eta}_{jk})^2.$$

Nun werden die normalisierten Werte der Genexpression bestimmt

$$Z_{jkl} = \frac{Y_{jkl} - \hat{\alpha}_j - X\hat{\beta}_j}{\hat{\sigma}_j}.$$

Im nächsten Schritt nehmen wir an, dass die Werte Z_{jkl} der Normalverteilung

$$Z_{jkl} \sim N(\gamma_{jk}, \delta_{jk}^2) \tag{3.23}$$

folgen. Weitere Annahmen lauten, dass die a-priori Verteilungen der Parameter γ_{jk} und δ_{jk}^2 einer Normalverteilung bzw. einer inversen Gammaverteilung genügen:

$$\gamma_{jk} \sim N(\gamma_k, \tau_k^2) \tag{3.24}$$

$$\delta_{jk}^2 \sim \Gamma^{-1}(\lambda_k, \Theta_k). \tag{3.25}$$

Die Hyperparameter $\gamma_k, \tau_k^2, \lambda_k$ und Θ_k werden dabei empirisch unter der Verwendung der Momentenmethode aus den Daten geschätzt (vergleiche Kapitel 3.4.3). Unter Verwendung

des Satzes von Bayes lassen sich nun die a-posteriori Erwartungswerte von γ_{jk} und δ_{jk}^2 wie folgt bestimmen:

$$E(\gamma_{jk} \mid Z_{jkl}, \delta_{jk}^2) = \frac{\tau_k^2 \sum_l Z_{jkl} + \delta_{jk}^2 \gamma_k}{n_k \tau_k^2 + \delta_{jk}^2} \tag{3.26}$$

$$E(\delta_{jk}^2 \mid Z_{jkl}, \gamma_{jk}) = \frac{\Theta_k + \frac{1}{2} \sum_l (Z_{jkl} - \gamma_{jk})^2}{\frac{n_k}{2} + \lambda_k - 1}. \tag{3.27}$$

Die beiden Behauptungen werden im Lemma 2 des Anhangs bewiesen. Somit lassen sich die beiden Parameter γ_{jk} und δ_{jk}^2 mit Hilfe der Schätzungen der Hyperparameter $\bar{\tau}_k, \bar{\gamma}_k, \bar{\Theta}_k$ und $\bar{\lambda}_k$ wie folgt annähern

$$\gamma_{jk}^* = \frac{\bar{\tau}_k^2 \sum_l Z_{jkl} + \delta_{jk}^{2*} \bar{\gamma}_k}{n_k \bar{\tau}_k^2 + \delta_{jk}^{2*}} \tag{3.28}$$

$$\delta_{jk}^{2*} = \frac{\bar{\Theta}_k + \frac{1}{2} \sum_l (Z_{jkl} - \gamma_{jk}^*)^2}{\frac{n_k}{2} + \bar{\lambda}_k - 1}. \tag{3.29}$$

Dabei hängt die Schätzung von γ_{jk}^* in (3.28) von der Schätzung δ_{jk}^{2*} in (3.29) ab und umgekehrt. Somit stellen die Gleichungen (3.28) und (3.29) keine Schätzung in einer geschlossenen Form, sondern sind als Iterationsvorschriften zu interpretieren. Man wählt einen sinnvollen Startwert von einem Parameter, z.B. $\frac{1}{n_k} \sum_l Z_{jkl}$ für γ_{jk}^*, und bestimmt eine Schätzung für δ_{jk}^{2*}, welche man für das Schätzen von γ_{jk}^* wiederum verwendet. Die Verfassser in Johnson u. a. (2007) geben an, dass bereits 30 Iterationen für eine Konvergenz genügen.

Nach dem Berechnen der Schätzungen für die Batch-Effekte werden nun die Beobachtungswerte adjustiert:

$$\bar{Y}_{jkl} = \hat{\sigma}_j \left(\frac{Z_{jkl} - \gamma_{jk}^*}{\delta_{jk}^*} \right) + \hat{\alpha}_j + X\hat{\beta}_j.$$

Anschaulich gesprochen zieht man den geschätzen additiven Effekt ab und teilt durch den multiplikativen Effekt. Danach führt man die Varianz des Fehlerterms wieder ein. Zuletzt werden die mittlere Expression und die Behandlungseffekte wieder aufaddiert. In Abbildung 10 wird der Algorithmus skizzenhaft erläutert.

1. Schätzen von $\alpha_j, \beta_j, \eta_{jk}$ und σ_j^2.
2. Normalisieren der Werte:
$$Z_{jkl} = \frac{Y_{jkl} - \hat{\alpha}_j - X\hat{\beta}_j}{\hat{\sigma}_j}.$$
3. Schätzen von Hyperparametern $\gamma_k, \tau_k^2, \lambda_k, \Theta_k$ mit Hilfe von Momentenmethode.
4. Schätzung des additiven Batch-Effekts δ_{jk}^* und des multiplikativen Batch-Effektes γ_{jk}^* durch a-posteriori Erwartungswert und Varianz der normalisierten Expression Z_{jkl}.
5. Adjustierung der Werte
$$\bar{Y}_{jkl} = \frac{\hat{\sigma}_j}{\delta_{jk}^*}(Z_{jkl} - \gamma_{jk}^*) + \hat{\alpha}_j + X\hat{\beta}_j.$$

Abbildung 10: ComBat Algorithmus

Wendet man die ComBat-Adjustierung auf die am Anfang des Kapitels angesprochenen Daten der UKN1-Klassifikationsstudie, erhält man einen Hauptkomponentenplot (Abbildung 46 im Anhang), bei welchem kein Batch-Effekt sichtbar ist. Ferner bilden die Proben je nach Behandlungsart (Kontrolle, behandelt mit quecksilberhaltigen Substand oder HDAC Inhibitor) relativ kompakte Gruppen. Dies ist ein Hinweis, dass der Batch-Effekt reduziert wurde.

3.4.3 Schätzen der Hyperparameter

In diesem Unterkapitel werden die Schätzer für die Hyperparameter $\gamma_k, \tau_k^2, \lambda_k$ und Θ_k aus (3.24) und (3.25) angegeben. Bezeichne dabei $\hat{\gamma}_{jk}$ die mittlere Expression des Probesets j aus dem Batch k:
$$\hat{\gamma}_{jk} = \frac{1}{n_k}\sum_l Z_{jkl}.$$
Dann lassen sich die Hyperparameter γ_k und τ_k^2 wie folgt schätzen:
$$\bar{\gamma_k} = \frac{1}{p}\sum_j \hat{\gamma}_{jk} \tag{3.30}$$
$$\bar{\tau_k}^2 = \frac{1}{p-1}\sum_j (\hat{\gamma}_{jk} - \bar{\gamma_k})^2. \tag{3.31}$$

Ferner bezeichne $\hat{\delta}_{jk}^2$ die Stichprobenvarianz für die Expression des Probesets j aus dem Batch k:

$$\hat{\delta}_{jk}^2 = \frac{1}{n_k - 1} \sum_l (Z_{jkl} - \hat{\gamma}_{jk})^2.$$

Nun schätzen wir die ersten beiden Momente $\bar{V}_k$ und $\bar{S}_k^2$ von $\hat{\delta}_{jk}^2$ wie folgt

$$\bar{V}_k = \frac{1}{p} \sum_j \hat{\delta}_{jk}^2$$
$$\bar{S}_k^2 = \frac{1}{p-1} \sum_j (\hat{\delta}_{jk}^2 - \bar{V}_k)^2.$$

Um die Hyperparameter λ_k und Θ_k jetzt zu schätzen, muß man $\bar{V}_k$ und $\bar{S}_k^2$ den theoretischen Momenten einer inversen Gammaverteilung $\Gamma^{-1}(\lambda_k, \Theta_k)$ gleichsetzen, nämlich dem Erwartungswert $\frac{\Theta_k}{\lambda_k - 1}$ und der Varianz $\frac{\Theta_k^2}{(\lambda_k-1)^2(\lambda_k-2)}$. Löst man dieses Gleichungssystem (siehe Lemma 3), erhält man

$$\bar{\lambda}_k = \frac{\bar{V}_k + 2\bar{S}_k^2}{\bar{S}_k^2} \tag{3.32}$$

$$\bar{\Theta}_k = \frac{\bar{V}_k^3 + \bar{V}_k \bar{S}_k^2}{\bar{S}_k^2}. \tag{3.33}$$

Diese Schätzungen werden nun für das Bestimmen von γ_{jk}^* und δ_{jk}^{2*} verwendet, welche für die Adjustierung von Expressionswerten $\bar{Y}_{jkl}$ benutzt werden.

3.5 Klassifikationsverfahren

Eine Zuordnung von Objekten in vorgegebenen Klassen stellt eine wichtige Aufgaben in der Statistik. Je nach Anwendungsgebiet wird dies als Klassifikation, Mustererkennung oder als Diskriminanz-Analyse genannt. Basierend auf der als Trainingsmenge bezeichneten Objektgruppe, deren Klasseneinteilung bekannt ist, wird eine Zuordnungsfunktion erstellt, welche eine neue Beobachtung aus der Testmenge in die Klassen einteilt. Anwendungsgebiete der Klassifikation neben der Biostatistik sind u.A. Quantifizierung der Kreditwürdigkeit (Kreditscoring), Handschrifterkennung und Spracherkennung.

Aus der Fülle der vorhandenen und etablierten Methoden zur Mustererkennung werden in den vorliegenden Analysen zwei Algorithmen genommen: Support Vector Machines (**SVM**) und Random Forests. Wenngleich keines der bekannten Verfahren nach dem No Free Lunch Theorem [Wolpert u. Macready (1997)] a priori besser ist, haben sich die oben genannten Verfahren in den meisten Fällen (man siehe z.B. Díaz-Uriarte u. De Andres (2006)) als sehr gute erwiesen.

3.5.1 Support Vector Machines

In diesem Abschnitt folgt eine kurze Einführung in die Theorie der Support-Vector-Maschinen (SVM). In der darauffolgenden Analyse wird eine bestimmte Version des SVM verwendet, die nur den linearen Kern verwendet. Deshalb wird an dieser Stelle auf die Präsentation der verschiedenen Kerne und des Kern-Tricks verzichtet. Ferner wird im Falle von hochdimensionalen Räumen von der linearen Separierbarkeit der Objekte ausgegangen.

Eine Support Vector Mashine (SVM) ist ein Klassifikator, welcher seit Einführung in Boser u. a. (1992) vermehrt verwendet wird. Während die Ergebnisse der Klassifikation nicht immer einfach zu interpretieren sind, z.B. bei der Verwendung des Gauß-Kerns, und die SVM deshalb auch als „black box classifier“ bezeichnet wird, zeichnet sich diese Entscheidungsregel durch (meistens) sehr gute Generalisierungseigenschaften aus. Es wird angenommen, dass der Grund für die guten Generalisierungseigenschaften von SVMs in der Verwendung der Idee vom „breiten Rand“ (Large-Margin-Classifier) liegt. Diese Idee wird nun anhand folgender Skizze (siehe Abbildung 11) erläutert.

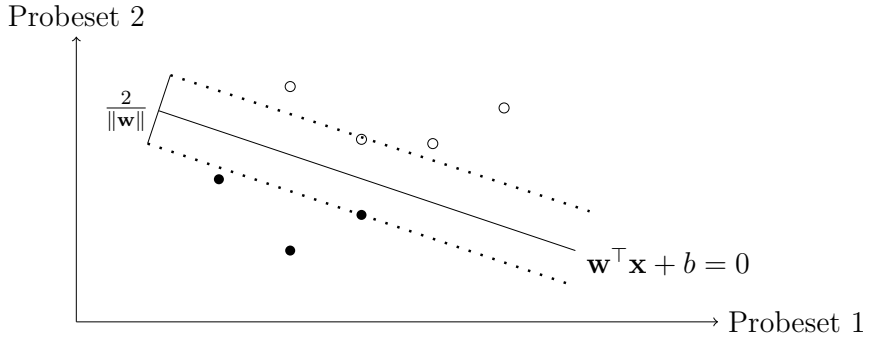

Abbildung 11: Trennung von zwei Klassen im $\mathbb{R}^2$ durch Hyperebene H_1. Die verschiedenen Klassen werden durch schwarze bzw. nicht ausgefüllte Kreise symbolisiert. Die durchgezogene Linie stellt die Hyperebene dar, welche durch den orthogonalen Vektor $\mathbf{w}$ und den Abstand b eindeutig definiert ist. Die gestrichelten Linien stellen die beiden Ränder dar, deren Abstand voneinander $\frac{2}{\|\mathbf{w}\|}$ beträgt.

Es werden dabei exemplarisch Objekte in $\mathbb{R}^2$ betrachtet, die in zwei Klassen eingeordnet sind. Sollen diese Klassen durch eine Hyperebene (Gerade in $\mathbb{R}^2$) linear trennbar sein, so gibt es grundsätzlich unendlich viele Möglichkeiten, dies zu tun. Nun wird die Hyperebene H_1 so gelegt, dass der von zu klassifizierenden Objekten freie Bereich möglichst groß wird. Dieser Bereich wird durch zu H_1 parallele Hyperebenen eingegrenzt und wird in der Literatur meistens als Margin (Rand) bezeichnet. Die Objekte, die auf den den Margin eingrenzenden Hyperebenen liegen, werden Stützvektoren (Support Vectors) genannt, sie „tragen“ den Rand und sind für die Lage von H_1 verantwortlich. Wird nun ein neues Ob-

jekt betrachtet, so hängt seine Zuordnung allein davon ab, in welcher der zwei Halbebenen bzgl. H_1 es liegt.

Für die mathematische Formulierung betrachten wir die Menge

$$\Big\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n) | \mathbf{x}_i \in X, y_i \in \{-1, +1\}\Big\},$$

die als Trainingsmenge bezeichnet wird. Der Koordinatenvektor $\mathbf{x}_i$ repräsentiert dabei das $i-$te Objekt, y_i ist das Klassenlabel. Die zwei Klassen trennende Hyperebene, deren Existenz momentan vorausgesetzt wird, ist durch den Normalenvektor $\mathbf{w}$ und den Abstand b eindeutig definiert. Nun sind $\mathbf{w}$ und b bis auf das Vielfache eindeutig und lassen sich so skalieren, dass für Objekte der Klasse mit $y_i = -1$ die Beziehung

$$\mathbf{w}^\top \mathbf{x}_i + b \leq -1 \tag{3.34}$$

und der Klasse $y_i = 1$ die Beziehung

$$\mathbf{w}^\top \mathbf{x}_i + b \geq 1 \tag{3.35}$$

gilt. Diese Beziehungen lassen sich beide durch die Gleichung

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \tag{3.36}$$

für alle $i = 1, ..., n$ zusammenfassen.

Nun betrachten wir jeweils einen Punkt auf dem jeweiligen Rand auf beiden Seiten. Diese zwei Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ erfüllen die Gleichungen

$$\mathbf{w}^\top \mathbf{x}_1 + b = -1$$
$$\mathbf{w}^\top \mathbf{x}_2 + b = 1.$$

Zieht man die erste Gleichung von der zweiten ab, bekommt man die Relation

$$\mathbf{w}^\top (\mathbf{x}_2 - \mathbf{x}_1) = 1 - (-1) = 2.$$

Anschaulich interpretiert bedeutet dies: Projizieren wir die Differenz $\mathbf{x}_2 - \mathbf{x}_1$ auf den Vektor $\mathbf{w}$, so erhalten wir den Wert 2. Normiert auf die Länge von $\mathbf{w}$, ergibt sich $\frac{2}{\|\mathbf{w}\|}$ als die Breite des Randes. Da man die Optimierungsaufgaben als Minimierungssproblem formuliert, betrachten wir den Kehrwert der Randbreite, den es nun unter Nebenbedingungen zu minimieren gilt. Somit lässt sich die optimale Hyperebene als Lösung des Optimierungsproblems

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2} \tag{3.37}$$

unter der Nebenbedingung $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$ für alle $1 \leq i \leq n$ konstruieren. Ein neuer Punkt $\mathbf{x}$ wird nun auf Grund seiner Lage bezüglich H_1 einer der beiden Klassen zugeordnet. Die Entscheidungsfunktion hat die Form

$$f(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}^\top \mathbf{x} + b). \tag{3.38}$$

Das Problem (3.37) ist quadratisch bzw. konvex und lässt sich mit Hilfe der Lagrange-Multiplikatoren lösen. Dabei wird der Normalenvektor als Linearkombination der Trainingsobjekte geschrieben:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i. \tag{3.39}$$

Dabei sind Koeffizienten α_i für viele $1 \leq i \leq n$ gleich Null. Somit wird in (3.39) nur über eine Teilmenge der Trainingspunkte summiert. Es lässt sich zeigen, dass dies genau die Support Vektoren sind, also diejenigen Punkte, die auf dem Rand liegen. Der Gewichtsvektor $\mathbf{w}$ lässt jetzt eine Anordnung der Variablen ihrer Wichtigkeit nach zu: Je größer der absolute Wert von w_j, desto wichtiger ist die Variable x_j.

Probabilistische Klassifikation In bestimmten Fällen ist es wünschenswert, zusätzlich zu einer Klassenzuordnung auch eine Klassenwahrscheinlichkeit zu erhalten. In dem Bereich des maschinellen Lernens wird dies als probabilistische Klassifikation bezeichnet. Während einige Lernmethoden auf eine natürliche Weise die Klassenwahrscheinlichkeiten bestimmen, wie z.B. ein Bayes-Klassifikator oder die logistische Regression, kann im Falle von Support Vektor Maschinen der vorzeichenbehaftete Abstand $d(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ des neuen Punktes $\mathbf{x}$ zu der Hyperebene H_1 in eine Wahrscheinlichkeit umgerechnet werden. Eine verbreitete Methode des Umrechnens stellt die Skalierung von Platt (1999) dar. Das Vorgehen wird nun kurz erläutert.

Als Erstes transformieren wir die Variable y_i, sodass die neue Variable t_i ihren Wertebereich zwischen 0 und 1 hat

$$t_i = \frac{y_i + 1}{2}.$$

Ihre Verteilung wird mit dem Modell der Binomialverteilung beschrieben. Die unbekannte Klassen-Wahrscheinlichkeit p_i hängt von der stetigen Einflussvariablen $d(\mathbf{x}_i)$ ab. Entsprechend der logistischen Regression lautet der Ansatz nun

$$p_i = \frac{1}{1 + \exp\Big(A \cdot d(\mathbf{x}_i) + B\Big)}, \tag{3.40}$$

wobei A und B zu bestimmen sind. Zur Lösung bestimmt man die zu minimierende negative Log-Likelihood-Funktion

$$\min_{A,B} - \sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i). \tag{3.41}$$

Die Minimierungsaufgabe in (3.41) lässt sich problemlos lösen, es sei denn, die vorzeichenbehaftete Abstände $d(\mathbf{x}_i)$ mit $t_i = 0$ sind perfekt von den $d(\mathbf{x}_i)$ mit $t_i = 1$ getrennt. Diese Situation wird auch als Hauck-Donner Phänomen bezeichnet, siehe [Ripley u. Venables (1994), S.198]. Für seine Methode schlägt Platt vor, t_i als Zufallsvariable zu betrachten, die mit einer gewissen Wahrscheinlichkeit auch entgegengesetzte Werte annehmen kann. Dies bedeutet, dass für Beobachtungen mit $y_i = 1$ ein Wert von $t_i = 1 - \epsilon_+$ für ein zu bestimmendes kleines $\epsilon_+ > 0$ verwendet wird. Um die Wahrscheinlichkeit des richtigen Labels für Beobachtungen mit $y_i = 1$ zu bestimmen, nehmen wir zuerst an, dass die a priori Wahrscheinlichkeitsverteilung betaverteilt mit Parametern $p = q = 1$ und somit nichtinformativ sei. Beobachten wir nun N_+ positive Proben, so ist die a posteriori Verteilung auch betaverteilt mit Parametern $p = 1 + N_+$ und $q = 1$ (siehe Lemma 4 im Anhang). Somit lautet der MAP-Schätzer für die Zielwahrscheinlichkeit der positiven Beobachtungen

$$t_+ = \frac{p}{p+q} = \frac{1+N_+}{2+N_+}.$$

Analog ist der MAP-Schätzer im Falle des Auftretens von N_- negativen Proben

$$t_- = \frac{1}{2+N_-}.$$

Diese Werte werden nun an Stelle von t_i in (3.41) verwendet. Da es sich dabei um Kubler-Leibler-Abstand zwischen t_i und $d(\mathbf{x}_i)$ handelt, lässt sich die Optimierungsaufgabe auch für das nicht-binäre t_i lösen.

3.5.2 Random Forests

In diesem Abschnitt werden die Grundlagen von Random Forests erläutert. In Breiman (2001) eingeführt, entwickelte sich dieses Verfahren zu einer beliebten Methode der Klassifizierung bzw. Regression. Die Methode ist ein Hybrid aus *Bagging* und *Random-Subspace*-Methode [Sammut u. Webb (2011)]. Analog zu Bagging wird ein Ensemble von B Entscheidungsbäumen erstellt, welche jeweils auf einer Teilmenge S der Trainingsdaten gebildet wurden (für eine kurze Erläuterung von Entscheidungsbäumen sei auf Kapitel 6.1.3 im Anhang verwiesen). Man geht davon aus, dass die Entscheidungsbäume approximativ unverfälscht sind, allerdings eine hohe Varianz aufweisen (in ihrem Buch Hastie u. a. (2011) bezeichnen die Autoren die Entscheidungsbäume als „notoriously noisy“). Durch das Bilden eines Mittelwertes bzw. einer Mehrheitsentscheidung wird die Varianz des Modells verkleinert. Betrachtet man einen Durchschnitt von B u.i.v Zufallsvariablen mit Varianz σ^2, so lautet seine Varianz $\frac{1}{B}\sigma^2$. Sind die Variablen allerdings gleichverteilt und paarwei-

se korreliert mit Korrelationskoeffizienten ρ, so ist die Varianz des Durchschnitts (siehe Lemma 5 im Anhang auf Seite 105)

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{3.42}$$

Erhöht sich die Anzahl der Bäume B, so konvergiert der zweite Summand gegen Null. Der erste Summand bleibt dagegen konstant und lässt die Varianz des Mittelwertes minimal $\rho\sigma^2$ sein. Die Random-Subspace-Methode versucht die Korrelation zwischen den Bäumen zu reduzieren, indem nur eine zufällig ausgewählte Menge der Variablen für die Konstruktion der Bäume verwendet wird. Somit versucht man die Reduktion der Varianz auch für die korrelierten Bäume zu verbessern. In Abbildung 12 wird der Algorithmus skizzenhaft erläutert, wobei mit $\lfloor \dots \rfloor$ die Abrundungsfunktion (Gaußklammer) bezeichnet wird.

1. Für $i = 1, ..., B$ wiederhole:
   a) Ziehe mit Zurücklegen eine Stichprobe S der Größe N aus der Trainingsmenge.
   b) Erstelle einen Entscheidungsbaum T_i für die Stichprobe durch rekursives Wiederholen folgender Schritte für jeden Endknoten, bis die minimale Knotenanzahl n_{min} erreicht ist:
      i. Wähle zufällig m aus p Variablen.
      ii. Bestimme die beste Variable bzw. Schranke aus den vorhandenen m Variablen.
      iii. Teile den Knoten in zwei Knoten auf.
2. Sei $\hat{C}_i(x)$ die Klassen-Vorhersage des Entscheidungsbaumes T_i für das neue Objekt x. Dann lautet die Mehrheitsentscheidung des Algorithmus:

$$\hat{C}_{RF}(x) = \lfloor \frac{1}{2} + \frac{1}{B}\sum_{i=1}^{B} \hat{C}_i(x) \rfloor$$

   Im Falle einer probabilistischen Klassifikation wird die Wahrscheinlichkeit $\hat{P}_{RF}(x)$ für das Objekt x wie folgt geschätzt:

$$\hat{P}_{RF}(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{C}_i(x).$$

Abbildung 12: Random Forests Algorithmus

Für die praktische Anwendung von Random Forests wurde das **R**-Paket *caret* verwendet. Es erlaubt eine kreuzvalidierte Optimierung der Parameter auf der Trainingsmenge

und eine probabilistische Klassifikation der Objekte aus der Testmenge. Die Wahrscheinlichkeit der Klasse mit dem Klassenlabel $y = 1$ wird dabei durch

$$\hat{P}_{RF}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{C}_i(x)$$

auf eine intuitive Weise geschätzt, wobei $\hat{C}_i(x)$ die Klass-Vorhersage des Entscheidungsbaumes T_i für das neue Objekt x ist. Der verwendete Skript ist im Anhang auf Seite 233 angegeben.

3.5.3 Sensitivitätsanalyse

Die dieser Analyse zu Grunde liegenden Expressionsdaten sind auf vielfache Weise verrauscht. Sowohl die biologische Variabilität der Zellen als auch die Technik der Genexpressionsmessung und deren praktische Umsetzung führen zu einer erhöhten Varianz der gemessenen Daten. Stellt man sich zur Aufgabe, die Beobachtungen auf Grund von Expressionswerten zu klassifizieren, erscheint es sinnvoll, das Auswirken dieses Rauschens auf die Leistung des Lernverfahrens zu untersuchen. In diesem Unterkapitel stellen wir eine Methode dar, welche es ermöglicht, das Abschneiden der Lernmethode nach einem zusätzlichen Verrauschen der Daten zu analysieren. Als Eingaben fungieren folgende Daten bzw. Parameter:

- Die Expressionsmatrix $X = (x_{ij})_{i=1,\ldots,n,j=1,\ldots,p}$ enthält die Messungen von p Genprodukten für n Proben.
- Die Anzahl der Variablen k gibt an, wie viele Probesets zur Klassifikation vorausgewählt werden.
- Der Rauschanteil (Noise) $t \in [0, 1]$ gibt den Anteil der Variablen an, die verrauscht werden. Der Wert $t = 0$ bedeutet, dass kein Verrauschen eingeführt wird, der Wert $t = 1$ bedeutet, dass alle Variablen verrauscht werden.
- Die Rauschstärke (Amplifier) ρ reguliert den Ausmaß des Rauschens. Je höher ρ ist, desto stärker werden die Daten verrauscht.

Zuerst werden die Daten X zufällig in Trainings- und Testmenge aufgeteilt. Dann werden die k Probesets mit der höchsten Varianz über die Proben in der Trainingsmenge bestimmt. Die beiden Mengen werden auf diese k Probesets reduziert und verrauscht, indem man zu einem Probeset x_{ij} einen Summanden aufaddiert, je nachdem, ob eine zufällig gezogene Zahl p_{ij} den Wert $1 - t$ übersteigt. Dieser Summand ist hierbei normalverteilt mit Erwartungswert 0 und einer Standardabweichung, welche multiplikativ von der Rauschstärke ρ und der ursprünglichen Standardabweichung von x_{ij} abhängt.

Nach dem Verrauschen der Daten wird das Lernverfahren auf der Trainingsmenge gebildet und die Klassen-Wahrscheinlichkeiten für die Testmenge bestimmt. Hier folgt eine schematische Übersicht des Vorgehens:

1. Definiere Rauschanteil (Noise) t und Rauschstärke (Amplifier) ρ.
2. Teile die Daten X in Trainings- und Testmenge.
3. Bestimme k Probesets mit der höchsten Varianz über die Proben im Trainingsset und reduziere die beiden Mengen auf diese Variablen.
4. Addiere Rauschen zu Training- und Testmenge:
   a) Bestimme die Standardabweichung der Probesets im reduzierten Trainingsset $\sigma_j = \mathrm{sd}(x_j)$ für $j = 1, ..., k$.
   b) Ziehe $n \times k$ unabhängige Realisierungen p_{ij} einer Zufallsvariable aus einer Rechtecksverteilung auf $[0, 1]$, wobei $i = 1, ..., n$ und $j = 1, ..., k$.
   c) Erzeuge ausgehend von p_{ij} Realisierungen z_{ij} einer standardnormalverteilten Zufallsvariablen mit der Inversionsmethode.
   d) Addiere Rauschen zu dem Expressionsniveau x_{ij}, falls die Realisierung p_{ij} größer als der Wert $1 - t$ ist:

$$x_{ij} = \begin{cases} x_{ij} + \rho\sigma_j z_{ij} & \text{falls } p_{ij} > 1 - t \\ x_{ij} & \text{falls } p_{ij} \leq 1 - t \end{cases}$$

5. Trainiere das Lernverfahren auf die geänderte Trainingsmenge.
6. Bestimme die Klassenwahrscheinlichkeiten für die geänderte Testmenge.

Abbildung 13: Sensitivitätsanalyse

Während der vorgestellte Algorithmus sowohl die Trainings- als auch die Testmenge verrauscht, werden bei der statistischen Auswertung auch Szenarien verfolgt, bei welchen nur die Trainingsmenge oder nur die Testmenge perturbiert werden.

4 Statistische Auswertung

In diesem Kapitel werden die vorgestellten Methoden auf die Daten angewendet und die Ergebnisse präsentiert. Alle Analysen wurden mit Hilfe der freien Programmiersprache für statistisches Rechnen **R** [R Core Team (2013)(Version 3.1.2)] durchgeführt. Die verwendeten Befehle sind als .R-Dateien im Anhang aufgelistet. Der erste Teil der Auswertung in den Unterkapiteln 4.1-4.4 befasst sich mit der Analyse von den in Kapiteln 2.3.1-2.3.4 vorgestellten Konzentrationsstudien. Entsprechend der auf Seite 4 vorgestellten Pipeline werden die Daten zuerst deskriptiv analysiert. Dazu werden Hauptkomponentenanalyse durchgeführt und Heatmaps abgebildet. Diese Ergebnisse liefern einen Überblick über die Qualität der Daten, sodass im Falle einer schlechten Qualität auf weitere Analysen verzichtet werden kann, wie im Falle der MeHg Chronic Konzentrationsstudie in Unterkapitel 4.3.

Sind deskriptiven Analysen zufriedenstellend, werden differentielle Probesets bestimmt. Dazu werden die jeweiligen Expressionen der behandelten Zellen mit den Expressionen der Kontrollen verglichen. Nach der vorgestellten Pipeline wären die nächsten Schritte eine Anreicherungsanalyse von Einträgen der Gene Ontology und eine Analyse der Bindestellen. Während in den vom Verfasser mitgeschriebenen Publikationen [Krug u. a. (2013), Waldmann u. a. (2014) und Balmer u. a. (2014)] dies auch gemacht wurde, wird in der vorliegenden Arbeit darauf verzichtet. Im Übrigen stimmen die Analysen in den Veröffentlichungen mit den in dieser Arbeit vorgestellten Analysen überein.

Der zweite Teil der Auswertung (Unterkapitel 4.5 und 4.6) befasst sich mit der Analyse der Klassifikationsstudien UKN1 und UKK, welche in den Abschnitten 2.3.6 und 2.3.7 eingeführt wurden. Als Lernverfahren werden zwei bekannte Diskriminanz-Methoden verwendet: Support Vector Machines und Random Forests. Für jede Kombination aus einer Studie und einem Lernverfahren werden folgende Fragestellungen nacheinander adressiert:

- Auswahl der Prädiktoren,
- Analyse der Generalisierungseigenschaft des Lernverfahrens mit Hilfe einer Kreuzvalidierung,
- Einfluss der Anzahl verwendeter Replikate,
- Einfluss von einem zusätzlichen Verrauschen der Daten (Sensitivitätsanalyse),
- Validierung des Lernverfahrens auf einem externen Datensatz.

Die ersten Schritte - deskriptive Verfahren und eventuelle Batch-Adjustierung - stimmen mit denen der Auswertung der Konzentrationsstudien überein. Im nächsten Schritt werden die Expressionswerte der Kontrollen von denen der behandelten Proben abgezogen. Dadurch erhofft man sich die durch die Hinzugabe von Lösungsmitteln hervorgeru-

fenen Effekte zu reduzieren, so dass die verschiedenen Behandlungsgruppen miteinander vergleichbar sind.

Im zweiten Schritt wird ermittelt, welche und wie viele Prädiktoren für das Bestimmen der Klassifikationsregel verwendet werden sollten. Einerseits befinden sich unter den insgesamt zur Verfügung stehenden über 50.000 Probesets auch solche, die keine biologisch relevante Information enthalten. Als Beispiel dazu dienen sogenannte *spike-in* Probesets. Von ihnen ist a priori bekannt, dass sie mit keinem der biologischen Einflussfaktoren assoziiert sind. Sie fungieren somit als negative Kontrollen. Ferner wurde in Golub u. a. (1999) gezeigt, dass die meisten Gene auf die Klassifikationsgüte keinen Einfluss haben. Dabei ist die Gefahr der Überanpassung um so größer, je mehr Variablen man zur Hand hat [Babyak (2004)]. Deshalb spielen Methoden der Dimensionsreduktion [Guyon u. a. (2006)] in der Microarray-Analyse eine wichtige Rolle. In vielen Untersuchungen werden sie vor der eigentlichen Klassifikation eingesetzt. Hier ist eine kleinere Übersicht:

- In Sreepada u. a. (2014) wurden zuerst die Hauptkomponenten gebildet und dann mit Hilfe von Genetischen Algorithmem [Rajashekaran u. Vijayalksmi (2004)] gefiltert.
- Lee u. a. (2011) schlugen eine Pipeline vor, wobei zuerst die Gene mit einem adaptiven genetischen Algorithmus selektiert wurden.
- In Kumar u. a. (2015) kam ein Ensemble aus ANOVA, Kruskal-Wallis- und Friedman-Test zum Einsatz, um die relevanten Variablen zu bestimmen.
- Ang u. a. (2015) schlugen eine SVM-basierte Methode zur Variablenselektion vor, welche die Proben mit sowohl bekannter als auch unbekannter Klassenzugehörigkeit (semiüberwachtes Lernen) verwendet.

In der vorliegenden Arbeit kommt folgende Methode zum Einsatz: Die Probesets werden ihrer Varianz über alle Proben nach geordnet und die ersten n werden zur Klassifikation genommen, wobei n festzulegen ist. Einerseits verwendet diese Methode keine Information über die Klassenzugehörigkeit, was die Gefahr der Überanpassung verringert. Zusätzlich werden die tatsächlich gemessenen Variablen genommen, sodass deren biologische Interpretation möglich ist.

Aus biologischer Sicht ist es ferner ratsam, die Anzahl der Prädiktoren nicht größer als 100 zu wählen. In diesem Falle wäre es möglich, die Expression von entsprechenden Genen unter Verwendung von quantitativem Echtzeit-PCR kostengünstiger zu messen. Somit wird der Suchraum zu Beginn für n auf den Bereich $\{1, 2, \ldots, 100\}$ reduziert. Zur Bestimmung der Anzahl von Probesets, welche für die Analyse verwendet werden, wird für $n = 1, ..., 100$ eine Klassifikationsregel gebildet, welche n Probesets enthält. Für jede Probe in der Testmenge wird die Wahrscheinlichkeit der Zugehörigkeiten zu den beiden Klassen bestimmt. In die Testmenge werden dabei alle Proben entweder einer Substanz

(Leave-one-), zwei (Leave-two-) oder drei (Leave-three-out) verschiedener Substanzen aufgenommen. Für die Beurteilung der Klassifikationsgüte wird der AUC-Wert (zur kurzen Erläuterung siehe Anhang auf Seite 100) genommen. Als stetiger Prädiktor wird dabei die Wahrscheinlichkeit der Zugehörigkeit zu der Klasse der HDAC-Inhibitoren verwendet und für jedes n die AUC-Maßzahl bestimmt. Zu einer visuellen Einschätzung der Abhängigkeit werden die AUC-Werte gegen die Anzahl verwendeter Probesets abgetragen. In Abbildung 20 sieht man den Verlauf im Falle einer Anwendung des SVM-Verfahrens auf die Daten der Klassifikationsstudie UKN1. Um die Verfahren mit 10, 50 oder 100 verwendeten Probesets miteinander zusätzlich graphisch zu vergleichen, werden ROC-Kurven erstellt, siehe Abbildung 21 auf Seite 57 (zu einer kurzen Einführung in die Theorie der ROC-Kurven sei der Leser auf Unterkapitel 6.1.2 im Anhang verwiesen). Dabei werden die Wahrscheinlichkeiten der HDACi-Klasse als ein stetiger prognostischer Faktor interpretiert, welcher unter Verwendung einer Schranke zur Klasseneinteilung hinzugezogen wird. Dabei kann diese Schranke von 0.5 unterschiedlich sein, sogar alle möglichen Werte im Bereich $(0, 1)$ annehmen.

Zu einer weiteren Präsentation von Klassifikationsergebnissen werden in einer Tabelle die Klassen-Wahrscheinlichkeiten aller Substanzen gezeigt. Die Werte stellen dabei die über alle Proben gemittelten Wahrscheinlichkeiten der Klasse für die Substanz dar, welche in der jeweiligen Zeile steht. Somit bedeutet ein Wert größer oder gleich 0.5, dass die entsprechende Proben im Mittel richtig klassifiziert werden. Der Spaltenname zeigt die Substanz an, welche in die Testmenge zusammen mit der Substanz in der Zeile aufgenommen wird. Somit sind die Ergebnisse der Leave-one-out Validierung auf der Hauptdiagonale der Tabelle zu finden.

Bei der Planung der Klassifikationsstudien werden die Wissenschaftler mit der Frage konfrontiert, wie viele Substanzen man in die Studie aufnimmt und wie viele technische Replikate man verwendet. Steht zu Beginn einer Studie die gesamte Anzahl der Proben fest, sind die beiden Zahlen negativ miteinander korreliert. Hierbei ist es aus biologischer Sicht erkenntnisförderlicher, die Unterschiede zwischen den biologischen Replikaten (in unserem Falle Substanzen) unter die Lupe zu nehmen und somit deren Anzahl zu erhöhen, was wiederum eine Reduzierung der Anzahl technischer Replikate nach sich ziehen würde. Um den Einfluss der Anzahl (oder genauer derer Reduzierung) verwendeter Replikate zu untersuchen, wird folgende Analyse durchgeführt: Es werden nicht alle technischen Replikate der jeweiligen Substanz für die Trainingsmenge verwendet, sondern vier, drei oder sogar nur zwei. Im Detail heißt es, dass für $n = 2, 3, 4$ jeweils n technische Replikate einer Trainingssubstanz gezogen und zum Trainieren verwendet werden. Falls 11 Substan-

zen zum Bilden des Klassifikators zur Verfügung stünden, für welche jeweils 5 Replikate existieren, so gäbe es für $n = 4$ insgesamt

$$\binom{5}{4}^{11} = 48828125$$

Möglichkeiten, eine reduzierte Trainingsmenge zu bilden. In der vorliegenden Arbeit werden nicht alle solche Möglichkeiten untersucht, sondern 100 zufällig gezogene. Das jeweilige Verfahren wird dann auf der reduzierten Trainingsmenge gelernt und zum Diskriminieren der Testmenge verwendet. Man erhält somit für jedes Replikat einer Testsubstanz 100 vorhergesagte Klassen-Wahrscheinlichkeiten. Zu deren deskriptiven Untersuchung wird eine Graphik verwendet, wie sie auf Seite 63 abgebildet ist. Die Stichprobenverteilung des jeweiligen Replikats wird durch einen Boxplot dargestellt. Auf der y-Achse ist die Wahrscheinlichkeit der Klasse der HDACi abgetragen. Auf der x-Achse ist die Anzahl verwendeter Replikate angezeigt. Auf eine weitere Analyse der Klassen-Wahrscheinlichkeiten wird verzichtet.

Im vorletzten Schritt wird die in dem Unterkapitel 3.5.3 vorgestellte Sensitivitätsanalyse durchgeführt. Dazu werden für die Rauschstärke (Amplifier) ρ die Werte 1, 2 und 3 genommen. Dieser Parameter reguliert das Ausmaß des Rauschens, indem er multiplikativ in den Summand eingeht, welcher auf die Expression aufaddiert wird. Der zweite Parameter $t \in [0, 1]$ wird als Noise bezeichnet und gibt den Anteil der verrauschten Variablen an. Für diese Analysen werden nacheinander 0, 50 und 100 % der k vorausgewählten Probesets perturbiert. Der Wert 0 bedeutet hierbei, dass kein Rauschen hinzugefügt wird, der Wert $t = 1$ impliziert, dass alle Probesets verrauscht werden. Der dritte Parameter k gibt an, wie viele Probesets vorausgewählt werden. Für die vorliegenden Analysen wird standardmäßig der Wert $k = 100$ genommen.

Da das hinzugefügte Rauschen stochastisch ist, wird das Verfahren für alle Parameterwerte von ρ und t 100 Mal wiederholt. Als Ergebnis erhält man für jedes Replikat der Testsubstanz 100 Klassen-Wahrscheinlichkeiten. Sie werden zu einer deskriptiven Analyse in einer Abbildung dargestellt, welche exemplarisch auf Seite 64 präsentiert wird. Auf der y-Achse ist die Wahrscheinlichkeit der Klasse HDACi angezeigt, auf der x-Achse des jeweiligen Bildabschnitts ist der Rauschanteil abgetragen. Für jeden der drei Werte der Rauschstärke wird ein anderer Bildabschnitt verwendet. Der Wert des Amplifiers steht dabei im oberen Bereich. Für die Darstellung der Wahrscheinlichkeiten wird die Darstellungsform des Boxplots verwendet.

Zuletzt geht man der Frage nach, inwiefern die erhaltenen Ergebnisse generalisierbar, d.h auf einen externen Datensatz übertragbar sind. Der Klassifikator, dessen Trainingsmenge der komplette Datensatz einer Studie darstellt, wird zur Klassifikation der Proben aus der VPA Chronic Konzentrationsstudie (vorgestellt in 2.3.2) bzw. VPA Zeitfensterstudie (vorgestellt in 2.3.5) herangezogen. Obwohl alle Proben der beiden Studien mit Val-

proinsäure behandelt wurden und somit einer einzigen Klasse zugehören, erscheint dieses Vorgehen insofern sinnvoll, als es eine Möglichkeit bietet, den Einfluss der Substanzkonzentration bzw. der Einwirkperiode auf die Klassen-Wahrscheinlichkeiten zu untersuchen.

4.1 VPA Chronic Konzentrationsstudie

Als Erstes werden die Ergebnisse der VPA Chronic Konzentrationsstudie präsentiert. Der Hauptkomponentenplot ist in Abbildung 6 dargestellt. Wie in Unterkapitel 3.1 bereits erörtert, eignet sich diese Abbildung sehr gut zur Repräsentierung der Daten: Die ersten beiden Hauptkomponenten erfassen über 95 % der Gesamtvarianz. Zu einer weiteren Visualisierung der Daten ist in Abbildung 7 eine Heatmap wiedergegeben. Die beiden Graphiken lassen auf gute Datenqualität schließen, da die mit gleichen Konzentration behandelten Zellen jeweils einen Cluster bilden. Ferner korreliert die Erhöhung der Konzentration mit der Steigerung der x-Koordinate auf dem Hauptkomponentenplot. Die Behandlung mit 650 µM zeigt allerdings ein abweichendes Verhalten: Während die anderen Proben sich mit steigender Konzentration nahezu perfekt nach rechts bewegen, weichen mit 650 µM inkubierte Zellen stark ab.

Im nächsten Schritt werden mit Hilfe des moderierten t-Tests für jede Konzentration getrennt die differentiell exprimierten Probesets ermittelt. Es lässt sich anhand von Tabelle 9 feststellen, dass die Anzahl sowohl hoch- als auch runterregulierter Probesets mit der steigenden Konzentration zunimmt. Ein Probeset wird dabei als signifikant dereguliert betrachtet, falls der absolute Wert vom logarithmierten Foldchange größer als $\log_2(1.5)$ und der adjustierte p-Wert kleiner als 0.05 war. Die Dosis 650 µM zeigt auch an dieser Stelle ein abweichendes Verhalten.

Konzentration in µM	25	150	350	450	550	650	800	1000
hochregulierten Probesets	0	0	273	554	955	656	1724	2070
runterregulierten Probesets	0	0	33	110	344	118	910	1273

Tabelle 9: Anzahl der differentiell exprimierten Probesets in der VPA Chronic Konzentrationsstudie. Ein Probeset wurde als differentiell exprimiert betrachtet, falls der absolute Wert vom Fold-Change größer als 1.5 und der adjustierte p-Wert kleiner als 0.05 war.

4.2 VPA Acute Konzentrationsstudie

In diesem Abschnitt werden die Ergebnisse der VPA Acute Konzentrationsstudie vorgestellt. Der Hauptkomponentenplot ist in Abbildung 14 dargestellt. Diese Graphik liefert ein starkes Indiz für das Vorhandensein eines Batch-Effektes. Die erste Hauptkomponente, welche knapp 55 % der Gesamtvarianz wiedergibt, erfasst den Unterschied zwischen zwei Gruppen, die sich nach Abfertigungsschüben aufteilen. So befinden sich auf der rechten

Seite der Abbildung zwei Fertigungslose unter den technischen Bezeichnungen **A** und **B**. Innerhalb dieser Gruppe fällt die Ordinate monoton mit steigender Konzentration ab. Somit erfasst die zweite Hauptkomponente, welche ca. 35 % der Gesamtvarianz erklärt, die Unterschiede im Expressionsniveau, deren Analyse das Hauptziel der Studie bildet. Analog verhält es sich mit dem Abfertigungsschub **D12_TW** auf der linken Seite, wobei bereits die Bezeichnung auf die Unterschiede zwischen den Fertigungslosen hinweist. Dieser Sachverhalt legt den Verdacht nahe, dass die erste Hauptkomponente den Batch-Effekt wiedergibt, wobei die zweite die biologisch relevanten Unterschiede erfasst. Die entsprechende Heatmap (siehe Abbildung 43 im Anhang) zeigt einen ähnlichen Sachverhalt.

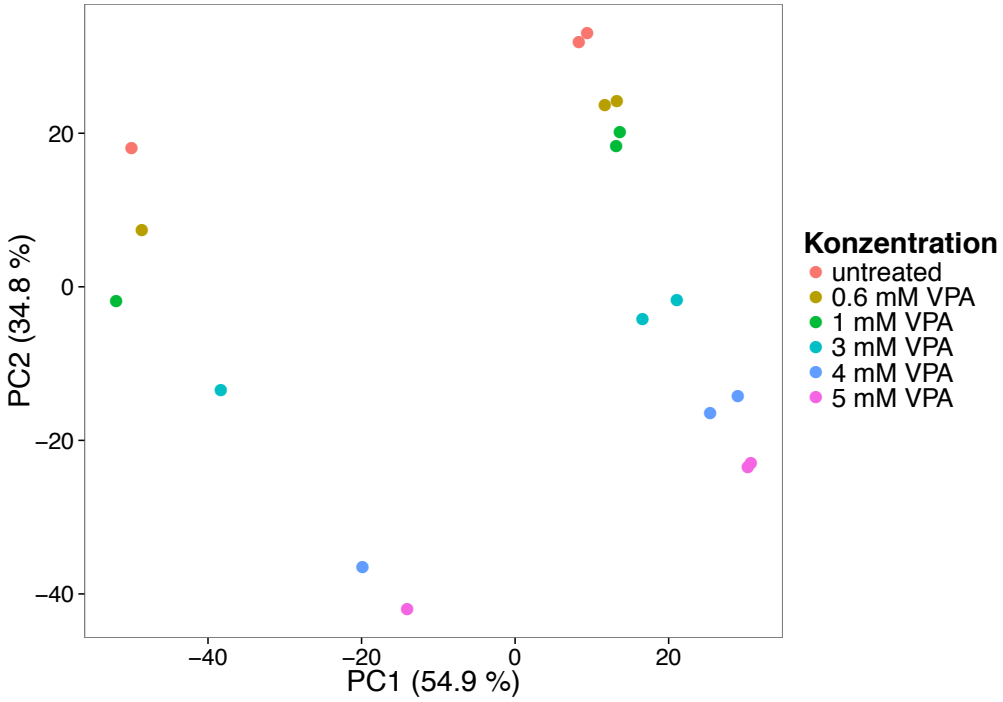

Abbildung 14: Hauptkomponentenplot der VPA Acute Konzentrationsstudie

Als nächster Schritt wurde die Batch-Korrektur mit **ComBat** durchgeführt. Der Hauptkomponentenplot der transformierten Daten ist in Abbildung 15 dargestellt.

Das Bild gibt nun das Verhalten wieder, welches bei der VPA Chronic Konzentrationsstudie (man möge die Hauptkomponentengraphiken vergleichen, siehe Abbildung 6) beobachtet wurde. Jetzt werden über 80 % der Gesamtvarianz durch die erste Hauptkomponente erklärt, und die Zunahme der Konzentration geht mit der Steigerung der Abszisse einher: Auf der linken Seite befinden sich die unbehandelten Proben und auf der rechten die Proben, welche der maximalen Dosis von 5 mM ausgesetzt wurden. Dabei gruppieren sich diejenigen Beobachtungen zusammen, welche mit der gleichen Konzentration der Valproinsäure behandelt wurden. Auch die Heatmap gibt diesen Sachverhalt wieder (siehe Abbildung 44 im Anhang).

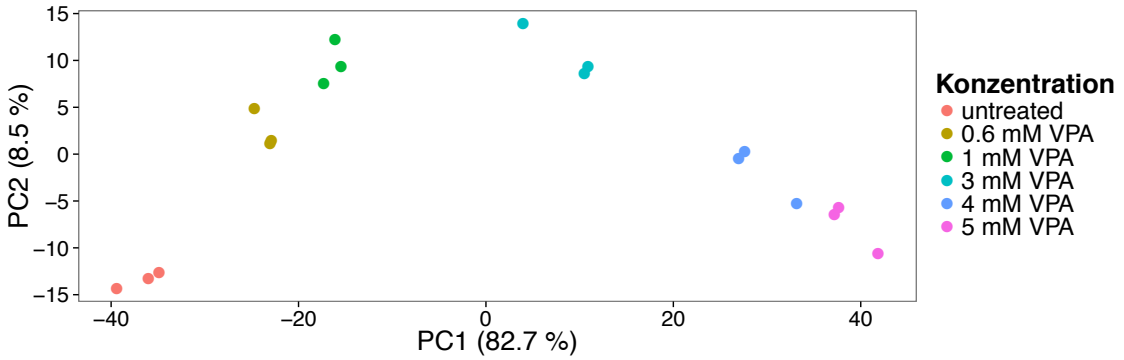

Abbildung 15: Hauptkomponentenplot der VPA Acute Konzentrationsstudie nach der Transformation mit ComBat

Im nächsten Schritt wurden analog dem in 4.1 vorgestellten Vorgehen mit Hilfe des moderierten t-Tests für jede Konzentration getrennt die differentiell expremierten Probesets ermittelt. Es lässt sich anhand von Tabelle 10 feststellen, dass die Anzahl sowohl hoch- als auch runterregulierter Probesets mit der steigenden Konzentration zunimmt. Ein Probeset wurde dabei als signifikant dereguliert betrachtet, falls der absolute Wert vom logarithmierten Foldchange größer als $\log_2(1.5)$ und der adjustierte p-Wert kleiner als 0.05 war.

Konzentration in mM	0.6	1	3	4	5
hochregulierten Probesets	679	1589	3150	3911	5019
runterregulierten Probesets	157	566	2695	4418	4738

Tabelle 10: Anzahl der differentiell exprimierten Probesets in der VPA Acute Konzentrationsstudie. Ein Probeset wurde als differentiell exprimiert betrachtet, falls der absolute Wert vom Fold-Change größer als 1.5 und der adjustierte p-Wert kleiner als 0.05 war. Man stellt fest, dass sich mit zunehmender Konzentration die Anzahl deregulierter Probesets vergrößert.

4.3 MeHg Chronic Konzentrationsstudie

In diesem Abschnitt werden die Ergebnisse der chronischen MeHg Konzentrationsstudie erläutert. Der Hauptkomponentenplot ist in Abbildung 16 dargestellt.

Abbildung 16 legt nahe, dass ein starker Batch-Effekt oder ein anderer Artefakt vorliegt. Als Erstes stellt man fest, dass die Proben sich in drei Gruppen unterteilen lassen. Jede Gruppe enthält dabei genau ein Replikat von jeder Konzentration. Während die beiden Gruppen am linken Rand ein insofern konsistentes Verhalten zeigen, als die Steigung der Konzentration jeweils mit der Steigung der y-Koordinate einhergeht, streuen die Proben aus der dritten Gruppe ohne ein erkennbares Muster. Ferner sind die Proben der dritten Gruppe für den größten Anteil (über 70%) der Gesamtvarianz verantwortlich. Die

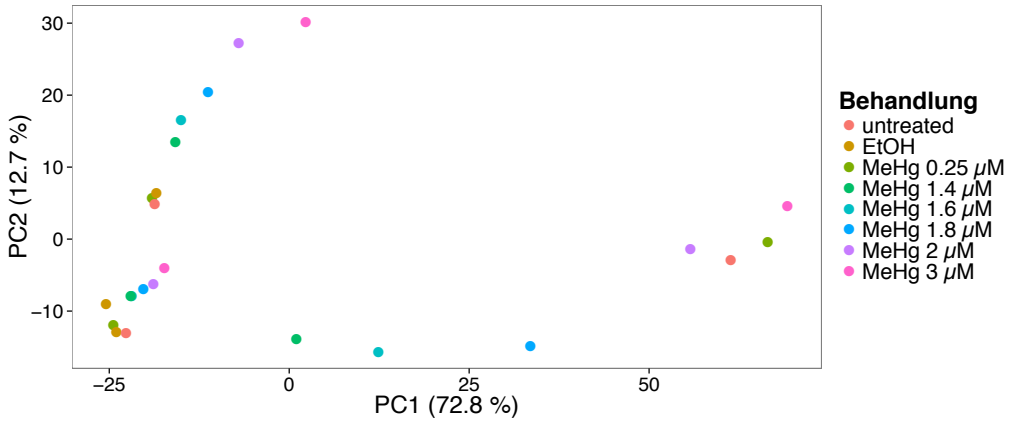

Abbildung 16: Hauptkomponentenplot der MeHg Chronic Konzentrationsstudie

Tatsache, dass die Steigerung der Konzentration einen derart kleinen Anteil der Gesamtvarianz erklärt, bekräftigt die Vermutung, das ein erntszunehmender Artefakt vorliegt. Dies führte zu dem Entschluss auf weitere Analysenschritte zu verzichten.

4.4 MeHg Acute Konzentrationsstudie

In diesem Abschnitt werden die Ergebnisse der akuten MeHg Konzentrationsstudie vorgestellt. Der Hauptkomponentenplot ist in Abbildung 17 dargestellt.

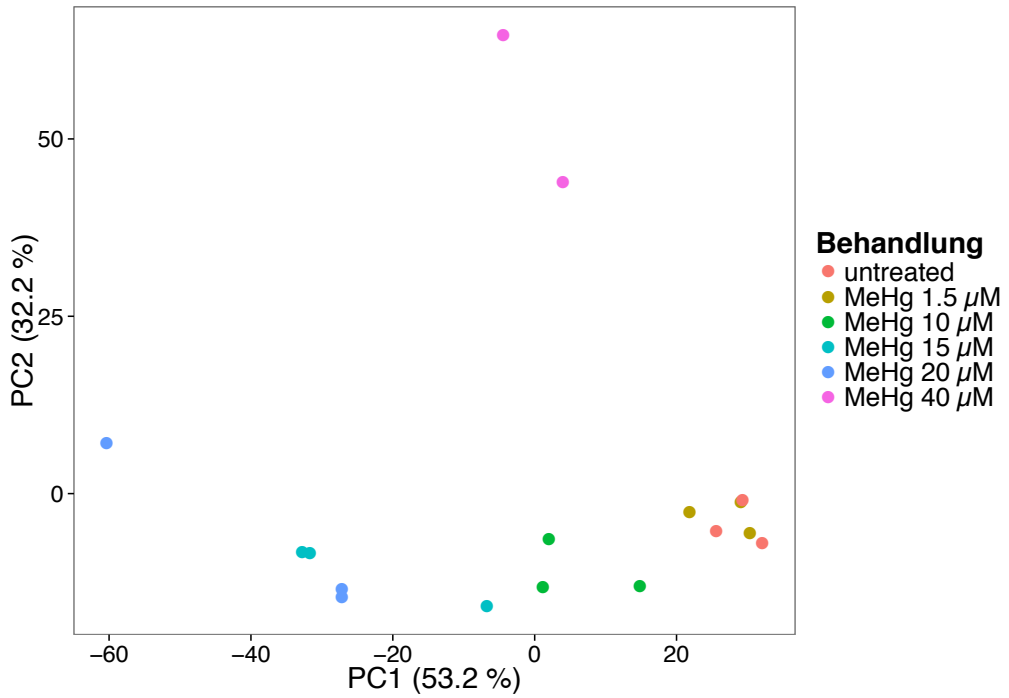

Abbildung 17: Hauptkomponentenplot der MeHg Acute Konzentrationsstudie

Man kann zuerst feststellen, dass im Gegensatz zu der akuten MeHg Konzentrationsstudie (analysiert in Unterkapitel 4.4) weder Batch-Effekt noch andere Artefakte sichtbar sind. Proben, welche mit der gleichen Konzentration behandelt wurden, weisen bis auf wenige Ausnahmen kleine Abstände zueinander auf. Ferner geht die Zunahme der Konzentration mit einer kontinuierlichen Abnahme der ersten Hauptkomponente einher. Die mit der höchsten Konzentration behandelte Proben zeigen allerdings ein abweichendes Verhalten.

Im nächsten Schritt werden analog dem in 4.1 vorgestellten Vorgehen mit Hilfe des moderierten t-Tests für jede Konzentration getrennt die differentiell expremierten Probesets ermittelt. Ein Probeset wurde dabei als signifikant dereguliert betrachtet, falls der absolute Wert vom logarithmierten Foldchange größer als $\log_2(1.5)$ und der adjustierte p-Wert kleiner als 0.05 war. In Tabelle 11 sind die entsprechenden Zahlen angegeben.

Konzentration in µM	1.5	10	15	20	40
hochregulierten Probesets	0	162	2284	1657	598
runterregulierten Probesets	0	206	3123	2412	1386

Tabelle 11: Anzahl der differentiell exprimierten Probesets in der MeHg Acute Konzentrationsstudie. Ein Probeset wurde als differentiell exprimiert betrachtet, falls der absolute Wert vom Fold-Change größer als 1.5 und der adjustierte p-Wert kleiner als 0.05 war.

4.5 Klassifikationsstudie UKN1

In diesem Unterkapitel werden die Analyseergebnisse der UKN1 Klassifikationsstudie dargestellt und erörtert. Dabei werden im ersten Schritt die Daten deskriptiv mit Hilfe der Hauptkomponentenanalyse und Heatmap-Graphiken analysiert. Es wurden sowohl alle 85 Proben deskriptiv untersucht als auch die Teilmenge von 50 behandelten Zelllinien, wobei die Kontrollen abgezogen werden. Diese transformierte Teilmenge wird im späteren Verlauf der Arbeit als kontrollbereinigte Daten umschrieben.

Wie bereits in Unterkapitel 3.4.1 besprochen, weisen die Ergebnisse bei allen 85 Beobachtungen auf einen Batch-Effekt hin (vgl. Abbildung 8 auf Seite 31 und Abbildung 45 im Anhang auf Seite 121). Demgegenüber lässt sich ein solcher Effekt bei den kontrollbereinigten Daten nur bedingt feststellen (siehe Hauptkomponentenplot in Abbildung 18, auch Abbildung 1C in Rempel u. a. (2015)). Es fällt auf, dass die quecksilberhaltigen Substanzen zwei Gruppen bilden: Links unten befinden sich die zu einem früheren Zeitpunkt hydridisierten unverblindeten Proben (MeHg, Thimerosal und $HgCl_2$) und mittig am oberen Rand sind die verblindeten Proben (PMA, PCMB und $HgBr_2$) zu finden. Die Tatsache, dass Microarray-Analysen dieser Gruppen zu verschiedenen Zeitpunkten stattfanden, lässt vermuten, dass diese Differenz sich eventuell nicht auf eine unterschiedliche phänotypische Wirkung zurückführen lässt, sondern auf einen Batch-Effekt. Diese Vermutung

wird durch die deutliche Steigerung der Menge der differentiell exprimierten Probesets bei verblindeten im Vergleich zu unverblindeten Mercurials erhärtet. Demgegenüber lässt sich ein ähnlicher Effekt bei HDACi nicht beobachten. Die Beobachtungen streuen am unteren Rand der Abbildung und lassen die Aufteilung in die Abfertigungsschübe nicht erkennen.

Man kann ferner feststellen, dass sich die beiden Klassen linear trennen lassen. Vorausgreifend lässt sich vermuten, dass die lineare SVM als ein geometrisch motiviertes Verfahren an diesem Datensatz gute Leistungen zeigt.

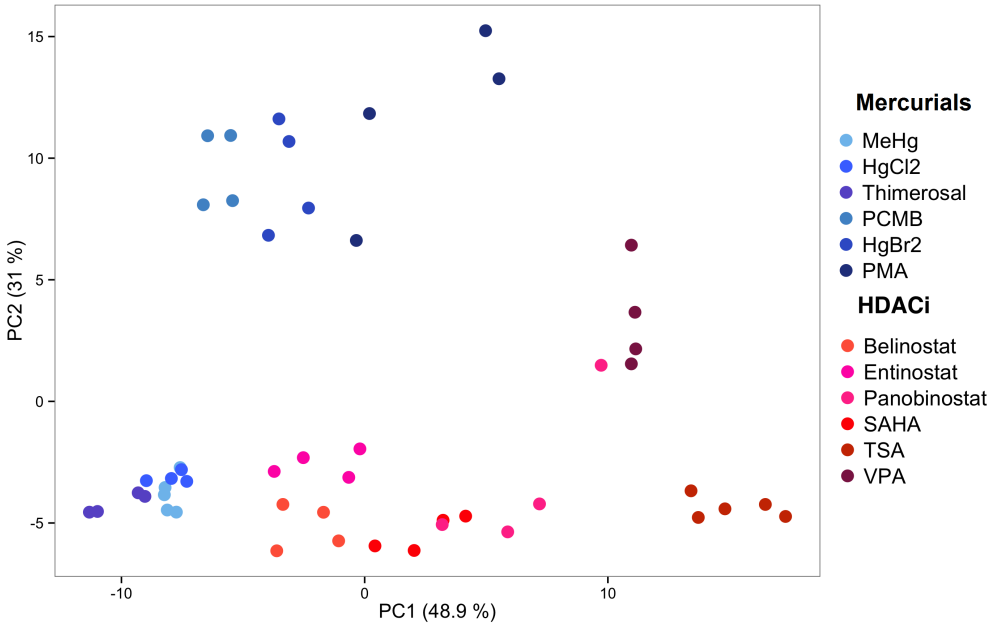

Abbildung 18: Darstellung der ersten beiden Hauptkomponenten der HKA für die UKN1 Klassifikationsstudie nach dem Abzug der entsprechenden Kontrollen. Die verschiedenen Substanzen sind durch verschiedenen Abstufungen der roten (im Falle von HDAC-Inhibitoren) und blauen (im Falle von quecksilberhaltigen Substanzen) Farbe gekennzeichnet. Die beiden Klassen sind voneinander durch eine Gerade trennbar.

Das Ergebnis der Cluster-Analyse wird in Abbildung 19 (siehe auch Abbildung 2A in Rempel u. a. (2015)) präsentiert. Das Dendrogramm am oberen Rand gibt die hierarchische Aufteilung der Proben wieder. Die farbige Leiste unterhalb des Dendrogramms zeigt sowohl die einzelnen Substanzen als auch deren Klassen an. Visuell lassen sich die Beobachtungen in 3 Gruppen aufteilen:

- Verblindete quecksilberhaltige Substanzen (PMA, PCMB und $HgBr_2$) bilden einen eigenen Cluster. Im Dendrogramm ist er durch Zweige am linken Rand repräsentiert.
- Zellen, welche mit TSA, VPA und in einem Falle mit Panobinostat behandelt wurden, gruppieren sich am rechten Rand des Dendrogramms zusammen.

- Die restlichen Beobachtungen bilden einen Cluster in der Mitte des Dendrogramms.

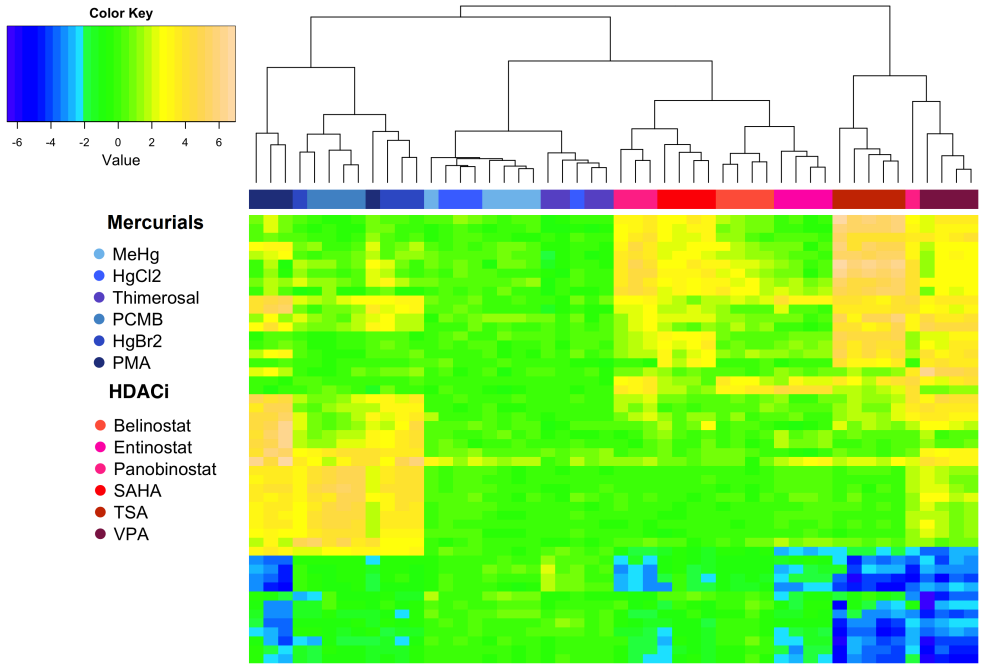

Abbildung 19: Heatmap der kontrollbereinigten Daten der UKN1 Klassifikationsstudie. Die farbigen Leisten unterhalb des Dendrogramms symbolisieren die Aufteilung der Proben nach Zugehörigkeiten zu verschiedenen Behandlungen.

Somit wird bei den kontrollbereinigten Daten auf eine zusätzliche Adjustierung des Batch-Effektes verzichtet. Eine ComBat-Transformation bei den ursprünglichen Daten scheint nur für repräsentative Zwecke nützlich zu sein (vgl. Hauptkomponentenplot der ComBat-transformierten Daten im Anhang in Abbildung 46).

4.5.1 Analyse des SVM Verfahrens

In diesem Abschnitt werden die Ergebnisse der Lernmethode SVM präsentiert. Als Erstes ist in Abbildung 20 die Abhängigkeit des AUC-Wertes von der Anzahl der verwendeten Prädiktoren abgebildet.

Im Falle einer Leave-one-out Validierung erkennt man, dass bereits eine kleine Anzahl von Probesets ausreicht, um einen AUC-Wert zu erreichen, welcher größer als 0.95 ist. Ab etwa 20 Probesets ist der AUC-Wert bis auf einige Ausnahmen konstant Eins. Dies bedeutet, dass bereits wenige Prädiktoren genügen, um die beiden Klassen voneinander zu diskriminieren. Diese Beobachtung erscheint angesichts des Hauptkomponentenplots (Abbildung 18) plausibel: Lineare Projektionen der kontrollbereinigten Proben lassen sich linear derart trennen, dass die quecksilberhaltigen Substanzen (blaue Punkte) von den HDAC-Inhibitoren (rote Punkte) separiert sind. Eine mögliche trennende Hyperebene

würde auf eine Gerade projiziert, welche im Hauptkomponentenplot von links unten nach rechts oben verlaufen würde (allerdings nur im Falle, dass der Normalenvektor der Hyperebene sich als lineare Kombination der beiden ersten Hauptkomponenten darstellen lässt).

Ferner erkennt man, dass die Abhängigkeiten im Falle einer Leave-two-out bzw. Leave-three-out Validierung bis zu einer Anzahl von etwa 40 bis auf wenige Ausnahmen monoton ansteigen und dann nahezu konstant bleiben. Betrachtet man die 3 Abhängigkeiten zusammen, so lässt sich kaum bestimmen, welche Anzahl von Probesets am besten geeignet ist. So erscheinen die Güteunterschiede für die Klassifikationsregel mit 50 bzw. 100 Probesets sehr klein, um von einer Überlegenheit der jeweiligen Regel zu sprechen. Um die Verfahren mit 10, 50 oder 100 verwendeten Probesets miteinander zusätzlich graphisch zu vergleichen, wurden die ROC-Kurven erstellt, siehe Abbildung 21.

Die ROC-Kurven in Abbildung 21 bestätigen die visuellen Eindrucke insofern, als dass die beiden Abhängigkeiten im Falle von 50 und 100 verwendeten Prädiktoren höher liegen als die unter Verwendung von 10 Probesets. Dies bedeutet, dass für jeden Wert der

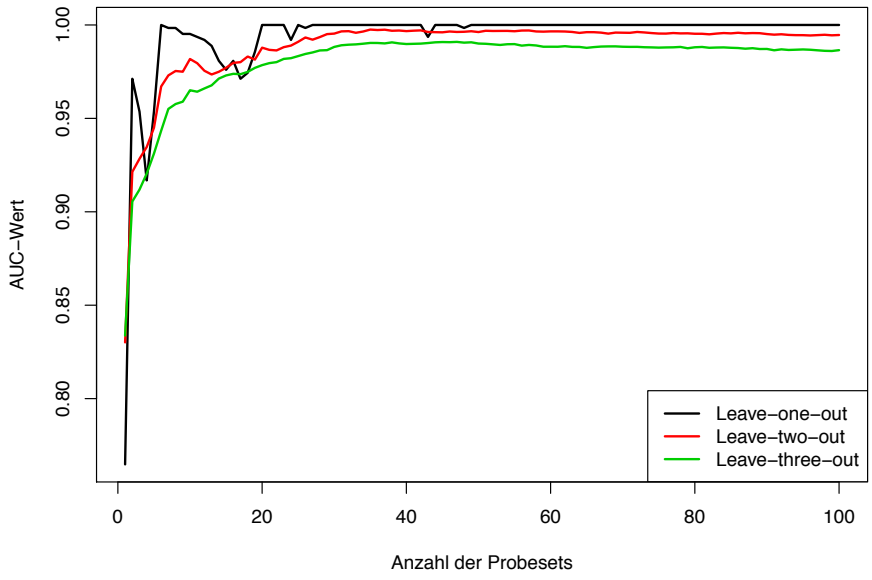

Abbildung 20: AUC-Wert in Abhängigkeit von der Anzahl verwendeter Probesets mit der höchsten Varianz über die Beobachtungen in der jeweiligen Trainingsmenge von UKN1. Zum Diskriminieren wird eine SVM verwendet. Die Proben einer, zwei oder drei verschiedener Substanzen werden jeweils in die Testmenge aufgenommen. Die jeweiligen Vorgehensweisen werden respektive als **Leave-one**, **Leave-two** oder **Leave-three-out** bezeichnet. In der Abbildung werden die entsprechenden Funktionen durch verschiedenen Farben hervorgehoben.

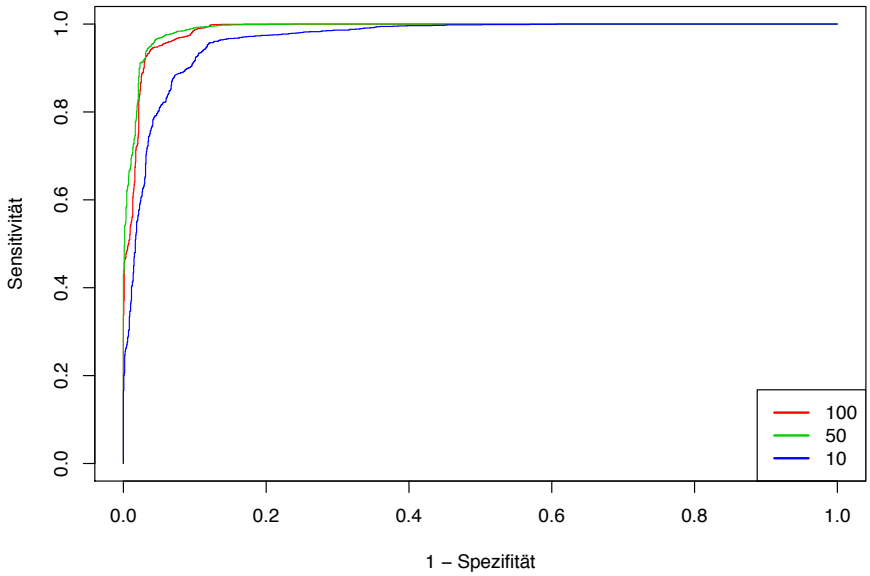

Abbildung 21: ROC-Kurven für das SVM-Verfahren mit 10, 50 und 100 Probesets mit der höchsten Varianz über die Beobachtungen in der Trainingsmenge von UKN1. Auf der x-Achse ist die Falsch-Positiv-Rate abgetragen, auf der y-Achse ist die Richtig-Positiv-Rate abgetragen. Die verschiedenen Varianten des SVM-Lernverfahrens sind durch verschiedene Farben gekennzeichnet.

Falsch-Positiv-Rate die beiden Verfahren eine nichtunterlegene Richtig-Positiv-Rate aufweisen und deshalb vorzuziehen sind. Vergleicht man nun die Verfahren mit 50 bzw. 100 verwendeten Genprodukten miteinander, so sind die Unterschiede derart klein, dass man von einer vergleichbaren Güte sprechen kann. Um sich allerdings auf eine bestimmte Zahl von Probesets zu konzentrieren, wurde entschieden, 100 Genprodukte zu verwenden. Man erhoffte sich, sowohl eine gewisse Robustheit zu gewährleisten als auch eine biologische Interpretierbarkeit zu ermöglichen. Somit wurden die späteren Berechnungen jeweils mit 100 Probesets mit der größten Varianz über die Beobachtungen in der Trainingsmenge durchgeführt.

In Tabelle 12 auf Seite 59 sind nun die Ergebnisse der Leave-one-out und Leave-two-out Analyse ausführlich präsentiert. Es lässt sich an Hand der Hauptdiagonale als Erstes feststellen, dass alle Substanzen in Rahmen einer Leave-one-out Auswertung jeweils der richtigen Klasse zugeordnet werden. In der Leave-two-out Auswertung werden allerdings 3 Substanzen falsch eingeordnet:

1. Bilden Belinostat und Entinostat die Testmenge, werden die beiden Substanzen fälschlicherweise als quecksilberhaltige Substanzen vorhergesagt.

2. Bilden PMA und $HgBr_2$ die Testmenge, so wird PMA fälschlicherweise als HDAC-Inhibitor erkannt.

	Belinostat	Entinostat	$HgBr_2$	$HgCl_2$	MeHg	Panobinostat
Belinostat	0.70	0.45	0.70	0.73	0.73	0.67
Entinostat	0.27	0.79	0.80	0.82	0.81	0.71
$HgBr_2$	0.94	0.96	0.97	0.94	0.95	0.96
$HgCl_2$	0.93	0.90	0.90	0.93	0.84	0.90
MeHg	0.90	0.87	0.85	0.77	0.86	0.87
Panobinostat	0.96	0.98	0.99	0.99	0.99	0.99
PCMB	0.97	0.98	0.91	0.97	0.97	0.98
PMA	0.77	0.71	0.26	0.72	0.66	0.76
SAHA	0.60	0.80	0.85	0.83	0.86	0.86
Thimerosal	0.93	0.92	0.92	0.85	0.89	0.92
TSA	1.00	1.00	1.00	1.00	1.00	1.00
VPA	0.95	0.94	0.95	0.95	0.95	0.95

	PCMB	PMA	SAHA	Thimerosal	TSA	VPA
Belinostat	0.69	0.62	0.62	0.72	0.79	0.62
Entinostat	0.81	0.81	0.78	0.83	0.81	0.78
$HgBr_2$	0.91	0.75	0.96	0.95	0.97	0.97
$HgCl_2$	0.91	0.89	0.92	0.86	0.93	0.91
MeHg	0.84	0.82	0.86	0.84	0.88	0.84
Panobinostat	0.99	0.99	0.98	0.99	1.00	0.99
PCMB	0.98	0.96	0.98	0.98	0.98	0.98
PMA	0.66	0.66	0.67	0.72	0.68	0.67
SAHA	0.83	0.83	0.84	0.80	0.82	0.82
Thimerosal	0.92	0.93	0.91	0.93	0.93	0.94
TSA	1.00	1.00	1.00	1.00	1.00	1.00
VPA	0.95	0.99	0.91	0.96	0.92	0.96

Tabelle 12: Ergebnisse der kreuzvalidierten Auswertung der Vorhersage des SVM-Verfahrens für die UKN1 Klassifikationsstudie. In den Zeilen und Spalten stehen in die Testmenge aufgenommenen Substanzen. Übrigen 10 bzw. 11 Substanzen bilden die Trainingsmenge. Die Werte in der Tabelle sind die gemittelten Wahrscheinlichkeiten der richtigen Klasse zugeordnet zu werden. Durch rote Farbe sind die falsch klassifizierten Substanzen hervorgehoben.

Um zu untersuchen, ob eine Substanz nur dann richtig klassifiziert werden, wenn eine andere bestimmte Substanz in die Trainingsmenge aufgenommen wird, entschied man sich, die Mächtigkeit der Testmenge zu erhöhen. So werden jeweils 3, 4 und schließlich 5 Substanzen nicht für die Konstruktion des Klassifikators verwendet. Für die Präsentation der Ergebnisse von dieser Auswertungsstrategie wird eine Darstellung gewählt, welche am Beispiel von Belinostat (Abbildung 22, siehe auch Abbildung 5C in Rempel u. a. (2015)) erörtert wird.

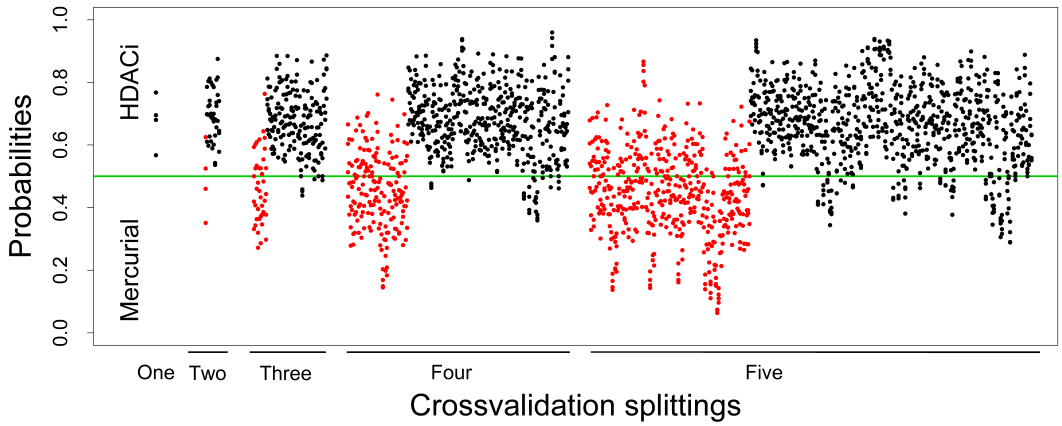

Abbildung 22: Wahrscheinlichkeiten-Graphik für die Vorhersage von Belinostat. Jedem Wert auf der x-Achse entspricht eine Aufteilung in Trainings- und Testmenge. Dabei werden nacheinander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz Belinostat stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von Belinostat, der Klasse der HDAC-Inhibitoren zu zugehören. Für richtige Vorhersagen weist der entsprechende Punkt einen Wert $y \geq 0.5$ auf und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben Belinostat auch Entinostat enthalten, sind durch rote Farbe hervorgehoben.

In Abbildung 22 sind Wahrscheinlichkeiten für eine Vorhersage als HDAC-Inhibitor für Belinostat abgetragen. Die grüne horizontale Linie markiert den Wert von 0.5, so dass die Proben oberhalb dieser Grenze richtig als HDAC-Inhibitore erkannt werden. Alle 4 Replikate werden klassifiziert, so dass 4 Punkte mit der gleichen Abszisse jeweils einer Aufteilung in Test- und Trainingsmenge entsprechen. Die Beschriftungen auf der x-Achse markieren Ergebnisse mit der Mächtigkeit der Testmenge gleich 1, 2, 3, 4 und 5. Die rot hervorgehobenen Punkte stehen für Ergebnisse der Diskriminierung, bei welcher Entinostat neben Belinostat mit in die Testmenge aufgenommen wird. Es ist deutlich zu erkennen, dass sobald Entinostat nicht bei der Konstruktion des Klassifikators verwendet wird, Belinostat vermehrt nicht als HDAC-Inhibitor vorhergesagt wird. Somit setzt sich die Tendenz aus Tabelle 22, dass für eine richtige Klassifizierung von Belinostat die Trai-

ningsmenge Entinostat enthalten muss, fort. Auf Grund des Wahrscheinlichkeiten-Plots für Entinostat stellt sich heraus, dass diese Beziehung wechselseitig ist: Sobald Belinostat nicht für das Trainieren des Klassifikators benutzt wird, sinkt die Erkennungsrate von Entinostat. Diese Abhängigkeit wird plausibel, wenn man den Hauptkomponenten-Plot der Daten in Abbildung 18 betrachtet: Belinostat und Entinostat sind die Substanzen, welche der Gruppe der unverblindeten quecksilberhaltigen Substanzen im linken unteren Eck am nächsten liegen. Entfernt man die beiden Stoffe aus der Trainingsmenge, so verschiebt sich die trennende Hyperebene der SVM in Richtung der HDAC-Inhibitoren, was eine Klassifizierung von Belinostat und Entinostat als quecksilberhaltige Substanz nach sich zieht.

Die Hauptkomponenten-Graphik liefert auch eine Hilfe zum Verständnis der Klassifizierung von PMA: Einerseits weist diese Substanz den größten Wert der ersten Hauptkomponente von allen quecksilberhaltigen Substanzen auf, andererseits ist $HgBr_2$ die quecksilberhaltige Substanz mit dem kleinsten Abstand zu PMA. Wird nun $HgBr_2$ aus der Trainingsmenge entfernt, wird die trennende Hyperebene näher an quecksilberhaltige Substanzen verlegt, und PMA wird fälschlicherweise als HDAC-Inhibitor erkannt. Dies wird bei der Betrachtung des Wahrscheinlichkeiten-Plots besonders deutlich (Abbildung 23): Sobald $HgBr_2$ mit in die Testmenge aufgenommen wird, was durch rote Farbe gekennzeichnet wird, weisen die Punkte eine viel höhere Wahrscheinlichkeit auf, als HDAC-Inhibitor vorhergesagt zu werden.

Die Wahrscheinlichkeiten-Graphiken für die anderen 11 Substanzen mit ausgewählt hervorgehobenen Stoffen sind im Anhang ab Seite 123 zu finden.

Einfluss der technischen Replikate In diesem Abschnitt werden die Ergebnisse der Analyse des Einflusses der Anzahl aufgenommener technischer Replikate vorgestellt. Wie bereits erörtert, werden dazu die technischen Replikate jeder der zwölf Substanzen vorhergesagt, wobei zum Trainieren jeweils 2, 3 oder 4 technische Replikate der jeweiligen Substanz insgesamt 100 mal gezogen werden. Zum Vergleich wird auch die komplette Trainingsmenge verwendet, wobei das Lernverfahren 100 mal auf derselben gelernt wird. Die Ergebnisse für die Substanz Belinostat sind in Abbildung 24 dargestellt.

Als Erstes lässt sich feststellen, dass die Vorhersagewahrscheinlichkeiten auch bei der Verwendung von allen technischen Replikaten variieren. Dies liegt daran, dass die Optimierung der Parameter randomisiert ist. Ferner sind die Replikate hinsichtlich ihrer vorhergesagten Klassenzugehörigkeit heterogen: Die mittleren Klassenwahrscheinlichkeiten unterscheiden sich um bis zu etwa 20 Prozent. So liegt der Median der Wahrscheinlichkeiten für das 2.Replikat bei ca. 0.8 und für das 3.Replikat bei ca. 0.6. Wird die Trainingsmenge reduziert, so verschlechtert sich die Güte: Der Median der jeweiligen Stichprobe wird kleiner und die Stichprobenverteilung wird an sich breiter. Die Ergebnisse dieser Analyse für die anderen Substanzen sind im Anhang ab Seite 133 präsentiert. Man kann

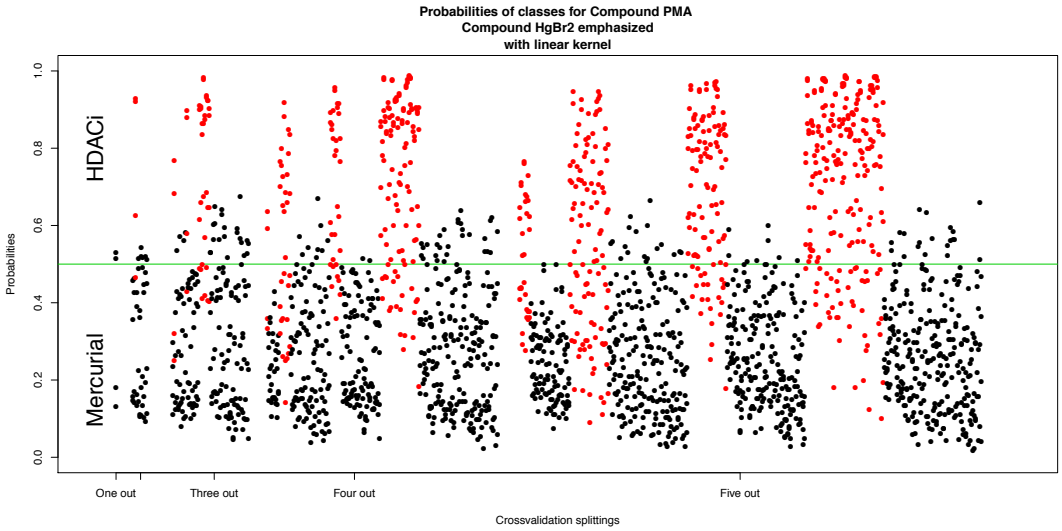

Abbildung 23: Wahrscheinlichkeiten-Graphik für die Vorhersage von PMA. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nacheinander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz PMA stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von PMA der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt einen Wert $y \leq 0.5$ auf und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben Belinostat auch $HgBr_2$ enthielten, sind durch rote Farbe hervorgehoben.

feststellen, dass bis auf die Ausnahme von PMA alle Proben in der Testmenge auch bei der Verwendung von lediglich 2 technischen Replikaten im Mittel richtig klassifiziert werden. Der R-Code ist auf Seite 234 angegeben.

Sensitivitätsanalyse In diesem Abschnitt werden die Ergebnisse der Sensitivitätsanalyse erörtert. Die Ergebnisse für die Substanz PMA sind in Abbildung 25 dargestellt.

Als Erstes fällt auf, dass die Vorhersagewahrscheinlichkeiten für die 4 Proben bei den unverrauschten Daten (Anteil der verrauschten Variablen $t = 0$) sich stark unterscheiden. Während 2 Proben relativ sicher als quecksilberhaltige Substanz erkannt werden, werden die anderen zwei Proben vermehrt als HDAC-Inhibitor klassifiziert. Diese Diskrepanz war bereits auf dem PCA-Plot 18 zu erkennen. Betrachtet man die Änderung von Klassen-Wahrscheinlichkeiten bei Vergrößerung von Rauschstärke und Festhalten von Noiseparameter $t = 0.5$, so stellt man fest, dass diejenigen Proben, welche unverrauscht richtig vorhergesagt werden, weiterhin richtig klassifiziert werden. Allerdings konvergieren die mittleren Vorhersagewahrscheinlichkeiten gegen den Wert von 0.5. Ferner streuen Wahrscheinlichkeiten stärker um den jeweiligen Mittelwert. Diejenigen Proben, welche unverrauscht falsch klassifiziert werden, zeigen ein ähnliches Verhalten. Erhöht man die Rauschstärke im Falle von $t = 1$, d.h. alle Probesets werden perturbiert, so beobachtet

Abbildung 24: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Belinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeten technischen Replikate in der Trainingsmenge angegeben, auf der y-Achse sind die Klassen-Wahrscheinlichkeiten abgetragen.

man bei allen Proben eine Konvergenz der mittleren Klassen-Wahrscheinlichkeiten gegen einen Wert von 0.5, wobei die einzelnen Wahrscheinlichkeiten stärker streuen.

Die Boxplot-Graphiken für die anderen 11 Substanzen sind im Anhang ab Seite 182 zu finden. Der R-Code ist auf Seite 237 angegeben.

4.5.2 Analyse von Random Forests

In diesem Kapitel werden die Ergebnisse der Auswertung von Random Forests auf die Klassifikationsstudie UKN1 präsentiert.

Als Erstes wird die Abhängigkeit des AUC-Wertes von der Anzahl der verwendenten Probesets untersucht. In Abbildung 26 werden die Abhängigkeiten graphisch dargestellt. Der verwendete R-Code ist im Anhang auf Seite 233 angegeben.

Vergleichbar mit Abbildung 20 lässt sich auch hier die optimale Anzahl von Probesets kaum angeben. Bereits bei wenigen Prädiktoren liegt der AUC-Wert oberhalb der Grenze von 0.9. Ihr jeweiliges Maximum erreichen die beiden Strategien bei ca. 40 Genprodukten, danach bleibt die interessierende Maßzahl etwa gleich. Um sich ein genaueres Bild zu verschaffen, werden in Abbildung 27 die ROC-Kurven für die Verfahren mit 10, 50 oder 100 verwendeten Prädiktoren dargestellt.

Die ROC-Kurven in Abbildung 27 spiegeln ein Verhalten wieder, welches dem in Abbildung 21 sehr ähnelt. Die beiden Kurven im Falle von 50 und 100 verwendeter Probesets liegen höher als die unter Verwendung von 10 Probesets. Vergleicht man nun die Verfahren mit 50 bzw. 100 verwendeten Genprodukten miteinander, so sind die Unterschiede sehr klein. Somit verschafft auch diese Analyse wenig Klarheit, ob eine Klassifikationsregel mit einer bestimmten Anzahl von Prädiktoren anderen überlegen ist. Auch in diesem Falle wird entschieden, 100 Genprodukte zu verwenden.

Als nächster Schritt werden in Tabelle 13 auf Seite 67 die Ergebnisse der Leave-one-out und Leave-two-out Kreuzvalidierung dargestellt. Man sieht, dass ähnlich zu SVM-

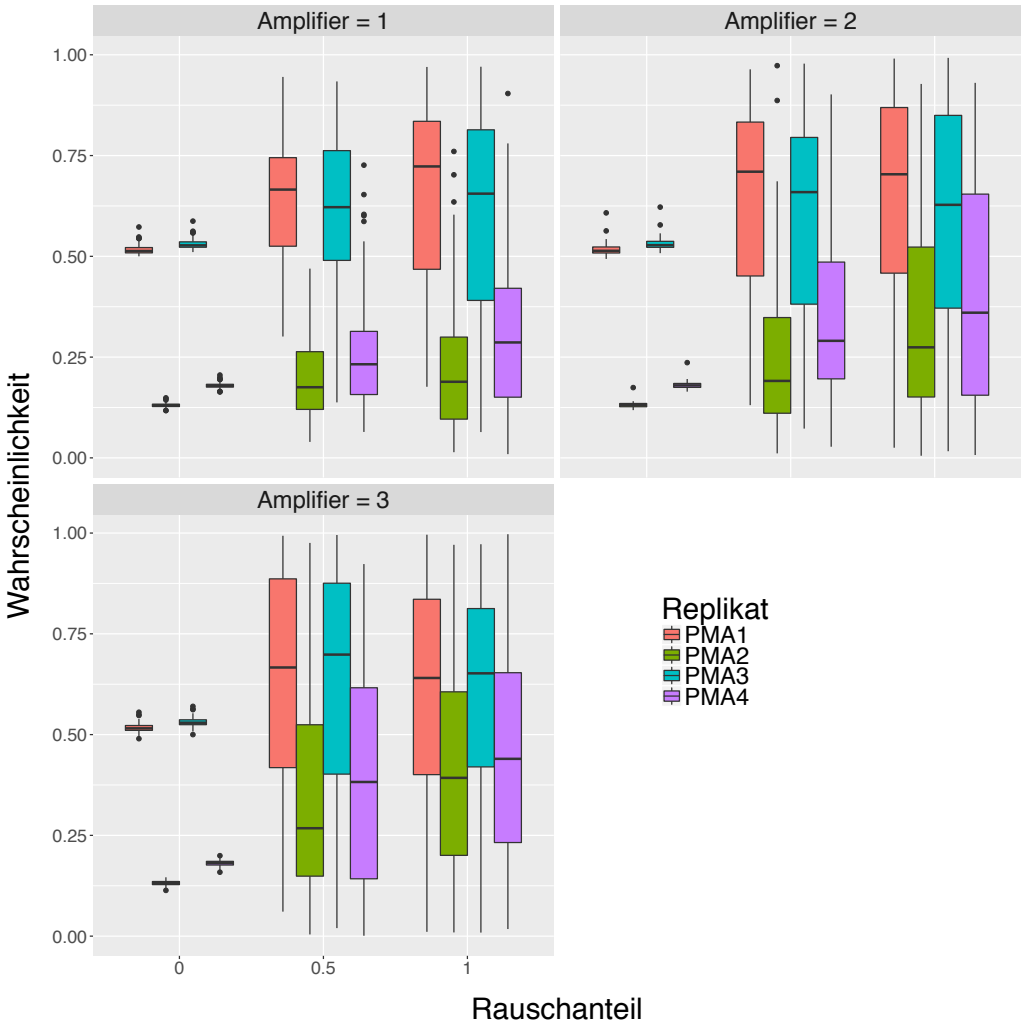

Abbildung 25: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA nach dem Verrauschen der Daten für Rauschenstärke 1 (links), Rauschenstärke 2 (mitte) und Rauschenstärke 3 (rechts). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen angegeben, auf der y-Achse sind die Klassen-Wahrscheinlichkeiten abgetragen.

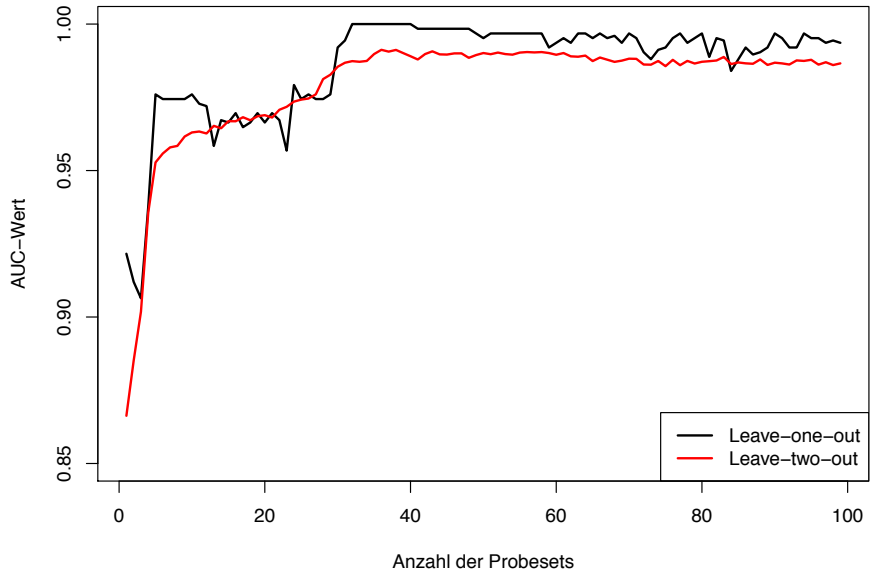

Abbildung 26: AUC-Wert in Abhängigkeit von der Anzahl verwendeter Probesets mit der höchsten Varianz über die Beobachtungen in der jeweiligen Trainingsmenge von UKN1. Zum Diskriminieren wird ein Random Forests verwendet.

Verfahren die Substanzen Belinostat und Entinostat ein Paar bilden: Die jeweilige Substanz wird nur dann richtig vorhergesagt, falls die andere in der Trainingsmenge ist. Die Substanz PMA wird in Abwesenheit von Entinostat, SAHA, PCMB, $HgCl_2$ oder Thimerosal fälschlicherweise als HDAC-Inhibitor erkannt, wobei die Wahrscheinlichkeiten knapp unterhalb der Grenze von 0.5 liegen. An dieser Stelle erscheint es sinnvoll, die ursprünglich ermittelte Wahrscheinlichkeit der Klasse HDACi als einen diagnostischen Faktor zu interpretieren und einen anderen Cutoff als 0.5 zu verwenden. Die ROC-Kurve könnte hier sinnvolle Hinweise liefern.

Einfluss der technischen Replikate In diesem Abschnitt wird der Einfluss der Anzahl aufgenommener technischer Replikate analysiert, und die Ergebnisse werden präsentiert. Die Ergebnisse für die Substanz MeHg sind in Abbildung 28 dargestellt.

Als Erstes lässt sich feststellen, dass die Stichprobenverteilungen der Vorhersagewahrscheinlichkeiten für die einzelne Proben auch bei der Verwendung aller technischer Replikate sich stark unterscheiden. Während die Proben 1, 4 und 5 mit einer großen Sicherheit als quecksilberhaltige Substanz erkannt werden, werden die Proben 2 und 3 nur mit einer mittleren Wahrscheinlichkeit von knapp über 60 Prozent als Mercurial klassifiziert. Wird die Trainingsmenge reduziert, so wird die Varianz grundsätzlich größer. Die jeweiligen Mediane der Stichproben scheinen allerdings bei der Verkleinerung der Trainingsmenge gegen einen Wert zu konvergieren. Ein ähnliches Verhalten zeigt auch die Substanz SA-

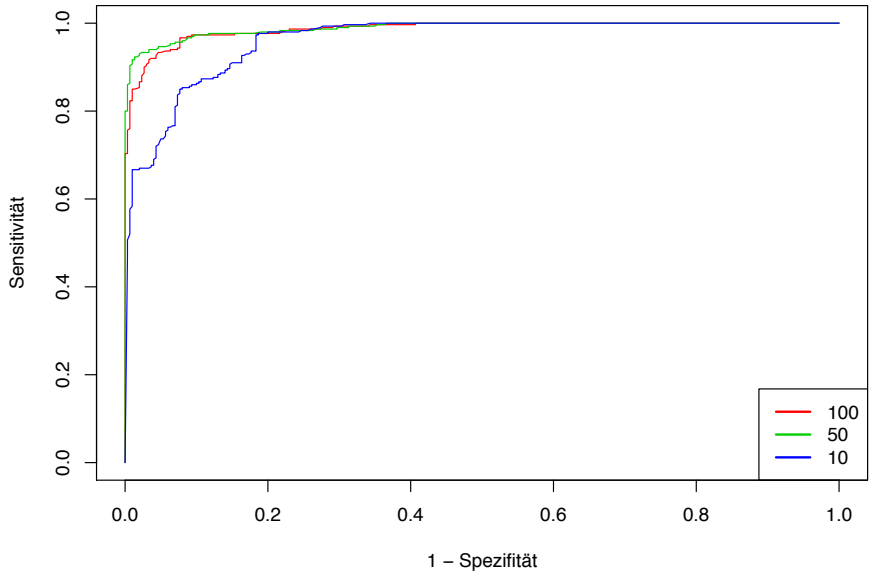

Abbildung 27: ROC-Kurven für das Random Forests Verfahren mit 10, 50 und 100 Probesets mit der höchsten Varianz über die Beobachtungen in der Trainingsmenge von UKN1. Auf der x-Achse ist die Falsch-Positiv-Rate abgetragen, auf der y-Achse ist die Richtig-Positiv-Rate angegeben.

	Belinostat	Entinostat	Panobinostat	SAHA	VPA	TSA
Belinostat	0.65	0.26	0.68	0.56	0.66	0.68
Entinostat	0.25	0.76	0.81	0.82	0.74	0.85
Panobinostat	0.99	0.99	1.00	1.00	1.00	1.00
SAHA	0.83	0.93	0.97	0.95	0.92	0.75
VPA	0.96	0.90	0.94	0.92	0.94	0.91
TSA	0.76	0.92	0.91	0.89	0.99	0.90
$HgBr_2$	0.85	0.81	0.81	0.90	0.94	0.80
PMA	0.54	0.49	0.50	0.45	0.51	0.62
PCMB	0.94	1.00	0.99	0.99	0.99	1.00
$HgCl_2$	0.99	0.99	0.99	0.99	0.99	1.00
MeHg	0.96	0.90	0.88	0.90	0.86	0.82
Thimerosal	1.00	1.00	0.99	1.00	1.00	1.00

	$HgBr_2$	PMA	PCMB	$HgCl_2$	MeHg	Thimerosal
Belinostat	0.75	0.72	0.67	0.70	0.74	0.68
Entinostat	0.76	0.64	0.76	0.86	0.81	0.85
Panobinostat	1.00	0.99	1.00	1.00	1.00	1.00
SAHA	0.96	0.87	0.95	0.94	0.95	0.96
VPA	0.94	0.95	0.94	0.94	0.94	0.95
TSA	0.99	0.98	0.93	0.91	0.92	0.91
$HgBr_2$	0.83	0.82	0.82	0.80	0.84	0.77
PMA	0.51	0.53	0.49	0.49	0.54	0.48
PCMB	0.96	0.94	1.00	1.00	0.98	1.00
$HgCl_2$	0.99	0.98	0.99	0.99	0.93	0.97
MeHg	0.82	0.92	0.90	0.75	0.88	0.78
Thimerosal	0.99	0.96	0.99	0.99	0.96	0.99

Tabelle 13: Ergebnisse der kreuzvalidierten Auswertung der Vorhersage des Random Forest Verfahrens für die UKN1 Klassifikationsstudie. In Zeilen und Spalten stehen in die Testmenge aufgenommenen Substanzen. Übrigen 10 bzw. 11 Substanzen bilden die Trainingsmenge. Somit stehen auf der Hauptdiagonalen von links oben nach rechts unten die Ergebnisse einer leave-one-out Kreuzvalidierung. Außerhalb der Hauptdiagonalen befinden sich die Resultate der leave-two-out Kreuzvalidierung. Die Werte in der Tabelle sind die gemittelten Wahrscheinlichkeiten der richtigen Klasse zugeordnet zu werden. Durch rote Farbe sind die falsch klassifizierten Substanzen hervorgehoben.

Abbildung 28: Boxplot-Graphiken für die HDACi-Vorhersagewahrscheinlichkeiten von MeHg in der UKN1 Klassifikationsstudie. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Zum Klassifizieren wurde das Random Forests verwendet.

HA, deren Analysen zusammen mit den anderen Behandlungen im Anhang ab Seite 144 zu finden sind. Ähnlich zu den Ergebnissen in dem Abschnitt 4.5.1 werden auch bei der Verwendung von Random Forests fast alle Proben in der Testmenge im Mittel richtig vorhergesagt. Der R-Code ist auf Seite 235 angegeben.

Sensitivitätsanalyse In diesem Abschnitt werden die Ergebnisse von Sensitivitätsanalyse präsentiert und erörtert. Die Ergebnisse für die Substanz PMA sind exemplarisch in Abbildung 29 dargestellt.

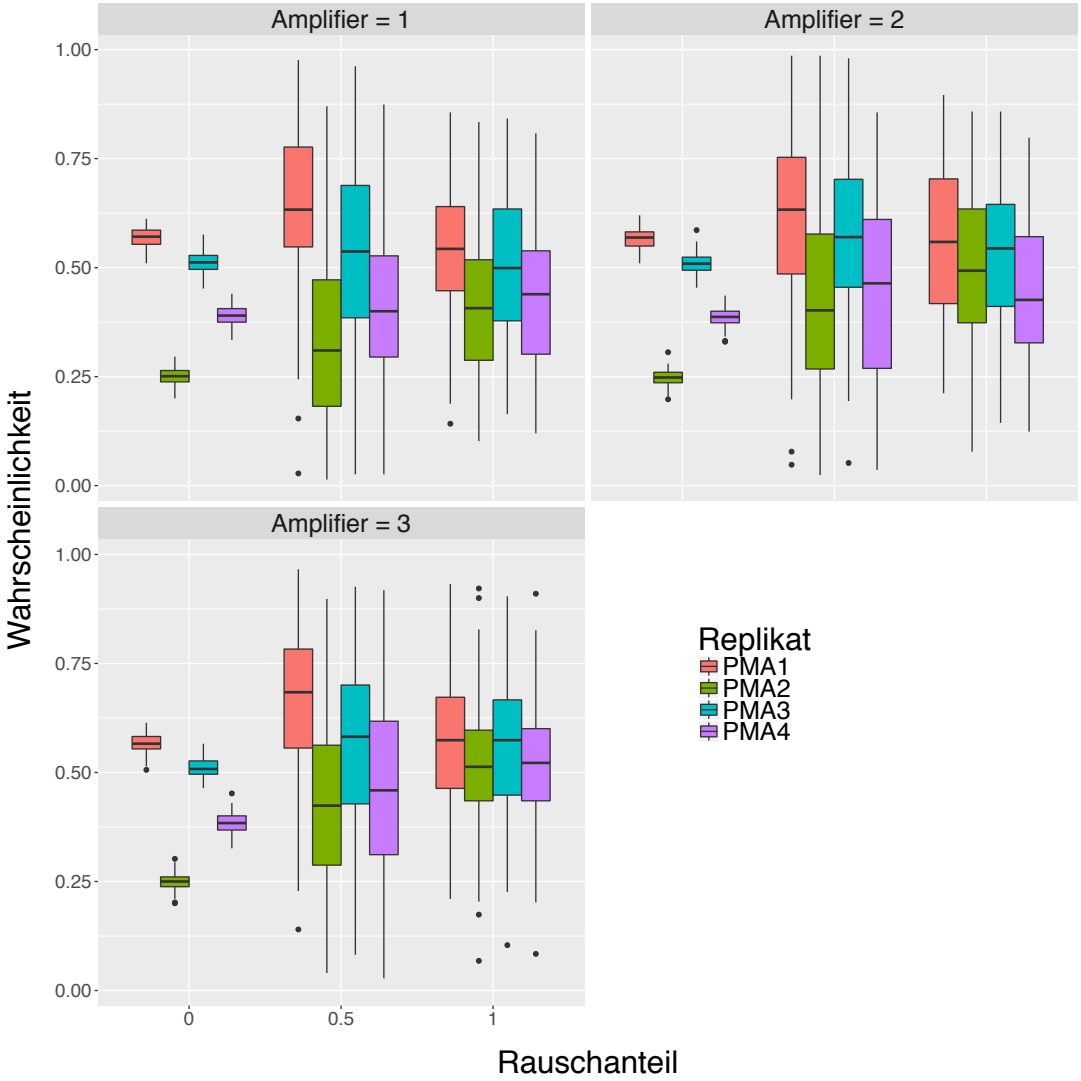

Abbildung 29: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA nach dem Verrauschen der Daten für Rauschstärke 1 (obere Reihe links), Rauschstärke 2 (obere Reihe rechts) und Rauschstärke 3 (untere Reihe links). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen angezeigt, auf der y-Achse sind die Vorhersagewahrscheinlichkeiten abgetragen.

Es fällt auf, dass sich die Ergebnisse von denen in Abbildung 25 teilweise unterscheiden. Auch hier unterscheiden sich die Vorhersagewahrscheinlichkeiten für die 4 Proben bei den unverrauschten Daten sehr stark. Während die Rauschstärke keinen sichtbaren Einfluss zu haben scheint, führt die Erhöhung des Rauschanteils zu einer deutlichen Konvergenz der vorhergesagten Wahrscheinlichkeiten gegen den Wert von 0.5. Die Darstellungen der Analyse für die anderen Substanzen sind im Anhang ab Seite 193 zu finden. Der R-Code ist auf Seite 237 angeben.

4.5.3 Anwendung auf UKN1 VPA-Konzentrationsstudie

Eine verbreitete Methode zur Prüfung der Generalisierungseigenschaft des Lernverfahrens stellt die Anwendung des Klassifikators auf einen externen Datensatz dar. Somit wird sichergestellt, dass die guten Ergebnisse des Verfahrens auf den ursprünglichen Daten nicht auf eine mögliche Überanpassung zurückzuführen sind. In dieser Arbeit wurde entschieden, die Perfomance der Klassifikationsregel auf den Daten der VPA Konzentrationsstudie aus Kapitel 2.3.1 zu validieren. Die gute Qualität des Datensatzes und verschiedene getestete Konzentrationen waren die wichtigsten Gründe für diese Entscheidung. In diesem Abschnitt stellen wir die Ergebnisse dieser Validierung vor.

Zu der ersten deskriptiven Einschätzung wurde der Hauptkomponentenplot der kontrollbereinigten Daten der UKN1 Klassifikationsstudie bestimmt. Dann wurden die kontrollbereinigten Daten der VPA Konzentrationsstudie erstellt: Die Expressionsprofile der Kontrolle wurden paarweise von den Messungen der einzelnen Konzentrationen abgezogen. Mit Hilfe der Rotationsmatrix U aus der Zerlegung (3.1) lassen sich die neuen Datenpunkte in den bestehenden Hauptkomponentenplot hineinprojizieren. Das Ergebnis dieser Berechnung ist in Abbildung 30 (siehe auch Abbildung 6A in Rempel u. a. (2015)) präsentiert.

Diese Abbildung bestätigt erneut die gute Qualität der VPA Konzentrationsstudie: Die mit der kleinsten Konzentration (25 µM) behandelte Proben befinden sich nahe dem Ursprung und haben somit den kleinsten Abstand zu den quecksilberhaltigen Substanzen. Mit steigender Konzentration erhöht sich die erste Koordinate und die Beobachtungen bewegen sich nach rechts. Die mit 650 µM behandelten Proben stellen nicht unerwartet wieder eine Ausnahme dar.

Im nächsten Schritt werden sowohl eine SVM als eine Random Forests Klassifikationsregel gebildet, wobei alle Daten aus der UKN1 Klassifikationsstudie als Trainingsmenge verwendet werden. Die einzelnen kontrollbereinigten Proben werden dann klassifiziert. Die pro einzelne Konzentration gemittelten Vorhersagewahrscheinlichkeiten sind in Tabelle 14 angegeben.

Als Erstes stellt man fest, dass die mit 25 µM und (bei der Verwendung von SVM) 150 µM behandelten Proben zu den Mercurials zugeordnet werden. Dies verwundet nicht insofern als es keine deregulierten Probesets bei diesen Proben identifiziert werden. Dies

Konzentration in µM	25	150	350	450	550	650	800	1000
HDACi-Wahrscheinlichkeit (SVM)	0.1	0.48	0.97	0.99	0.99	0.99	0.99	0.99
HDACi-Wahrscheinlichkeit (RF)	0.08	0.52	0.91	0.96	0.98	0.94	0.96	0.92

Tabelle 14: Mittleren Vorhersagewahrscheinlichkeiten für die Proben aus der VPA Chronic Konzentrationsstudie. Die Klassifikation erfolgte mit SVM und Random Forests, wobei alle Daten der UKN1 Klassifikationsstudie zum Erstellen verwendet wurde.

bedeutet, dass diese Konzentrationen sehr geringe genotypische Wirkung hervorrufen. Demgegenüber verursachen die höheren Konzentrationen von VPA, dass immer mehr Gene dereguliert werden (man beachte Tabelle 9 in Kapitel 4.1). Entsprechende Proben werden auch von den Klassifikationsregeln als HDAC-Inhibitoren mit einer hohen Wahrscheinlichkeit erkannt. Dies bestärkt die Hypothese, dass die HDAC-Inhibitoren eine vergleichbare Wirkung auf das Transkriptom ausüben. Statistische Lernverfahren sind in der Lage ein gemeinsames Muster in der Expressionsprofilen zu erkennen, sodass unbekannte Proben der richtigen Klasse zugeordnet werden können.

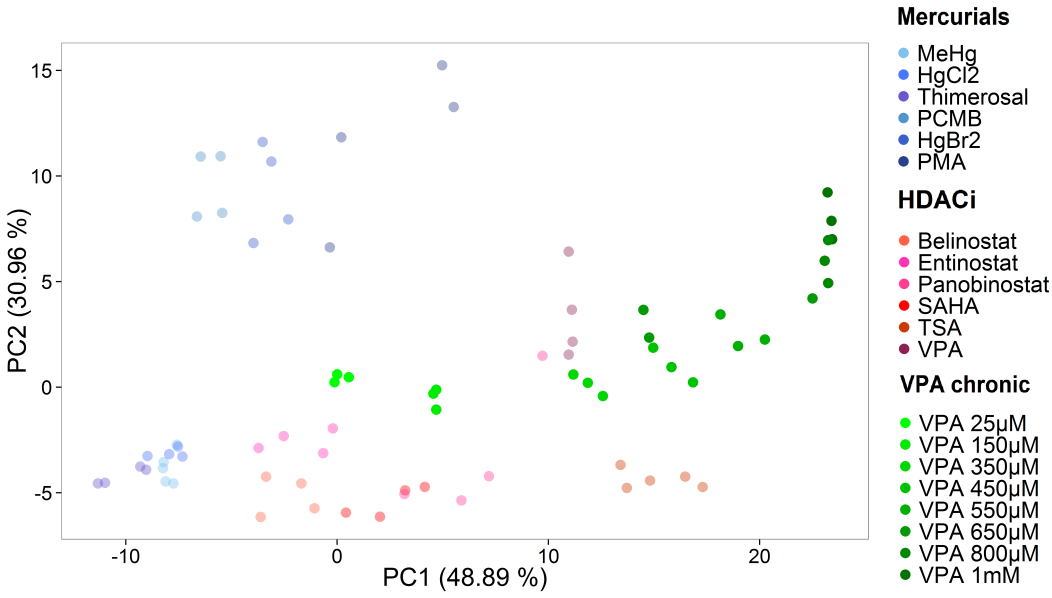

Abbildung 30: Der Hauptkomponentenplot der Daten aus der UKN1 Klassifikationsstudie. Die Daten aus der VPA chronischen Konzentrationsstudie sind mit Hilfe der Rotationsmatrix reinprojiziert und sind durch grüne Farbe gekennzeichnet. Die steigende Konzentrationen sind durch dunklere Farbe hervorgehoben.

4.6 Klassifikationsstudie UKK

In diesem Unterkapitel werden die Analysenergebnisse der UKK Klassifikationsstudie dargestellt und erörtert. Analog zur UKN1 Klassifikationsstudie werden im ersten Schritt die Daten deskriptiv mit Hilfe der Hauptkomponenten- und Heatmap-Graphiken analysiert. Dabei lassen die Ergebnisse bei allen 100 Beobachtungen einen starken Batch-Effekt vermuten, siehe den Hauptkomponentenplot in Abbildung 31 auf Seite 72 und die Heatmap auf Seite 177. Alle Proben auf der linken Seite des Hauptkomponenten-Plots gehören zu dem ersten Batch. Die restlichen Beobachtungen bilden einen Cluster auf der rechten Seite in Abbildung 31. Der Batch-Effekt ist somit für einen großen Anteil der Gesamtvarianz verantwortlich: Die erste Hauptkomponente erklärt über 70% der Gesamtvarianz. Für die späteren Analysen der Daten ist somit eine Reduktion des Batch-Effekts nötig. Ferner ist festzustellen, dass innerhalb der beiden Gruppen keine Cluster vorliegen. Sowohl die zur Kontrolle herangezogene Substanzen, welche in Abbildung durch Abstufungen der grünen Farbe gekennzeichnet sind, als auch quecksilberhaltigen Substanzen (Abstufungen der blauen Farbe) und HDAC-Inhibitoren (Abstufungen der roten Farbe) streuen über den jeweiligen Cluster ohne visuell erkennbare Gesetzmäßigkeit.

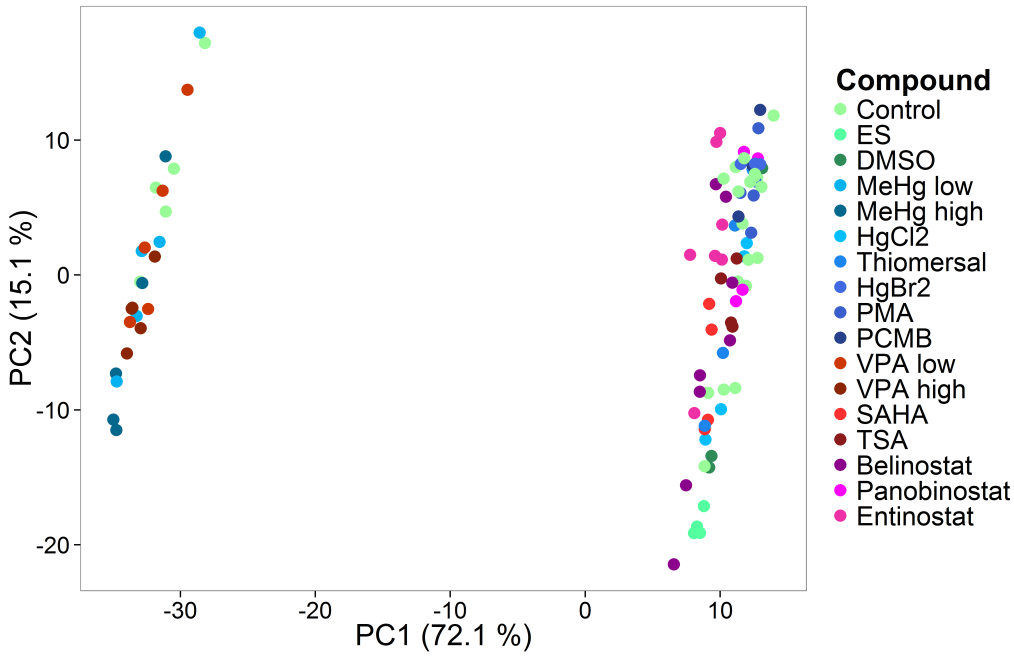

Abbildung 31: Darstellung der ersten beiden Hauptkomponenten der HKA der UKK Klassifikationsstudie. Für die Graphik wurden 100 Probesets mit der größten Varianz über alle Proben verwendet. Ein starker Batch-Effekt wird durch die erste Hauptkomponente wiedergegeben. Die verschiedenen Substanzen sind durch verschiedene Abstufungen der roten (im Falle von HDAC-Inhibitoren), blauen (im Falle von quecksilberhaltigen Substanzen) und grünen (im Falle von Kontrollen) Farbe gekennzeichnet.

Zieht man die entsprechenden Kontrollen ab und entfernt die nichtdifferenzierten Stammzellen (ESC), erhält man den Hauptkomponentenplot in Abbildung 32. Die erste Hauptkomponente erklärt dabei ca. 52% der Varianz und die zweite knapp 17% der Varianz. Die quicksilberhaltigen Substanzen befinden sich dabei im unteren Bereich der Abbildung, während die meisten HDAC-Inhibitoren im oberen Abschnitt zu finden sind. Zieht man eine imaginäre Diagonale von links unten nach rechts oben, würde sie die beiden Klassen fast perfekt trennen. Eine Ausnahme bildet allerdings die Substanz Panobinostat, welche sich von den anderen Repräsentanten der Klasse stark absetzt. Der Hauptkomponentenplot ändert sich unwesentlich beim Verwenden von Combat (siehe Abbildung 102 im Anhang auf Seite 178). Somit wird analog zu der UKN1 Klassifikationsstudie bei den kontrollbereinigten Daten auf eine zusätzliche Adjustierung des Batch-Effektes verzichtet. Die entsprechende Heatmap ist im Anhang auf Seite 179 zu finden (Abbildung 103).

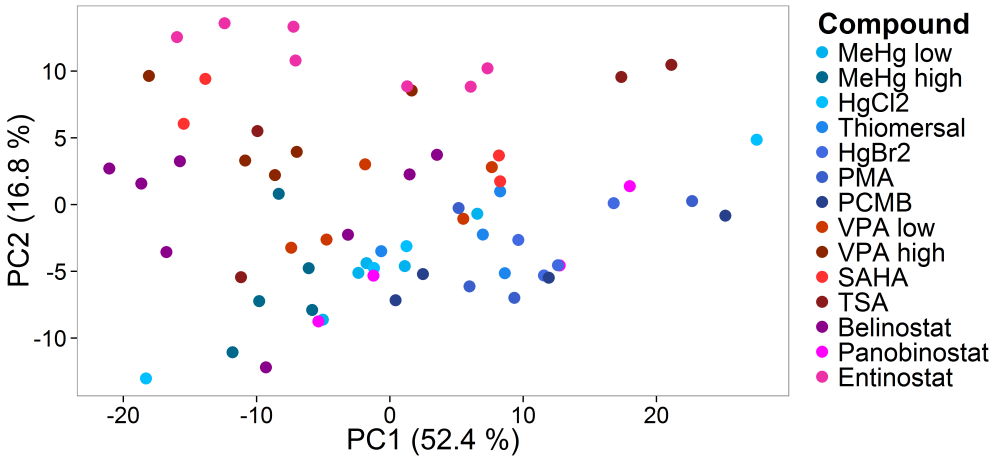

Abbildung 32: Hauptkomponentenplot der UKK Klassifikationsstudie nach Abzug entsprechender Kontrollen von den behandelten Proben. Bei dieser Analyse wurden 100 Probesets mit der höchsten Varianz verwendet. Analog zu Abbildung 18 sind die quecksilberhaltigen Substanzen durch Abstufungen der blauen Farbe und die HDAC-Inhibitoren durch Abstufungen der roten Farbe gekennzeichnet.

4.6.1 Analyse des SVM-Verfahrens

In diesem Unterkapitel werden die Ergebnisse der Anwendung von SVM auf die Klassifikationsstudie UKK vorgestellt. Im ersten Schritt wird die Abhängigkeit des AUC-Wertes von der Anzahl der verwendeten Probeset untersucht. Die entsprechenden Kurven sind in Abbildung 33 dargestellt. Anders als bei der UKN1 Studie werden hier der AUC-Wert nicht nur für $n = 1, \ldots, 100$ Probesets bestimmt, sondern auch für $n = 150, 200, 300, 400$. Diese Werte werden aufsteigend und nicht maßstabsgetreu aufgetragen.

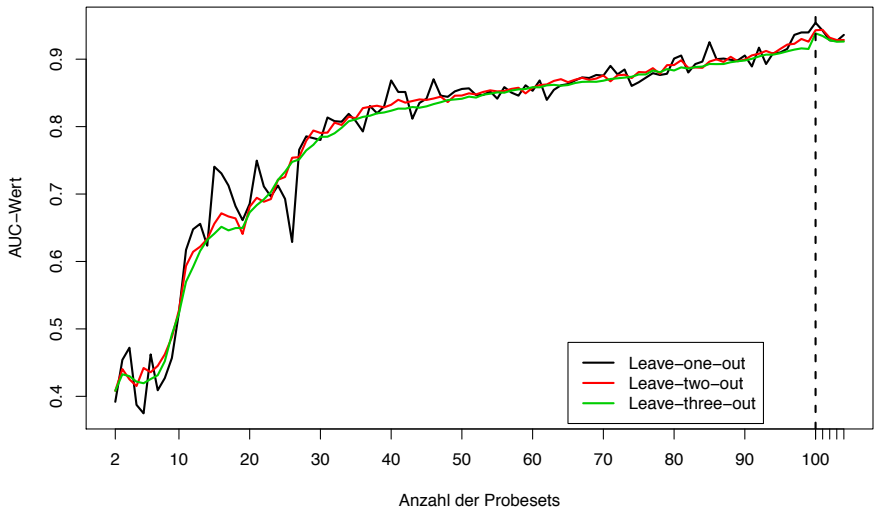

Abbildung 33: AUC-Wert in Abhängigkeit von der Anzahl verwendeter Probesets mit der höchsten Varianz über die Beobachtungen in der jeweiligen Trainingsmenge von UKK. Die Proben einer, zwei oder drei verschiedener Substanzen werden jeweils in die Testmenge aufgenommen. In der Abbildung werden die entsprechenden Funktionen durch verschiedene Farben hervorgehoben.

Die Kurven zeigen nun ein zu vorhergehenden Analysen anderes Verhalten. Bis auf wenige Ausnahmen steigen die drei Funktionen zuerst monoton an. Während bei der UKN1 Studie auch wenige Prädiktoren ausreichen, um einen AUC-Wert von über 0.9 zu erreichen, weist die Klassifikationsregel mit 10 Prädiktoren einen AUC-Wert von ca. 0.5 auf, was auf ein zufälliges Zuordnen der Proben hindeutet. Die Funktionen erreichen bei etwa 100 Probesets ihre maximalen Werte von ca. 0.95. Danach fallen die Kurven leicht ab. Es scheint, dass die optimale Wahl für die Anzahl der Prädiktoren bei etwa 100 liegt. Dies würde mit der Wahl bei der UKN1 Studie übereinstimmen. Um die Verfahren mit 10, 50 oder 100 verwendeten Probesets miteinander graphisch zu vergleichen, werden ferner die ROC-Kurven erstellt, siehe Abbildung 34.

Diese Grafik bestätigt die durch Abbildung 33 visuell gewonnenen Eindrücke. Die Klassifikationsregel mit 10 Probesets zeigt eine ROC-Kurve, die fast mit der Winkelhalbierenden übereinstimmt. Dies bedeutet, dass für eine beliebigen Wahl einer Schranke die Identität

$$\text{Spezifität} + \text{Sensitivität} = 1$$

gilt. Dies bedeutet, dass es sich um ein zufälliges Verfahren handelt. Die Abhängigkeiten für 50 bzw. 100 Probesets liegen wesentlich überhalb der ROC-Kurve für 10 Prädiktoren, wobei die letztere deutlich höher liegt. Dies entspricht Abbildung 33. Basierend auf diesen Analysen wird auch hier entschieden, 100 Probesets als Prädiktoren zu verwenden. Dies

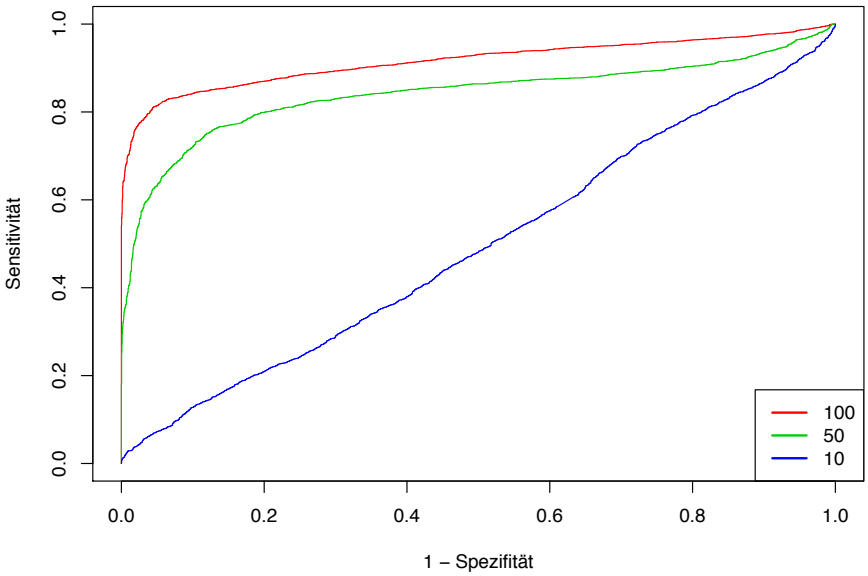

Abbildung 34: ROC-Kurven für das SVM-Verfahren mit 10, 50 und 100 Probesets mit der höchsten Varianz über die Beobachtungen in der Trainingsmenge von UKK. Auf der x-Achse ist die Falsch-Positiv-Rate abgetragen, auf der y-Achse ist die Richtig-Positiv-Rate abgetragen.

scheint eine sinnvolle Klassifikationsgüte sicherzustellen und ermöglicht eine einheitliche Empfehlung für die Auswertungen anderer Studien.

Im nächsten Schritt wird das SVM-Verfahren trainiert. Dabei werden sowohl Leave-one-out als auch Leave-two-out Kreuzvalidierungen durchgeführt. In Tabelle 15 auf Seite 78 sind die Ergebnisse dargestellt. Daraus folgt, dass bis auf eine Ausnahme alle Substanzen in beiden Auswertungsszenerien richtig vorhergesagt werden. Die Substanz Panobinostat sticht allerdings deutlich hervor. Auch im Rahmen einer Leave-one-out Auswertung werden die entsprechenden Proben zu der Klasse der quecksilberhaltigen Substanzen zugeordnet. Dieses Ergebnis stimmt mit Abbildung 32 insofern überein, als sich dort die mit Panobinostat behandelten Proben stark von der HDAC-Inhibitoren absetzen.

Einfluss der technischen Replikate In diesem Unterkapitel wird der Einfluss der Anzahl aufgenommener technischer Replikate analysiert, und die Ergebnisse werden präsentiert. Die Ergebnisse für die Substanz Belinostat sind exemplarisch in Abbildung 35 präsentiert.

Als Erstes fällt auf, dass die Proben sehr heterogen sind. Die Mediane der Stichprobenverteilungen von Klassenwahrscheinlichkeiten variieren im Laufe der Leave-one-out Auswertung sehr stark: Während die Replikate 5 und 7 einen Median über 0.9 aufweisen und somit mit großer Wahrscheinlichkeit als HDAC-Inhibitoren vorhergesagt werden, zeigt die sechste Probe einen Median unter 0.25 und wird somit als Mercurial klassifiziert.

Werden jeweils drei oder zwei technische Replikate der Trainingssubstanz verwendet, verschiebt sich die Masse der Verteilungen von Vorhersagewahrscheinlichkeiten hin zu einem Wert von 0.5, wobei die Varianz sich vergrößert. Dabei scheinen die Vorhersagen der (ursprünglich falsch klassifizierten) Probe 6 am stärksten zu schwanken. Dagegen sind die Verteilungen der Vorhersagen der Probe 7, die am sichersten als HDAC-Inhibitor erkannt wurde, viel kompakter. Die Boxplot-Abbildungen für die anderen Substanzen zeigen ein ähnliches Verhalten und sind ab Seite 152 im Anhang zu finden. Der R-Code ist auf Seite 234 angegeben.

Sensitivitätsanalyse Hier werden die Ergebnisse der Sensitivitätsanalyse präsentiert. Die Ergebnisse für die Substanz Belinostat sind in Abbildung 36 dargestellt.

Für die unverrauschten Proben gilt die Bemerkung über die Verteilungen von Vorhersagewahrscheinlichkeiten bei der Verwendung von allen technischen Replikaten in der Trainingsmenge (vgl. Abschnitt 4.6.1). Erhöht man den Anteil der verrauschten Variablen (Noise), so erhöht sich die Stichprobenstandardabweichung der Verteilungen und die Mediane konvergieren gegen den Wert von 0.5. Dieses Verhalten ist bei der Erhöhung der Rauschstärke ρ (Amplifier) stärker ausgeprägt. So liegen die Mediane der Wahrschein-

Abbildung 35: Boxplot-Graphiken für die HDACi-Vorhersagewahrscheinlichkeiten von Belinostat in der UKK Klassifikationsstudie. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist die Anzahl der verwendeten technischen Replikate in der Trainingsmenge angezeigt, auf der y-Achse ist die Vorhersagewahrscheinlichkeiten abgetragen.

lichkeitsverteilungen aller Proben bei maximalen Rauschanteil und -stärke im Bereich zwischen 0.3 und 0.7. Die Abbildungen für die restlichen 11 Substanzen sind im Anhang ab Seite 217 zu finden. Der R-Code ist auf Seite 237 angegeben.

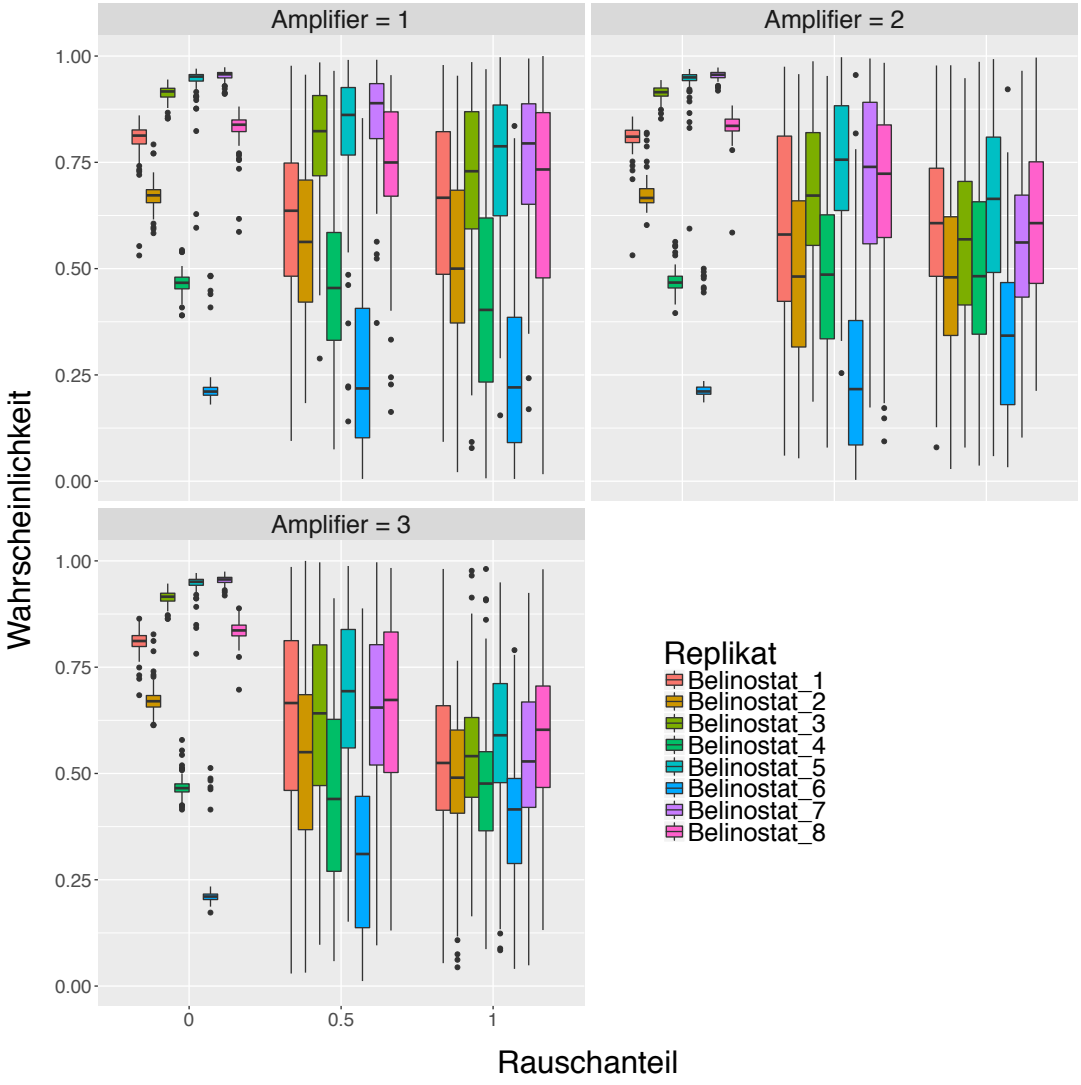

Abbildung 36: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Belinostat nach dem Verrauschen der Daten für Rauschstärke 1 (links), Rauschstärke 2 (mitte) und Rauschstärke 3 (rechts). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen abgetragen, auf der y-Achse die Vorhersagewahrscheinlichkeiten.

	Belinostat	Entinostat	Panobinostat	SAHA	VPA low	VPA high	TSA
Belinostat	0.86	0.87	0.65	0.79	0.58	0.73	0.62
Entinostat	0.99	1.00	1.00	1.00	1.00	1.00	1.00
Panobinostat	0.09	0.14	0.19	0.18	0.19	0.29	0.15
SAHA	0.96	0.98	0.99	0.96	0.96	0.96	0.91
VPA low	0.70	0.80	0.69	0.88	0.69	0.73	0.62
VPA high	0.98	0.99	1.00	0.98	0.88	0.99	0.99
TSA	0.93	0.88	0.95	0.88	0.80	0.92	0.91
$HgBr_2$	0.82	0.83	0.90	0.86	0.81	0.87	0.83
PMA	0.87	0.81	0.93	0.72	0.83	0.83	0.88
PCMB	0.85	0.82	0.91	0.87	0.81	0.78	0.84
$HgCl_2$	0.76	0.81	0.92	0.82	0.77	0.85	0.81
MeHg low	0.80	0.64	0.93	0.79	0.83	0.84	0.76
MeHg high	0.85	0.96	0.81	0.91	0.80	0.86	0.92
Thimerosal	0.88	0.84	0.94	0.89	0.82	0.86	0.78

	$HgBr_2$	PMA	PCMB	$HgCl_2$	MeHg low	MeHg high	Thimerosal
Belinostat	0.65	0.81	0.73	0.65	0.84	0.88	0.83
Entinostat	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Panobinostat	0.27	0.27	0.19	0.23	0.21	0.18	0.16
SAHA	0.97	0.93	0.98	0.97	0.94	0.98	0.94
VPA low	0.72	0.79	0.77	0.73	0.73	0.77	0.68
VPA high	0.98	0.95	0.99	0.99	0.99	0.94	0.98
TSA	0.90	0.92	0.84	0.78	0.93	0.93	0.87
$HgBr_2$	0.87	0.74	0.82	0.87	0.84	0.84	0.85
PMA	0.76	0.83	0.78	0.80	0.75	0.81	0.63
PCMB	0.82	0.74	0.81	0.71	0.78	0.83	0.83
$HgCl_2$	0.84	0.78	0.66	0.80	0.83	0.80	0.71
MeHg low	0.75	0.78	0.74	0.77	0.80	0.62	0.77
MeHg high	0.68	0.84	0.89	0.90	0.67	0.87	0.81
Thimerosal	0.78	0.74	0.93	0.55	0.80	0.84	0.78

Tabelle 15: Ergebnisse der kreuzvalidierten Auswertung der Vorhersage des SVM-Verfahrens für die UKK Klassifikationsstudie. In Zeilen und Spalten stehen in die Testmenge aufgenommenen Substanzen. Die Werte in der Tabelle sind die über alle Proben gemittelten Wahrscheinlichkeiten der richtigen Klasse zugeordnet zu werden. Durch rote Farbe sind die falsch klassifizierten Substanzen hervorgehoben.

4.6.2 Analyse von Random Forests

In diesem Unterkapitel werden die Resultate des Random Forests Verfahrens erörtert. Zuerst wird die Abhängigkeit des AUC-Wertes von der Anzahl der verwendeten Probesets in Abbildung 37 graphisch untersucht.

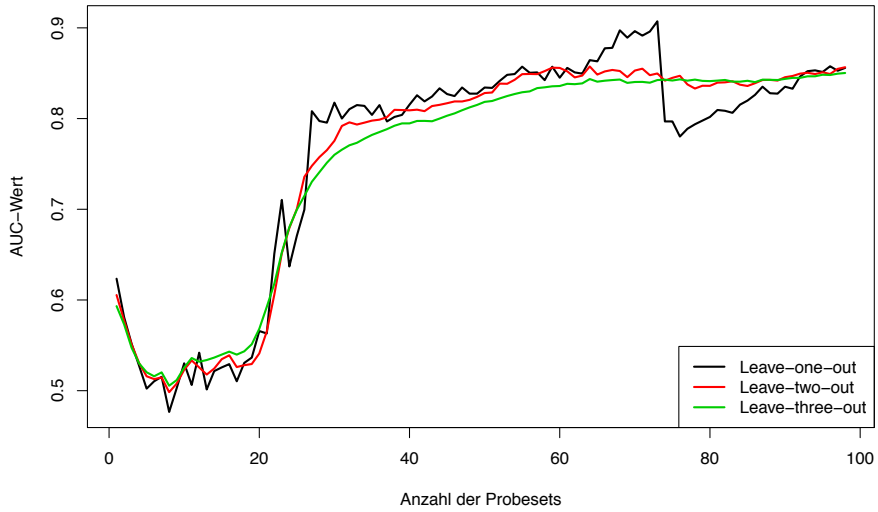

Abbildung 37: AUC-Wert in Abhängigkeit von der Anzahl verwendeter Probesets mit der höchsten Varianz über die Beobachtungen in der jeweiligen Trainingsmenge von UKK. Zum Diskriminieren wird ein Random Forests verwendet. Die Proben einer, zwei oder drei verschiedener Substanzen werden jeweils in die Testmenge aufgenommen. Die jeweilige Vorgehensweise wird respektive als Leave-one-out, Leave-two-out oder Leave-three-out bezeichnet. In der Abbildung werden die entsprechenden Funktionen durch verschiedenen Farben hervorgehoben.

Diese Abhängigkeiten tragen allerdings wenig zur Klärung der Frage nach der optimalen Anzahl der Prädiktoren bei: Die Klassifikationsregeln mit bis zu 20 Probesets führen zu AUC-Werten kleiner als 0.6. Danach steigen alle drei Kurven stark an und erreichen bei 30 Genprodukten mit AUC-Werten von etwa 0.8 ein Plateau. Eine weitere Erhöhung der Prädiktorenanzahl scheint wenig zu verändern. Um die Verfahren mit 10, 50 oder 100 verwendeten Probesets miteinander nochmal graphisch zu vergleichen, werden die ROC-Kurven erstellt und in Abbildung 38 dargestellt.

Diese Abbildung bekräftigt die durch Graphik 37 visuell gewonnenen Eindrücke. Die Klassifikationsregel mit verwendeten 10 Probesets zeigt eine ROC-Kurve, die etwa linear mit der Steigung eins verläuft und somit fast mit der Winkelhalbierenden übereinstimmt. Dies weist auf ein zufälliges Verfahren hin. Die beiden Verfahren mit 50 und 100 Prädiktoren zeigen eine bessere Klassifikationsgüte, wobei das letztere besser zu sein scheint. Dies veranlasste dazu, bei den späteren Auswertungen ein Verfahren zu verwenden, welches 100

Probesets mit der höchsten Varianz über die Trainingsmenge verwendet. Somit wird die Trainingsmenge für beide Verfahren bei beiden Studien auf die gleiche Weise bestimmt, wodurch der Entschluss bekräftigt wird.

Im nächsten Schritt wurde ein Random Forests trainiert. Dabei wurden sowohl Leave-one-out als auch Leave-two-out Kreuzvalidierungen durchgeführt. In Tabelle 16 sind die Ergebnisse dargestellt. Analog zu den in Tabelle 15 präsentierten Ergebnissen ordnet auch das Random Forests Verfahren die Substanz Panobinostat der Klasse der quecksilberhaltigen Substanzen zu. Die Tatsache, dass beide Verfahren auch im Falle einer Leave-one-out Validierung die Substanz falsch klassifizieren, scheint ein Beleg zu sein, dass Panobinostat in seiner genotypischen Wirkung nach einer vierzehntägigen Einwirkperiode den quecksilberhaltigen Substanzen nahe liegt. Ferner wird die Substanz Belinostat von Random Forests ausschließlich Mercurials zugeordnet. Bis auf eine Ausnahme werden alle anderen Substanzen - auch im Falle von Leave-two-out - richtig klassifiziert.

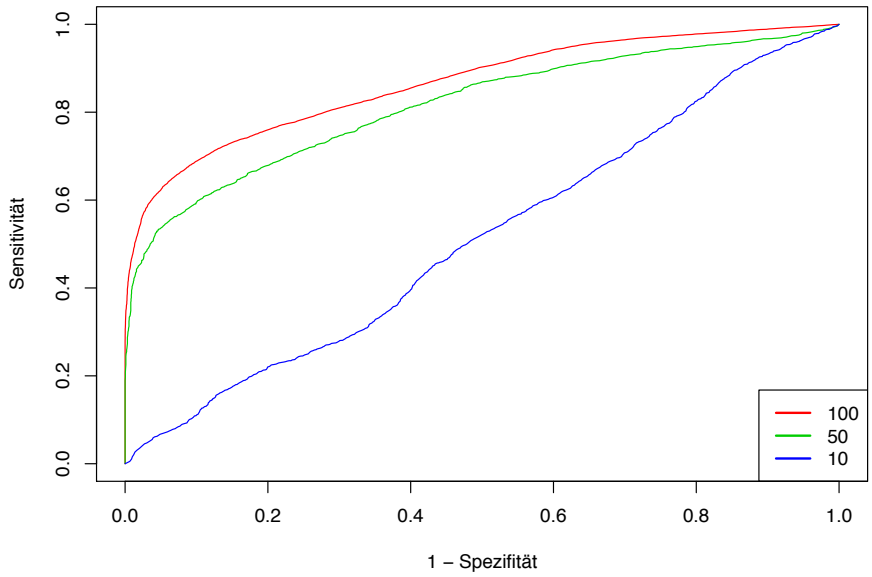

Abbildung 38: ROC-Kurven für das Random Forests Verfahren mit 10, 50 und 100 Probesets mit der höchsten Varianz über die Beobachtungen in der Trainingsmenge von UKK. Auf der x-Achse ist die Falsch-Positiv-Rate abgetragen, auf der y-Achse ist die Richtig-Positiv-Rate abgetragen.

	Belinostat	Entinostat	Panobinostat	SAHA	VPA low	VPA high	TSA
Belinostat	0.35	0.49	0.20	0.33	0.26	0.25	0.32
Entinostat	0.99	0.98	0.98	0.96	0.97	0.93	0.98
Panobinostat	0.35	0.42	0.41	0.39	0.40	0.34	0.41
SAHA	0.94	0.89	0.88	0.86	0.89	0.86	0.82
VPA low	0.76	0.59	0.68	0.72	0.81	0.80	0.79
VPA high	0.97	0.84	0.97	0.90	0.87	0.95	0.96
TSA	0.78	0.56	0.51	0.55	0.64	0.61	0.63
$HgBr_2$	0.73	0.69	0.72	0.68	0.70	0.63	0.67
PMA	0.90	0.81	0.82	0.66	0.79	0.88	0.78
PCMB	0.78	0.52	0.79	0.72	0.74	0.64	0.57
$HgCl_2$	0.84	0.61	0.69	0.61	0.61	0.63	0.68
MeHg low	0.94	0.93	0.91	0.81	0.88	0.82	0.88
MeHg high	0.90	0.72	0.59	0.55	0.75	0.63	0.80
Thimerosal	0.97	0.91	0.97	0.94	0.94	0.90	0.95

	$HgBr_2$	PMA	PCMB	$HgCl_2$	MeHg low	MeHg high	Thimerosal
Belinostat	0.30	0.32	0.38	0.24	0.33	0.32	0.36
Entinostat	0.99	0.99	0.99	0.98	0.99	1.00	0.99
Panobinostat	0.41	0.40	0.40	0.47	0.42	0.30	0.43
SAHA	0.87	0.86	0.90	0.92	0.84	0.91	0.87
VPA low	0.88	0.86	0.91	0.87	0.84	0.82	0.82
VPA high	0.93	0.94	0.97	0.95	0.95	0.96	0.91
TSA	0.64	0.59	0.67	0.72	0.62	0.58	0.62
$HgBr_2$	0.68	0.55	0.50	0.68	0.63	0.67	0.63
PMA	0.67	0.75	0.73	0.77	0.65	0.80	0.69
PCMB	0.55	0.67	0.70	0.71	0.66	0.75	0.70
$HgCl_2$	0.61	0.64	0.61	0.62	0.58	0.52	0.53
MeHg low	0.88	0.63	0.72	0.86	0.85	0.51	0.89
MeHg high	0.63	0.67	0.67	0.64	0.42	0.59	0.50
Thimerosal	0.92	0.86	0.94	0.85	0.90	0.85	0.96

Tabelle 16: Ergebnisse der kreuzvalidierten Auswertung der Vorhersage des Random Forest Verfahrens für die UKK Klassifikationsstudie. In Zeilen und Spalten stehen in die Testmenge aufgenommenen Substanzen. Übrigen 10 bzw. 11 Substanzen bilden die Trainingsmenge. Somit stehen auf der Hauptdiagonalen von links oben nach rechts unten die Ergebnisse einer leave-one-out Kreuzvalidierung. Außerhalb der Hauptdiagonalen befinden sich die Resultate der leave-two-out Kreuzvalidierung. Die Werte in der Tabelle sind die gemittelten Wahrscheinlichkeiten der richtigen Klasse zugeordnet zu werden. Durch rote Farbe sind die falsch klassifizierten Substanzen hervorgehoben.

Einfluss der technischen Replikate In diesem Unterkapitel wird der Einfluss der Anzahl aufgenommener technischer Replikate analysiert, und die Ergebnisse werden präsentiert. Die Ergebnisse für die Substanz Belinostat sind auf Abbildung 39 dargestellt.

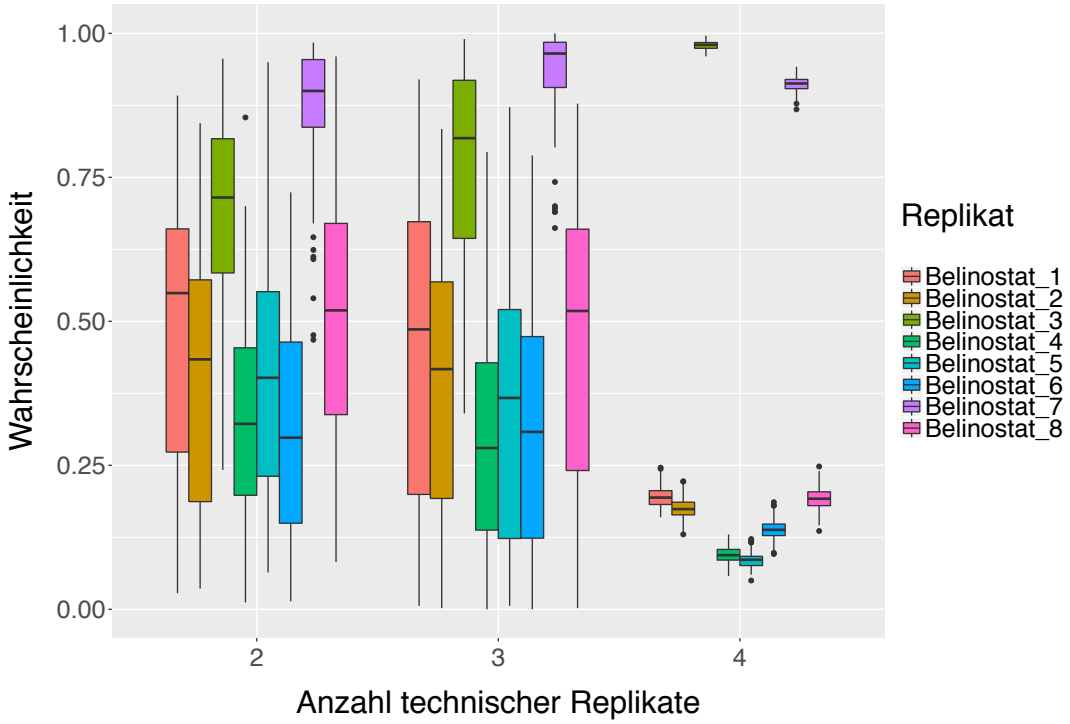

Abbildung 39: Boxplot-Graphiken für die HDACi-Vorhersagewahrscheinlichkeiten von Belinostat in der UKK Klassifikationsstudie. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Zum Klassifizieren wurde das Random Forests verwendet.

Vergleicht man diese Abbildung mit der entsprechenden Präsentation der Ergebnisse für die SVM (Abbildung 35 auf Seite 76), so stellt man fest, dass die beiden Verfahren auch bei der Hinzunahme der kompletten Trainingsmenge einzelne Replikate ganz unterschiedlich vorhersagen. Während das SVM-Verfahren die Replikate mit laufenden Nummern 1, 2, 3, 5, 7 und 8 im Mittel korrekt vorhersagt, klassifiziert das RF-Verfahren nur die Replikate 3 und 7 im Mittel richtig. Die anderen Proben werden mit einer Wahrscheinlichkeit von über 0.75 als quecksilberhaltig vorausgesagt. Reduziert man die Trainingsmenge, verschiebt sich die Masse der Verteilungen von Vorhersagewahrscheinlichkeiten zu einem Wert von 0.5, wobei die Varianz sich vergrößert. Die Boxplot-Abbildungen für die anderen Substanzen zeigen ein ähnliches Verhalten und sind ab Seite 165 im Anhang zu finden. Der R-Code ist auf Seite 235 angegeben.

Sensitivitätsanalyse In diesem Unterkapitel werden die Ergebnisse der Sensitivitätsanalyse für das Random Forests Verfahren präsentiert. Dabei wird wie in den früheren Kapiteln 4.5.1 , 4.5.2 und 4.6.1 vorgegangen: Für die Rauschstärke (Amplifier) ρ werden 3 Werte genommen: $\rho = 1, 2, 3$. Der Anteil (Noise) t der verrauschten Variablen wird variiert: Es werden 0, 50 und 100 % der Probesets perturbiert. Die Ergebnisse für die Substanz $HgBr_2$ sind exemplarisch in Abbildung 40 dargestellt.

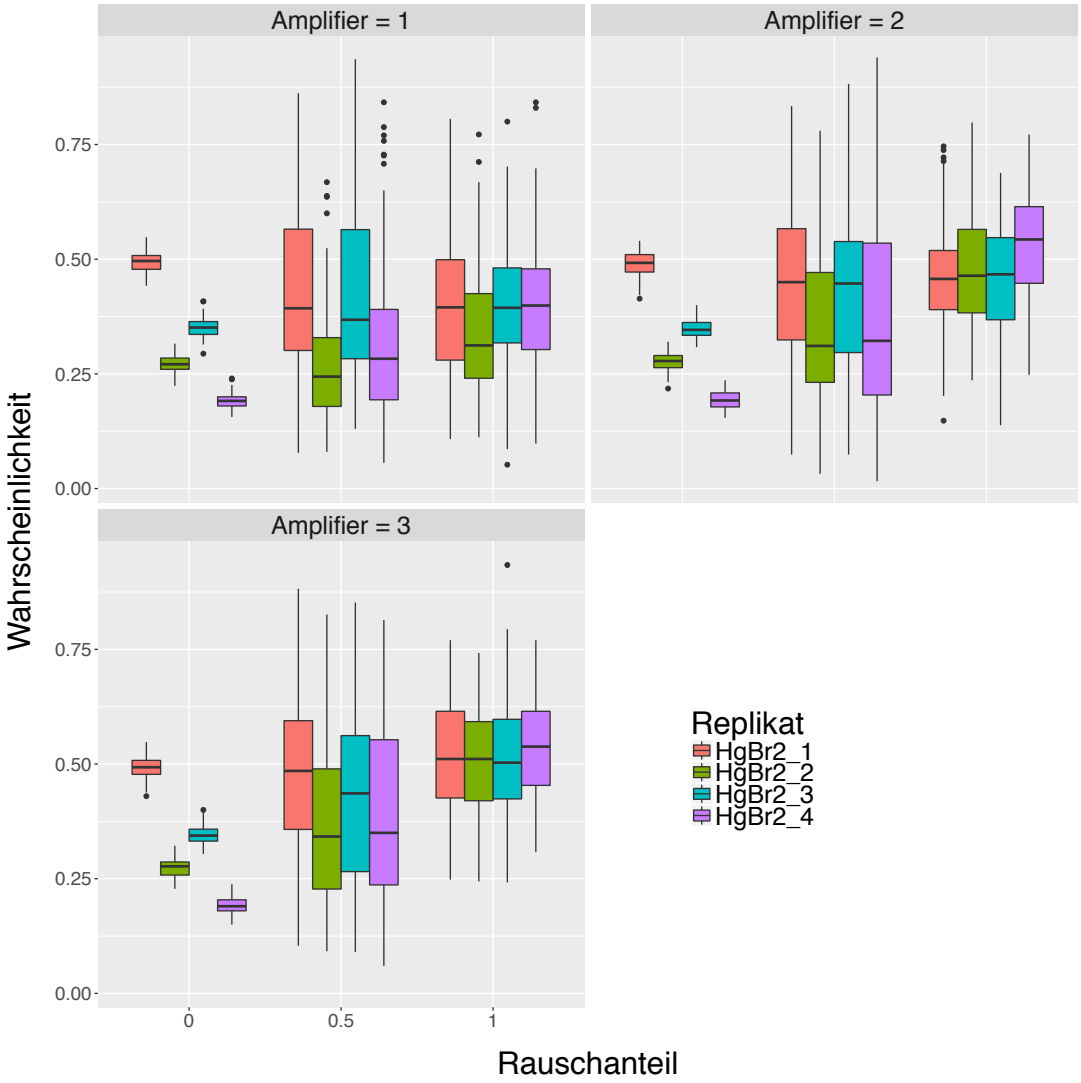

Abbildung 40: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgBr_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forests verwendet.

Die Ergebnisse dieser Analyse ähneln sehr stark den Resultaten vorhergehender Kapitel: Im Falle von unverrauschten Daten unterscheiden sich die Verteilungen von Vorhersagewahrscheinlichkeiten teilweise deutlich. Während das vierte Replikat sicher als Mercurial klassifiziert wird, liegt der Median der Verteilung im Falle von dem ersten Replikat bei 0.5. Dies bedeutet, dass etwa in der Hälfte der Fälle die Probe falsch klassifiziert wird. Wird der Anteil der verrauschten Variablen bei gleichbleibender Rauschstärke erhöht, so konvergiert die Masse der Verteilung gegen einen Wert von 0.5, wobei die Varianz zurückgeht. Dies bedeutet, dass bei starkem Verrauschen die Klassifikationsregel die Klassenlabels immer zufälliger verteilt: Die Zugehörigkeitswahrscheinlichkeiten gehen gegen 0.5. Den gleichen Effekt beobachtet man bei der Erhöhung der Rauschenstärke bei gleichbleibendem Anteil der perturbierten Variablen. Die Abbildungen für die restlichen 11 Substanzen sind im Anhang ab Seite 204 zu finden. Der R-Code ist auf Seite 237 angeben.

4.6.3 Anwendung auf UKK Zeitfensterstudie

In diesem Abschnitt wird die Generalisierungsfähigkeit der beiden Klassifikationsmethoden exemplarisch an Hand der VPA Zeitfensterstudie untersucht. Da die fürs Estellen der Lernregel verwendete Zellen am Tage 14 gesammelt wurden, nimmt man zur Validierung diejenigen Proben heran, die für 14 Tage der Valproinsäure bzw. dem Lösungsmittel ausgesetzt wurden. Wie man Tabelle 7 auf Seite 16 entnehmen kann, gibt es insgesamt 24 solche Proben.

Analog zu vorhergehendem Abschnitt wird der Hauptkomponentenplot der kontrollbereinigten Daten der UKK Klassifikationsstudie erstellt. Als Erstes werden die Expressionsprofile der Kontrolle paarweise von den Expressionen der behandelten Zellen abgezogen. Mit Hilfe der Rotationsmatrix U aus der Zerlegung (3.1) lassen sich die neuen Datenpunkte in den bestehenden Hauptkomponentenplot projizieren. Das Ergebnis dieser Berechnung ist in Abbildung 41 präsentiert. Dabei sind übersichtlichkeitshalber nur mit VPA behandelten Zellen dargestellt. Die vollständige Abbildung ist im Anhang auf Seite 181 präsentiert.

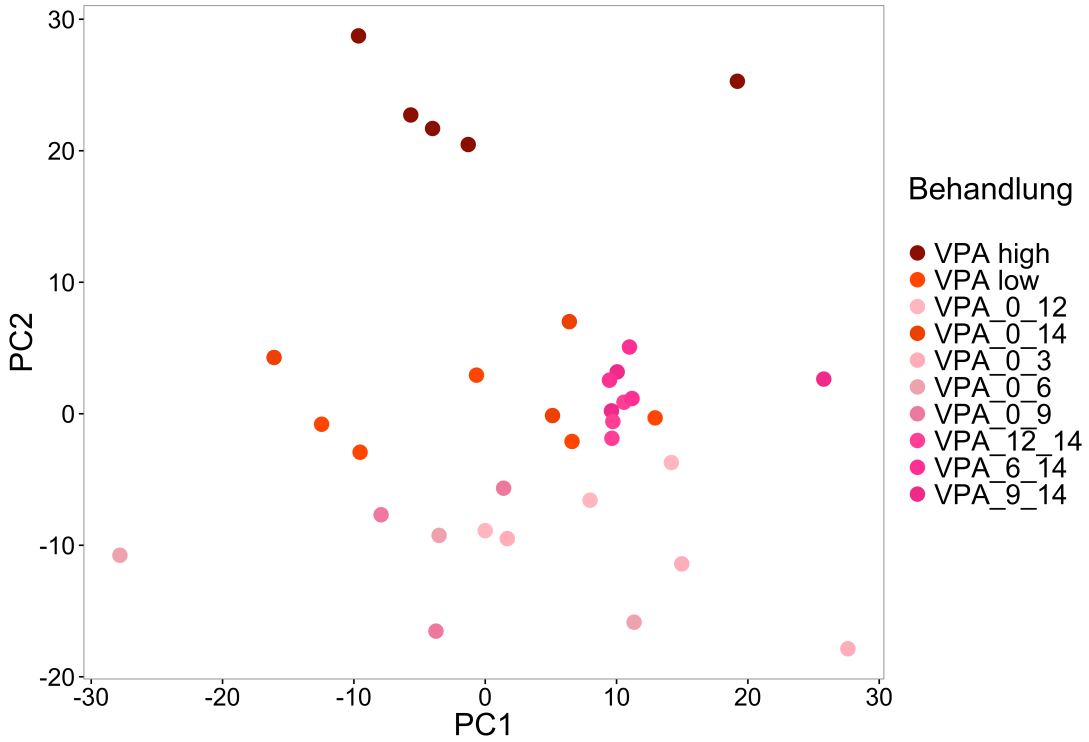

Abbildung 41: Der Hauptkomponentenplot der Daten aus der UKK Klassifikationsstudie. Die Daten aus der VPA Zeitfenserstudie sind mit Hilfe der Rotationsmatrix reinprojiziert. Für eine bessere Darstellung der Ergebnisse sind nur mit VPA behandelten Zellen gelassen.

Diese Hauptkomponentenabbildung beinhaltet 10 Proben aus der UKK Klassifikationsstudie. Sie sind mit kleineren bzw. höheren Konzentration der Valproinsäure behandelt.

Ferner sind 24 Proben aus der VPA Zeitfensterstudie abgebildet. Die erste Zahl in der Legende bezeichnet dabei den Beginn der Behandlung mit Valproinsäure und die zweite Zahl das Ende. So bedeutet z.B. der Name VPA_6_14, dass die entsprechenden Zellen für 14 Tage dem Lösungsmittel ausgesetzt wurden. Von Tag 6 bis Tag 14 wurde zusätzlich die Valproinsäure hinzugefügt.

Während die erste Hauptkomponente eher die Varianz zwischen den Replikaten wiedergibt, eignet sich die zweite Hauptkomponente besser, um die Unterschiede zwischen den einzelnen Gruppen darzustellen. So befinden sich mit der höchsten Konzentration der Valproinsäure behandelte Zellen am oberen Rand der Abbildung. Am unteren Rand der Grafik sind die Proben zu finden, welche im Rahmen der Zeitfensterstudie erstellt wurden. Bezeichnenderweise wurden sie unterschiedlich lang mit Valproinsäure behandelt, jedoch vor dem vierzehnten Tag gesammelt. Die am Tag 14 gesammelten Proben haben Ordinaten um den Wert Null, unabhängig von der Dauer der VPA-Behandlung. Dies lässt vermuten, dass nicht die Länge des Zeitfensters von entscheidender Relevanz ist, sondern die Lage bezüglich der vierzehntägigen Periode der Analyse. So fallen diejenigen Zellen, welche lediglich zwei Tage lang der Valproinsäure ausgesetzt wurden, mehr mit den Zellen zusammen, die 14 Tage lang behandelt wurden, als die Zellen, deren Behandlungsfenster 12 Tage betrug, vorausgesetzt diese zwei Tage fallen auf das Ende der vierzehntägigen Periode.

Im nächsten Schritt werden sowohl eine SVM als auch eine Random Forests Klassifikationsregel gebildet, wobei alle Daten aus der UKK Klassifikationsstudie als Trainingsmenge verwendet werden. Die Proben aus der VPA Zeitfensterstudie werden dann vorhergesagt. Die pro Gruppe gemittelten Vorhersagewahrscheinlichkeiten sind in Tabelle 17 angegeben.

Begin der Einwirkperiode	Ende der Einwirkperiode	HDACi-Vorhersage (SVM)	HDACi-Vorhersage (RF)
0	3	0.133	0.296
0	6	0.163	0.188
0	9	0.265	0.264
0	12	0.236	0.209
0	14	0.941	0.831
6	14	0.890	0.735
9	14	0.768	0.671
12	14	0.722	0.746

Tabelle 17: Mittleren Vorhersagewahrscheinlichkeiten für die Proben aus der VPA Zeitfensterstudie. Die Klassifikation erfolgte mit SVM und Random Forests, wobei alle Daten der UKK Klassifikationsstudie zum Erstellen verwendet wurde.

Die Ergebnisse der Vorhersage stimmen mit der Hautpkomponentengrafik insofern überein, als sämtliche Proben, deren Einwirkperiode sich bis zum Tag 14 erstreckte, von beiden Verfahren richtig als HDAC-Inhibitor klassifiziert werden. Demgegenüber werden all die

anderen Proben falsch klassifiziert. Obwohl eine biologische Interpretation noch aussteht, kann man vermuten, dass die Stammzellen sich von der Wirkung der Valproinsäure „erholen“ können, sobald die Substanz nicht mehr hinzugefügt wird. Diese Untersuchung liefert einen Hinweis darauf, dass der Zeitraum von lediglich zwei Tagen dazu ausreichend ist, denn die Zellen mit einer „Erholungsphase“ nach dem Einwirken von VPA werden nicht als HDAC-Inhibitoren erkannt. Demgegenüber scheint die Einwirkperiode von zwei Tagen genügend, um für die HDAC-Inhibitoren typische Muster zu initiieren. Die Tatsache, dass die entsprechenden Zellen richtig vorhergesagt werden, liefert einen Beleg für diese Vermutung. Man kann die Hypothese aufstellen, dass eine längere Einwirkperiode sich kumulativ auswirkt, denn die vom SVM-Verfahren berechneten Wahrscheinlichkeiten sind grundsätzlich um so größer, je großer die Einwirkperiode ist. Lediglich das Zeitfenster vom Tag Null bis Tag 12 stellt eine Ausnahme dar. Demgegenüber korrelieren die vom RF-Verfahren bestimmten Wahrscheinlichkeiten nicht mit der Dauer der Substanzzugabe.

5 Zusammenfassung

In diesem Kapitel werden die Ergebnisse dieser Arbeit zusammengefasst und erläutert. Innerhalb der Untersuchung setzte man sich zum Ziel, eine Abfolge von Verfahren bereitzustellen, welche man bei der Analyse von hochdimensionalen biologischen Daten sequentiell anwenden kann. Es handelt sich dabei um Messungen der Genexpression mit Hilfe eines Chips des Herstellers Affymetrix. Diese Analysen stellen die Wissenschaft vor eine große Herausforderung. Komplexe biologische Prozesse, eine große Anzahl von Variablen, welche mit einander korreliert sein können und wenige Beobachtungen erschweren die Untersuchungen und deren Interpretation.

Demgegenüber stellen die Statistik bzw. das maschinelle Lernen verschiedene Verfahren und Methoden zur Verfügung, um die hochdimensionale Daten zu analysieren. Der Fortschritt in den biologisch motivierten exakten Wissenschaften ließ die Verknüpfungen zwischen den Fachgebieten Biologie, Mathematik, Informatik und Statistik entstehen und stark expandieren. Diese Konvergenz von verschiedenen wissenschaftlichen Bereichen ruft nach einer interdisziplinären Herangehensweise an die biologisch motivierten Aufgaben. Dies gilt vor allem für Fragestellungen in Bezug auf zelluläre Prozesse. Die mikroskopische Skala, komplex regulierte Netzwerke und das eng verbundene Zusammenspiel von Transkriptomik, Proteomik und Epigenetik machen die Analysen und deren Interpretation sehr herausfordernd. Affymetrix Technologie, RNA-Sequenzierung und ChIP-Sequenzierung generieren eine große Menge von Daten, zu deren Analyse viele statistische Werkzeuge benötigt werden. Statistische Methoden und die Verfahren des maschinellen Lernens ermöglichen die Untersuchung von diesen Daten und dienen der Erkennung von verschiedenen Mustern und der Informationsgewinnung. Dieser Prozess der „Offenlegung“ der Information wird in der Literatur als **Data Mining** bezeichnet.

In dieser Arbeit wurde es zum Ziel gesetzt, statistische Methoden zur Normalisierung, Aufarbeitung, Analyse und Visualisierung hochdimensionaler biologischer Daten vorzustellen, zu erörtern und auf praktische Beispiele anzuwenden. Diese Methoden lassen sich sowohl unabhängig von einander benutzen als auch in Rahmen einer sequenziellen Analyse. Diese Abfolge von Anwendungen wird in dieser Arbeit als „Pipeline“ bezeichnet. Der Verfasser erhofft sich, dass ihre Anwendung es ermöglicht, ein möglichst umfassendes Bild von den Daten zu bekommen und verborgene Kenntnisse und Muster offenzulegen. So erlauben Hauptkomponentenanalysen und Clustering erste deskriptive Eindrücke zu gewinnen. Obwohl diese Ergebnisse nicht überinterpretiert werden sollten, lassen sie mitunter verlässliche Schlüsse über die Qualität der Daten zu. So konnten mit deren Hilfe sowohl mögliche Ausreißer (Behandlung mit 650 µM in VPA chronischen Konzentrationsstudie, siehe Abbildung 6) als auch einen Batch-Effekt (UKN1 Konzentrationsstudie, siehe Abbildung 8) aufgezeigt werden.

Die Problematik des Batch-Effektes stellt wiederum eine große Herausforderung für die Interpretation von hochdimensionalen Daten dar. Auf Grund von technischen Limitierungen bzw. Wiederholungen von Experimenten können nicht alle Proben auf einmal bearbeitet werden. Dies zieht nach sich, dass oft ein nicht zu vernachlässigender Anteil der Varianz in den Messungen technischer Natur ist. Dies erschwert die Analyse und die Interpretation der Ergebnisse, insbesondere dann, wenn die unabhängigen Variablen (Geschlecht, Alter, Behandlungsgruppe) mit der abhängigen (Expressionsmessungen) über die Einteilung in die Abfertigungsschübe konfundiert sind. Um dies zu vermeiden, sollen die Batch-Gruppen randomisiert und homogen sein. Bei den vorliegenden Daten ist es insofern gewährleistet, dass alle Abfertigungslose alle interessierenden Behandlungsgruppen (Kontrollen, Mercurials und HDAC-Inhibitoren) enthalten. Trotzdem ist ein großer Anteil an Varianz auf die Batch-Effekte zurückzuführen. Dies ist vor allem dadurch ersichtlich, dass die Hauptkomponentenanalysen die Einteilung in die Abfertigungslose widerspiegeln (siehe Abbildung 8 in dem Unterkapitel 4.5). In dieser Arbeit beschränkte sich der Verfasser auf den paarweisen Abzug der Kontrollen von den behandelten Proben, um den Batch-Effekt zu reduzieren. Eine andere Möglichkeit stellt das ComBat-Verfahren dar, welches nicht nur einen additiven, sondern auch einen multiplikativen Effekt berücksichtigt.

Nächster Schritt in der Pipeline ist die Bestimmung von differentiell exprimierten Genen. Statistisch gesehen handelt es sich um den Vergleich von Mittelwerten zweier Stichproben. Der moderierte t-Test weist dabei einige Vorteile gegenüber dem üblichen Zweistichproben-t-Test auf, z.B. die verbesserte Schätzung der Stichprobenstandardabweichung und den Umgang mit kleineren Stichproben. Die praktische Umsetzung der Methode ist im R-Paket *limma* realisiert.

Die biologische Interpretation der im vorhergehenden Schritt ermittelten Genlisten stellt die Wissenschaft vor eine weitere Herausforderung. Während eine manuelle Durchsicht

der Ergebnisse sehr zeitaufwendig und ineffizient ist, bietet die Anreicherungsanalyse der biologischen Signaturen eine vielversprechende Alternative. Basierend auf dem Wissen, welches in Form von verschiedenen Ontologien gespeichert ist, lässt sich die Momentaufnahme des Transkriptoms vorliegend in der Form von Genlisten sinnvoll interpretieren. In dieser Arbeit wurde dazu die Anreicherung von Gene Ontology Gruppen und Zielgenen von Transkriptionsfaktoren vorgestellt. Während diese Methoden in den Publikationen Rempel u. a. (2015), Waldmann u. a. (2014) und Balmer u. a. (2014) zur Anwendung kamen, wird in dieser Arbeit auf eine Präsentation entsprechender Ergebnisse verzichtet. Die praktische Umsetzung erfolgte mit Hilfe des R-Pakets *topGO* und der Web-Anwendung *oPOSSUM*.

Eines der wichtigsten Gebiete des Data Mining stellt die Klassifikation dar. Dabei versucht man basierend auf einer Ansammlung von Objekten, deren Klassenzugehörigkeit bekannt ist, eine Regel zu erstellen, welche unbekannte Beobachtungen der richtigen Klasse zuordnet. In Rahmen dieser Arbeit wurden zwei Klassen von Substanzen diskriminiert, die Klasse der quecksilberhaltigen Substanzen (Mercurials) und die Histon-Deacetylase-Inhibitoren (HDAC-Inhibitoren oder HDACi). Dies sind toxikologische Substanzen, deren Einfluss auf den Reifeprozess der menschlichen Stammzellen von großer Wichtigkeit ist. Die einzelnen Fragestellungen, die im Rahmen der Analyse zu beantworten waren, lauteten:

- Auswahl der Prädiktoren,
- Analyse der Generalisierungseigenschaft des Lernverfahrens mit Hilfe der Kreuzvalidierung,
- Einfluss der Anzahl verwendeter Replikate,
- Einfluss von einem zusätzlichen Verrauschen der Daten,
- Validierung des Lernverfahrens auf einem externen Datensatz.

Diese Punkte wurden anhand von zwei verschiedenen Datensätzen (UKN1 und UKK) und zwei verschiedenen Lernverfahren (SVM und Random Forests) sequentiell untersucht. Es wurden sowohl Ähnlichkeiten als auch Unterschiede hinsichtlich der Vorhersagegüte der Algorithmen auf verschiedenen Datensätzen festgestellt. So stellte sich heraus, dass die Wahl von 100 Probesets mit der höchsten Varianz innerhalb der Trainingsmenge für alle Paare von Verfahren und Datensatz sinnvoll ist. Die Abhängigkeiten der AUC-Werte von den Anzahl der verwendeten Variablen gaben ein starkes Indiz dafür ab, vor allem bei dem UKK Datensatz.

Die Generalisierungseigenschaft des jeweiligen Lernverfahrens wurden in Rahmen einer Kreuzvalidierung untersucht. Dabei wurden die Replikate von einer, zwei oder drei Substanzen aus der Trainingsmenge entfernt, die Klassifikationsregel gebildet und die ausgelassenen Substanzen klassifiziert. Die Ergebnisse wurden sowohl in Form von Tabellen als

auch in Form von den Abbildungen mit den Vorhersagewahrscheinlichkeiten für einzelne Substanzen (siehe z.B. Abbildung 22) präsentiert. Es hat sich Folgendes herausgestellt:

- Im Rahmen einer leave-one-out Validierung wurden bei der Verwendung sowohl von SVM als auch Random Forests alle Substanzen in der UKN1 Studie richtig klassifiziert. Bei der UKK Studie wurden bei der Verwendung von SVM bis auf Panobinstat alle Substanzen richtig vorhergesagt. Bei der Verwendung von Random Forests wurden Panobinostat und Belinostat falsch zugeordnet.

- Im Rahmen einer leave-two-out Validierung innerhalb der UKN1 Studie stellte sich bei der Verwendung von SVMs heraus, dass bestimmte Substanzen andere Substanzen für die richtige Diskriminierung brauchen. So werden die mit Belinostat behandelten Proben in Abwesenheit von Entinostat falsch vorhergesagt und umgekehrt. Dies ist aus Sicht des Verfassers einerseits auf eine große Ähnlichkeit zwischen diesen Substanzen zurückzuführen. Andererseits kann man vermuten, dass das Fehlen ähnlicher Substanzen in der Trainingsmenge der Grund für dieses Ergebnis ist. Die Verwendung von Random Forests lieferte ähnliche Ergebnisse.

- Die Analyse des Einflusses der Anzahl verwendeter Replikate brachte zu Tage, dass sich die Vorhersagen für die einzelne Proben stark unterscheiden. So konnte man beobachten, dass während einige Proben mit einer Wahrscheinlichkeit von über 90 % der richtigen Klasse zugeordnet wurden, andere hingegen falsch klassifiziert wurden (siehe Beispiel Belinostat auf Abbildung 36). Reduziert man die Trainingsmenge, so verschlechtert sich die Vorhersagegüte: Die Vorhersagewahrscheinlichkeiten konvergieren gegen den Wert von 0.5, und die Varianz erhöht sich. Trotzdem werden die Proben in der Testmenge bis auf wenige Ausnahmen im Mittel richtig vorhergesagt, sogar unter Verwendung von zwei technischen Replikaten jeweiliger Trainingssubstanz.

- Es zeigte sich, dass im Falle von einem moderaten zusätzlichen Verrauschen der Daten (Rauschenstärke $\rho = 1$) eine korrekte Klassifikation in der Regel möglich ist. Dies ist ein Zeichen für die Robustheit der angewandten Verfahren.

- Die guten Generalisierungseigenschaften beider Verfahren wurden insofern bestätigt, als bei einer Auswertung auf einem externen Datensatz diejenigen Proben richtig klassifiziert wurden, deren Exposition mit der der Trainingsmenge vergleichbar war. So wurden im Falle des Testsystems UKN1 diejenigen Proben richtig vorhergesagt, welche mit den Konzentrationen behandelt wurden, die auch eine Änderung des Expressionprofils induziert hatten (siehe Tabelle 14 auf Seite 71). Im Falle des Testsystems UKK wurden diejenigen Proben korrekterweise als HDAC-Inhibitoren klassifiziert, deren Exposition bis zum Ende der Analyse dauerte, so dass die Zellen sich von der Wirkung der Substanz nicht erholen konnten.

Zusammenfassend kann man sagen, dass die ins Auge gefassten Fragestellungen beantwortet werden konnten. Die Ergebnisse dieser Analyse wurden in der Fachliteratur veröffentlicht (z.B. in Rempel u. a. (2015)). Der Verfasser erhofft sich, dass diese Arbeit das interdisziplinäre Zusammenwirken von Biologen und Statistikern verstärkt. Als Empfehlung für die zukünftigen Studien kann man vorschlagen, dass mehr verschiedene Substanzen aufgenommen werden sollten, sogar auf Kosten einer eventuellen Reduzierung der Anzahl technischer Replikate. Die erörterten Analysen und die erstellte Pipeline kann ferner auf die anderen hochdimensionalen Daten übertragen werden, z.B. auf die durch RNA- oder ChIP-Sequenzierungstechnik gewonnenen Datensätze.

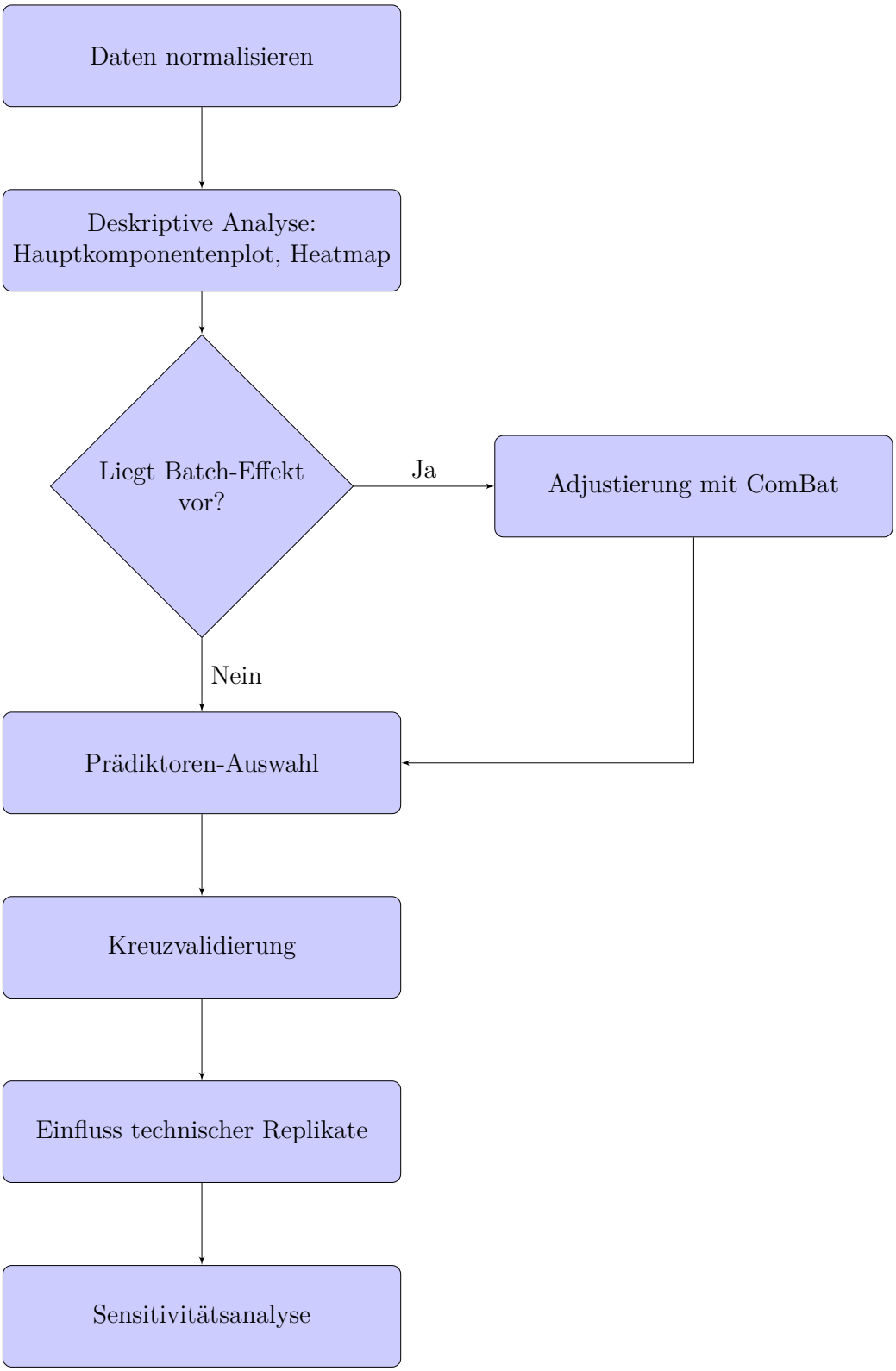

Abbildung 42: Eine schematische Abfolge von Schritten zur Klassifikation hochdimensionaler Microarray-Daten

Literaturverzeichnis

[Alexa u. Rahnenführer 2010] Alexa, Adrian ; Rahnenführer, Jorg: topGO: enrichment analysis for gene ontology. In: *R package version* 2 (2010), Nr. 0

[Ang u. a. 2015] Ang, Jun C. ; Haron, Habibollah ; Hamed, Haza Nuzly A.: Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In: *Current Approaches in Applied Artificial Intelligence*. Springer, 2015, S. 468–477

[Ashburner u. a. 2006] Ashburner, M ; Ball, C ; Blake, J u. a.: Gene ontology: tool for the unification of biology. The gene ontology consortium database resources of the national center for biotechnology information. In: *Nucleic Acids Research* 34 (2006)

[Babyak 2004] Babyak, Michael A.: What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. In: *Psychosomatic medicine* 66 (2004), Nr. 3, S. 411–421

[Balmer u. a. 2014] Balmer, Nina V. ; Klima, Stefanie ; Rempel, Eugen ; Ivanova, Violeta N. ; Kolde, Raivo ; Weng, Matthias K. ; Meganathan, Kesavan ; Henry, Margit ; Sachinidis, Agapios ; Berthold, Michael R. u. a.: From transient transcriptome responses to disturbed neurodevelopment: role of histone acetylation and methylation as epigenetic switch between reversible and irreversible drug effects. In: *Archives of toxicology* 88 (2014), Nr. 7, S. 1451–1468

[Balmer u. a. 2012] Balmer, Nina V. ; Weng, Matthias K. ; Zimmer, Bastian ; Ivanova, Violeta N. ; Chambers, Stuart M. ; Nikolaeva, Elena ; Jagtap, Smita ; Sachinidis, Agapios ; Hescheler, Jürgen ; Waldmann, Tanja u. a.: Epigenetic changes and disturbed neural development in a human embryonic stem cell-based model relating to the fetal valproate syndrome. In: *Human molecular genetics* 21 (2012), Nr. 18, S. 4104–4114

[Bandyopadhyay u. a. 2014] Bandyopadhyay, Supriyo ; Mallik, Saurav ; Mukhopadhyay, Amit: A survey and comparative study of statistical tests for identifying differential expression from microarray data. In: *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 11 (2014), Nr. 1, S. 95–115

[Beeson u. Lippman 2006] Beeson, Diane ; Lippman, Abby: Egg harvesting for stem cell research: medical risks and ethical problems. In: *Reproductive BioMedicine Online* 13 (2006), Nr. 4, S. 573–579

[Binetti u. a. 2008] Binetti, Roberto ; Costamagna, Francesca M. ; Marcello, Ida: Exponential growth of new chemicals and evolution of information relevant to risk control. In: *ANNALI-ISTITUTO SUPERIORE DI SANITA* 44 (2008), Nr. 1, S. 13

[Boser u. a. 1992] Boser, Bernhard E. ; Guyon, Isabelle M. ; Vapnik, Vladimir N.: A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory* ACM, 1992, S. 144–152

[Breiman 2001] Breiman, Leo: Random forests. In: *Machine learning* 45 (2001), Nr. 1, S. 5–32

[Breiman u. a. 1984] Breiman, Leo ; Friedman, Jerome ; Stone, Charles J. ; Olshen, Richard A.: *Classification and regression trees*. CRC press, 1984

[Bryne u. a. 2008] Bryne, Jan C. ; Valen, Eivind ; Tang, Man-Hung E. ; Marstrand, Troels ; Winther, Ole ; Piedade, Isabelle da ; Krogh, Anders ; Lenhard, Boris ; Sandelin, Albin: JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. In: *Nucleic acids research* 36 (2008), Nr. suppl 1, S. D102–D106

[Carlson] Carlson, Marc: *GO.db: A set of annotation maps describing the entire Gene Ontology*. – R package version 3.2.2

[Chambers u. a. 2009] Chambers, Stuart M. ; Fasano, Christopher A. ; Papapetrou, Eirini P. ; Tomishima, Mark ; Sadelain, Michel ; Studer, Lorenz: Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. In: *Nature biotechnology* 27 (2009), Nr. 3, S. 275–280

[Chen u. a. 2011] Chen, Chao ; Grennan, Kay ; Badner, Judith ; Zhang, Dandan ; Gershon, Elliot ; Jin, Li ; Liu, Chunyu: Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. In: *PloS one* 6 (2011), Nr. 2, S. e17238

[Delmar u. a. 2005] Delmar, Paul ; Robin, Stéphane ; Daudin, Jean J.: VarMixt: efficient variance modelling for the differential analysis of replicated gene expression data. In: *Bioinformatics* 21 (2005), Nr. 4, S. 502–508

[Díaz-Uriarte u. De Andres 2006] Díaz-Uriarte, Ramón ; De Andres, Sara A.: Gene selection and classification of microarray data using random forest. In: *BMC bioinformatics* 7 (2006), Nr. 1, S. 3

[Drăghici 2010] Drăghici, Sorin: *Data analysis tools for DNA microarrays*. CRC Press, 2010

[Gelman u. a. 2003] Gelman, Andrew ; Carlin, John B. ; Stern, Hal S. ; Rubin, Donald B.: *Bayesian Data Analysis*. CRC Press, 2003

[Golub u. a. 1999] Golub, Todd R. ; Slonim, Donna K. ; Tamayo, Pablo ; Huard, Christine ; Gaasenbeek, Michelle ; Mesirov, Jill P. ; Coller, Hilary ; Loh, Mignon L. ; Downing, James R. ; Caligiuri, Mark A. u. a.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. In: *science* 286 (1999), Nr. 5439, S. 531–537

[Guyon u. a. 2006] Guyon, Isabelle ; Gunn, Steve ; Nikravesh, Masoud ; Zadeh, L: Feature extraction. In: *Foundations and applications* (2006)

[Harris u. a. 2014] Harris, Catherine A. ; Scott, Alexander P. ; Johnson, Andrew C. ; Panter, Grace H. ; Sheahan, Dave ; Roberts, Mike ; Sumpter, John P.: Principles of Sound Ecotoxicology. In: *Environmental science & technology* 48 (2014), Nr. 6, S. 3100–3111

[Hartung u. a. 2005] Hartung, Joachim ; Elpelt, Bärbel ; Klösener, Karl-Heinz: *Statistik: Lehr-und Handbuch der angewandten Statistik; mit zahlreichen, vollständig durchgerechneten Beispielen*. Oldenbourg Verlag, 2005

[Hastie u. a. 2011] Hastie, Trevor J. ; Tibshirani, Robert J. ; Friedman, Jerome H.: *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2011

[Irizarry u. a. 2003a] Irizarry, Rafael A. ; Bolstad, Benjamin M. ; Collin, Francois ; Cope, Leslie M. ; Hobbs, Bridget ; Speed, Terence P.: Summaries of Affymetrix GeneChip probe level data. In: *Nucleic acids research* 31 (2003), Nr. 4, S. e15–e15

[Irizarry u. a. 2003b] Irizarry, Rafael A. ; Hobbs, Bridget ; Collin, Francois ; Beazer-Barclay, Yasmin D. ; Antonellis, Kristen J. ; Scherf, Uwe ; Speed, Terence P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. In: *Biostatistics* 4 (2003), Nr. 2, S. 249–264

[Jagtap u. a. 2011] Jagtap, S ; Meganathan, K ; Gaspar, J ; Wagh, V ; Winkler, J ; Hescheler, J ; Sachinidis, A: Cytosine arabinoside induces ectoderm and inhibits mesoderm expression in human embryonic stem cells during multilineage differentiation. In: *British journal of pharmacology* 162 (2011), Nr. 8, S. 1743–1756

[Jeanmougin u. a. 2010] Jeanmougin, Marine ; De Reynies, Aurélien ; Marisa, Laetitia ; Paccard, Caroline ; Nuel, Gregory ; Guedj, Mickael: Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. In: *PloS one* 5 (2010), Nr. 9, S. e12336

[Johnson u. a. 2007] Johnson, W E. ; Li, Cheng ; Rabinovic, Ariel: Adjusting batch effects in microarray expression data using empirical Bayes methods. In: *Biostatistics* 8 (2007), Nr. 1, S. 118–127

[Krug u. a. 2013] Krug, Anne K. ; Kolde, Raivo ; Gaspar, John A. ; Rempel, Eugen ; Balmer, Nina V. ; Meganathan, Kesavan ; Vojnits, Kinga ; Baquié, Mathurin ; Waldmann, Tanja ; Ensenat-Waser, Roberto u. a.: Human embryonic stem cell-derived test systems for developmental neurotoxicity: a transcriptomics approach. In: *Archives of toxicology* 87 (2013), Nr. 1, S. 123–143

[Kumar u. a. 2015] Kumar, M. ; Rath, N.K. ; Swain, A. ; Rath, S.K.: Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor, 2015, 301-310. – cited By 0

[Lazar u. a. 2012] Lazar, Cosmin ; Meganck, Stijn ; Taminau, Jonatan ; Steenhoff, David ; Coletta, Alain ; Molter, Colin ; Weiss-Solís, David Y. ; Duque, Robin ; Bersini, Hugues ; Nowé, Ann: Batch effect removal methods for microarray gene expression data integration: a survey. In: *Briefings in bioinformatics* (2012), S. bbs037

[Lee u. a. 2011] Lee, Chien-Pang ; Lin, Wen-Shin ; Chen, Yuh-Min ; Kuo, Bo-Jein: Gene selection and sample classification on microarray data based on adaptive genetic algorithm/k-nearest neighbor method. In: *Expert Systems with Applications* 38 (2011), Nr. 5, 4661 - 4667. `http://dx.doi.org/http://dx.doi.org/10.1016/j.eswa.2010.07.053`. – DOI http://dx.doi.org/10.1016/j.eswa.2010.07.053. – ISSN 0957–4174

[Leek u. a. 2010] Leek, Jeffrey T. ; Scharpf, Robert B. ; Bravo, Héctor C. ; Simcha, David ; Langmead, Benjamin ; Johnson, W E. ; Geman, Donald ; Baggerly, Keith ; Irizarry, Rafael A.: Tackling the widespread and critical impact of batch effects in high-throughput data. In: *Nature Reviews Genetics* 11 (2010), Nr. 10, S. 733–739

[Leek u. Storey 2007] Leek, Jeffrey T. ; Storey, John D.: Capturing heterogeneity in gene expression studies by surrogate variable analysis. In: *PLoS Genet* 3 (2007), Nr. 9, S. e161

[McLaren 2001] McLaren, Anne: Ethical and social considerations of stem cell research. In: *Nature* 414 (2001), Nr. 6859, S. 129–131

[Meganathan u. a. 2012] Meganathan, Kesavan ; Jagtap, Smita ; Wagh, Vilas ; Winkler, Johannes ; Gaspar, John A. ; Hildebrand, Diana ; Trusch, Maria ; Lehmann, Karola ; Hescheler, Jürgen ; Schlüter, Hartmut u. a.: Identification of thalidomide-specific transcriptomics and proteomics signatures during differentiation of human embryonic stem cells. In: *PloS one* 7 (2012), Nr. 8, S. e44228

[Murie u. a. 2009] Murie, Carl ; Woody, Owen ; Lee, Anna Y. ; Nadon, Robert: Comparison of small n statistical tests of differential expression applied to microarrays. In: *BMC bioinformatics* 10 (2009), Nr. 1, S. 45

[Platt 1999] Platt, John C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in large margin classifiers* Citeseer, 1999

[Quinlan 1986] Quinlan, J. R.: Induction of decision trees. In: *Machine learning* 1 (1986), Nr. 1, S. 81–106

[R Core Team 2013] R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2013. `http://www.R-project.org/`. – ISBN 3-900051-07-0

[Rajashekaran u. Vijayalksmi 2004] Rajashekaran, S ; Vijayalksmi, GA: *Neural Networks, Fuzzy Logic and Genetic Algorithms*. Prentice-Hall of India Pvt. Ltd, 2004

[Rempel u. a. 2015] Rempel, Eugen ; Hoelting, Lisa ; Waldmann, Tanja ; Balmer, Nina V. ; Schildknecht, Stefan ; Grinberg, Marianna ; Gaspar, John Antony D. ; Shinde, Vaibhav ; Stöber, Regina ; Marchan, Rosemarie u. a.: A transcriptome-based classifier to identify developmental toxicants by stem cell testing: design, validation and optimization for histone deacetylase inhibitors. In: *Archives of toxicology* 89 (2015), Nr. 9, S. 1599–1618

[Ripley u. Venables 1994] Ripley, Brian D. ; Venables, William N.: *Modern applied statistics with S-Plus*. Springer-Verlag New York, NY, 1994

[Sammut u. Webb 2011] Sammut, Claude ; Webb, Geoffrey I.: *Encyclopedia of machine learning*. Springer, 2011

[Scherer 2009] Scherer, Andreas: *Batch effects and noise in microarray experiments: sources and solutions*. Bd. 868. John Wiley & Sons, 2009

[Smyth u. a. 2004] Smyth, Gordon K. u. a.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. In: *Stat Appl Genet Mol Biol* 3 (2004), Nr. 1, S. 3

[Sreepada u. a. 2014] Sreepada, Rama S. ; Vipsita, Swati ; Mohapatra, Puspanjali: An efficient approach for classification of gene expression microarray data. In: *Emerging Applications of Information Technology (EAIT), 2014 Fourth International Conference of* IEEE, 2014, S. 344–348

[Sui u. a. 2007] Sui, Shannan J H. ; Fulton, Debra L. ; Arenillas, David J. ; Kwon, Andrew T. ; Wasserman, Wyeth W.: oPOSSUM: integrated tools for analysis of regulatory motif over-representation. In: *Nucleic acids research* 35 (2007), Nr. suppl 2, S. W245–W252

[Tamm u. a. 2006] Tamm, Christoffer ; Duckworth, Joshua ; Hermanson, Ola ; Ceccatelli, Sandra: High susceptibility of neural stem cells to methylmercury toxicity: effects on cell survival and neuronal differentiation. In: *Journal of neurochemistry* 97 (2006), Nr. 1, S. 69–78

[van Thriel u. a. 2012] Thriel, Christoph van ; Westerink, Remco H. ; Beste, Christian ; Bale, Ambuja S. ; Lein, Pamela J. ; Leist, Marcel: Translating neurobehavioural endpoints of developmental neurotoxicity tests into< i> in vitro</i> assays and readouts. In: *Neurotoxicology* 33 (2012), Nr. 4, S. 911–924

[Tusher u. a. 2001] Tusher, Virginia G. ; Tibshirani, Robert ; Chu, Gilbert: Significance analysis of microarrays applied to the ionizing radiation response. In: *Proceedings of the National Academy of Sciences* 98 (2001), Nr. 9, S. 5116–5121

[Waldmann u. a. 2014] Waldmann, Tanja ; Rempel, Eugen ; Balmer, Nina V. ; König, André ; Kolde, Raivo ; Gaspar, John A. ; Henry, Margit ; Hescheler, Jürgen ; Sachinidis, Agapios ; Rahnenführer, Jörg u. a.: Design principles of concentration-dependent transcriptome deviations in drug-exposed differentiating stem cells. In: *Chemical research in toxicology* 27 (2014), Nr. 3, S. 408–420

[Watkins u. a. 2010] Watkins, John B. ; Klaassen, Curtis D. ; Acosta, Daniel: *Casarett & Doull's essentials of toxicology.* McGraw-Hill, 2010

[Weinstein 2008] Weinstein, John N.: Biochemistry. A postgenomic visual icon. In: *Science (New York, NY)* 319 (2008), Nr. 5871, S. 1772–1773

[Weinstein u. a. 1997] Weinstein, John N. ; Myers, Timothy G. ; O'Connor, Patrick M. ; Friend, Stephen H. ; Fornace, Albert J. ; Kohn, Kurt W. ; Fojo, Tito ; Bates, Susan E. ; Rubinstein, Lawrence V. ; Anderson, N L. u. a.: An information-intensive approach to the molecular pharmacology of cancer. In: *Science* 275 (1997), Nr. 5298, S. 343–349

[Wingender 2008] Wingender, Edgar: The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. In: *Briefings in bioinformatics* 9 (2008), Nr. 4, S. 326–332

[Witten u. Tibshirani 2007] Witten, Daniela ; Tibshirani, Robert: A comparison of fold-change and the t-statistic for microarray data analysis. In: *Technical Report, Stanford University.* (2007)

[Wolpert u. Macready 1997] Wolpert, David H. ; Macready, William G.: No free lunch theorems for optimization. In: *Evolutionary Computation, IEEE Transactions on* 1 (1997), Nr. 1, S. 67–82

Anhang

Im Anhang sind Hintegrundinformationen zu den verwendeten statistischen Methoden zu finden. Ferner sind Abbildungen und Tabellen dargestellt, die Übersichtlichkeit halber nicht in den Hauptteil aufgenommen werden. Zuletzt wird der verwendeter R-Code präsentiert.

6.1 Theoretischer Hintergrund

In diesem Abschnitt werden die theoretischen Grundlagen von Receiver Operating Characteristic und dem AUC-Wert erörtert. Ferner werden die Entscheidungsbäume kurz vorgestellt. Die inverse Gammaverteilung wird präsentiert und ihre ersten Momente berechnet. Zuletzt werden bestimmte im Hauptteil aufgestellte Behauptungen bewiesen.

6.1.1 Inverse Gammaverteilung

In diesem Unterkapitel wird die inverse Gammaverteilung vorgestellt. Die inverse Gammaverteilung ist eine kontinuierliche Wahrscheinlichkeitsverteilung über die Menge der positiven reellen Zahlen. Sie ist die Verteilung vom Kehrwert einer gammaverteilten Zufallsvariable. So sei die Variable X gammaverteilt mit Parametern α und β. Dann lautet die Dichtefunktion für $x > 0$ wie folgt

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right),$$

wobei mit Γ die Gamma-Funktion

$$\Gamma(x) = \int_0^\infty t^{x-1} \exp(-t) \mathrm{d}t$$

bezeichnet wird. Wir bestimmen nun die Dichtefunktion der Variablen $Y = \frac{1}{X}$ mit Hilfe der Transformationssatzes. Es gilt

$$\begin{aligned} f_Y(y) &= f_X\left(\frac{1}{y}\right) \left| \frac{\mathrm{d}}{\mathrm{d}y} y^{-1} \right| \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{-\alpha+1} \exp\left(-\frac{1}{y\beta}\right) y^{-2} \\ &= \frac{\beta^{-\alpha}}{\Gamma(\alpha)} y^{-\alpha-1} \exp\left(-\frac{1}{y\beta}\right). \end{aligned}$$

Man sagt, die Variable Y folge einer inversen Gammaverteilung mit den Parametern α und $\frac{1}{\beta}$:

$$Y \sim \Gamma^{-1}(\alpha, \frac{1}{\beta}).$$

Wir bestimmen nun die Momente der Zufallsvariablen $X \sim \Gamma^{-1}(\alpha, \beta)$ wie folgt: für $\alpha > n$ gilt mit der Substitution

$$t = \frac{\beta}{x} \Rightarrow \mathrm{d}x = -\beta t^{-2}\mathrm{d}t$$

folgende Relation

$$\begin{aligned}
\mathrm{E}(X^n) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^n x^{-\alpha-1} \exp\left(-\frac{x}{\beta}\right)\mathrm{d}x \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{n-\alpha-1} \exp\left(-\frac{x}{\beta}\right)\mathrm{d}x \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\beta^{\alpha-n}} t^{\alpha-n-1} \exp(-t)\mathrm{d}t \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \\
&= \frac{\beta^n}{(\alpha-1)(\alpha-2)\cdots(\alpha-n)}.
\end{aligned}$$

Somit gilt für $\alpha > 1$ die Beziehung

$$\mathrm{E}(X) = \frac{\beta}{\alpha-1} \tag{6.1}$$

und für $\alpha > 2$ die Identität

$$\mathrm{E}(X^2) = \frac{\beta^2}{(\alpha-1)(\alpha-2)} \tag{6.2}$$

und somit

$$\mathrm{var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}. \tag{6.3}$$

6.1.2 Receiver Operating Characteristic und AUC-Wert

In diesem Unterkapitel erläutern wir kurz die Grundzüge der Receiver Operating Characteristic (ROC)-Kurve. Die ROC-Kurve stellt visuell die Abhängigkeit der Performance eines Tests bzw. eines Lernverfahrens von einem bestimmten Parametern dar und dient somit der Bewertung. Für jeden möglichen Parameterwert bestimmt man die beiden Maßzahlen: Sensitivität und Spezifität. Dazu erstellt man zuerst eine Kontingenztafel mit zwei binären nominalen Variablen: der Vorhersage des Lernverfahrens und der tatsächlichen Klasse.

Basierend auf den berechneten Häufigkeiten werden die beiden Maßzahlen wie folgt berechnet:

$$\text{Sensitivität} = P(\text{positiv erkannt}|\text{tatsächlich positiv}) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{6.4}$$

$$\text{Spezifität} = P(\text{negativ erkannt}|\text{tatsächlich negativ}) = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}. \tag{6.5}$$

	D+	D-
T+	TP	FP
T-	FN	TN

Tabelle 18: Kontingenztafel mit Variablen Vorhersage des Verfahrens (Ausprägungen: Test positiv und Test negativ) und tatsächliche Klasse (Ausprägungen: Diagnose positiv und Diagnose negative). Es treten 4 Fälle auf: richtig positiv (TP), richtig negativ (TN), falsch positiv (FP) und falsch negativ (FN).

Nun werden in einem Diagramm Sensitivität gegen die Falsch-Positiv-Rate abgetragen. Dabei wird die Falsch-Positiv-Rate als $1 - \text{Spezifität}$ berechnet. Es resultiert eine aufsteigende Kurve, siehe Abbildung 21 auf der Seite 57.

Eine ROC-Kurve nahe der Winkelhalbierenden deutet auf einen Zufallsprozess hin: es bedeutet, dass für jeden Parameterwert die Richtig-Positiv-Rate der Falsch-Positiv-Rate gleicht, was einem zufälligen Raten entspricht. Ist das Lernverfahren in der Lage für einen bestimmten Parameter die beiden Klassen perfekt von einander zu trennen, bedeutet dies, dass Sensitivität den Wert eins einnimmt, während die Falsch-Positiv-Rate gleich Null ist. Visuell heißt es, dass die ROC-Kurve senkrecht auf den Punkt (0,1) ansteigt und dann parallel zu der x-Achse bis zum Punkt (1,1) verläuft.

Zu einer ROC-Kurve lässt sich die Fläche unterhalb der Kurve berechnen. Dieser Wert wird als Area under Curve (AUC) bezeichnet. Er kann zwischen 0 und 1 liegen, wobei der Wert von 0.5 insofern am schlechtesten ist, als er auf ein zufälliges Raten hindeutet. Der Wert 1 bedeutet, dass die ROC-Kurve durch den Punkt (0,1) verläuft, und das Verfahren somit die beiden Klassen perfekt trennt. Der AUC-Wert lässt auch als die Wahrscheinlichkeit interpretieren, dass ein zufällig gezogenes Paar von einer negativen und einer positiven Probe insofern richtig klassifiziert wird, als der zur Vorhersage genommenes Parameter bei der positiven Probe höher als bei der negativen liegt.

6.1.3 Entscheidungsbäume

Ein Entscheidungsbaum ist ein Klassifizierungsmodell, welches in seinem Aufbau einem nach unten wachsenden Baum ähnelt. Zuordnung eines neuen Objektes fängt von dem obersten Knoten - die Wurzel - an, wobei man den Wert der Variablen betrachtet, welche zu der Wurzel korrespondiert. Man wandert dann entlang des Astes, welcher dem bestimmten Wert der Variablen zugeordnet ist, und kommt bei einem neuen Knoten mit einer anderen Variablen an. Dieses Vorgehen wird dann so lange wiederholt, bis man an dem Endknoten - dem Blatt - angelangt ist. Das Blatt weist dann das Objekt einer Klasse zu.

Die Entscheidungsbäume werden vom Wurzel aus aufgebaut. Dazu steht *Rekursive Partitionierung* als Algorithmus zur Verfügung. Das Verfahren wählt die beste Variable für die Wurzel, partitioniert die Trainingsmenge und fügt dem Baum die korrespondierende

Knoten und Äste zu. Entsprechend einem **top-down** Prinzip werden nun die Knoten betrachtet, die auf einer tieferen Ebene liegen. Bei jedem Schritt wird die Variable bzw. das Attribut gesucht, mit welchem sich die Objekte am besten klassifizieren lassen.

Typische Kriterien zur Attributauswahl verwendet eine Funktion zur Messung der Heterogenität des Knotens, d.h. ob ein Knoten nur Repräsentanten einer Klasse enthält. Die zwei bekannten Heterogenitätsmaßen sind Entropie [Quinlan (1986)] und Gini-Koeffizient [Breiman u. a. (1984)], welche wie folgt definiert sind:

$$\text{Entropie}(T) = -\sum_{i}^{c} \frac{|T_i|}{T} \cdot \log_2\left(\frac{|T_i|}{T}\right)$$

$$\text{Gini}(T) = 1 - \sum_{i}^{c}\left(\frac{|T_i|}{T}\right)^2,$$

wobei T die Trainingsmenge ist, und T_i die Trainingsmenge der Objekte der Klasse c_i ist.

Demzufolge ist der endgültige Baum als Ergebnis lokaler Selektionen aufgebaut, welche nur einen Teil der Trainingsmenge betrachten. Dies kann keinen globalen Optimum garantieren. Demgegenüber liegt die Schnelligkeit des Verfahrens.

6.1.4 Lemmata

Lemma 1. *Es gilt die Beziehung 3.7. Dies bedeutet, dass für alle $c > 0$ gilt*

$$X \sim \chi^2_\nu \Rightarrow cX \sim \Gamma(k = \frac{\nu}{2}, \Theta = 2c).$$

Beweis. Es gelte die Relation:

$$X \sim \chi^2_\nu.$$

Somit lautet die Dichte von X wie folgt

$$f(x) = \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \exp(-\frac{x}{2}).$$

Somit ist für die Verteilungsfunktion der Variablen cX die folgende Gleichung erfüllt:

$$\begin{aligned} F_{cX}(t) = P(cX \leq t) = P(X \leq \frac{t}{c}) &= \int_0^{\frac{t}{c}} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \exp(-\frac{x}{2})\, dx \\ &= \int_0^t \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2}) c^{\frac{\nu}{2}-1}} y^{\frac{\nu}{2}-1} \exp(-\frac{y}{2c}) \frac{1}{c} dy \\ &= \int_0^t \frac{1}{(2c)^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})} y^{\frac{\nu}{2}-1} \exp(-\frac{y}{2c})\, dy. \end{aligned}$$

Dabei wurde die Transformation $y = cx$ verwendet. Der letzte Intergrand stellt dabei die Dichte einer gammaverteilten Zufallsvariablen mit den Parametern $k = \frac{\nu}{2}$ und $\Theta = 2c$. Somit gilt für die Variable cX die Behauptung. □

Lemma 2. *Es gelten die Beziehungen 3.26 und 3.27, d.h. die a-posteriori Erwartungswerte der Parameter γ_{jk} und δ^2_{jk} lauten:*

$$E(\gamma_{jk} \mid Z_{jkl}, \delta^2_{jk}) = \frac{\tau_k^2 \sum_l Z_{jkl} + \delta^2_{jk}\gamma_k}{n_k\tau_k^2 + \delta^2_{jk}} \tag{6.6}$$

$$E(\delta^2_{jk} \mid Z_{jkl}, \gamma_{jk}) = \frac{\Theta_k + \frac{1}{2}\sum_l (Z_{jkl} - \gamma_{jk})^2}{\frac{n_k}{2} + \lambda_k - 1}. \tag{6.7}$$

Beweis. Es gilt nach dem Satz von Bayes für die a-posteriori Verteilung von γ_{jk} folgende Beziehung

$$\begin{aligned} p(\gamma_{jk} \mid Z_{jkl}, \delta^2_{jk}) &\propto p(Z_{jkl} \mid \gamma_{jk}, \delta^2_{jk}) p(\gamma_{jk}) \\ &\propto \exp(-\frac{1}{2\delta^2_{jk}} \sum_l (Z_{jkl} - \gamma_{jk})^2) \exp(-\frac{1}{2\tau^2}(\gamma_{jk} - \gamma_k)^2) \\ &= \exp\Big(-\frac{1}{2\delta^2_{jk}}\big(\sum_l Z^2_{jkl} - 2\sum_l Z_{jkl}\gamma_{jk} + n_k\gamma_{jk}\big) - \frac{1}{2\tau^2}(\gamma^2_{jk} - 2\gamma_{jk}\gamma_k + \gamma_k^2)\Big) \\ &\propto \Big(-\frac{1}{2}\Big(\frac{n_k\tau_k^2 + \delta^2_{jk}}{\delta^2_{jk}\tau_k^2}\Big)\Big(\gamma^2_{jk} - 2\Big(\frac{\tau_k^2 \sum_l Z_{jkl} + \delta^2_{jk}\gamma_k}{n_k\tau_k^2 + \delta^2_{jk}}\Big)\gamma_{jk}\Big)\Big). \end{aligned}$$

Mit Hilfe der quadratischen Ergänzung lässt sich der letzte Ausdruck als Dichtefunktion einer Normalverteilung mit dem Erwartungswert

$$E(\gamma_{jk} \mid Z_{jkl}, \delta_{jk}^2) = \frac{\tau_k^2 \sum_l Z_{jkl} + \delta_{jk}^2 \gamma_k}{n_k \tau_k^2 + \delta_{jk}^2}$$

erkennen. Analog lässt sich nach dem Satz von Bayes für die a-posteriori Verteilung von δ_{jk}^2 folgende Beziehung herleiten

$$\begin{aligned} p(\delta_{jk}^2 \mid Z_{jkl}, \gamma_{jk}) &\propto p(Z_{jkl} \mid \gamma_{jk}, \delta_{jk}^2) p(\delta_{jk}^2) \\ &\propto (\delta_{jk}^2)^{-\frac{n_k}{2}} \exp(-\frac{1}{2\delta_{jk}^2} \sum_l (Z_{jkl} - \gamma_{jk})^2)(\delta_{jk}^2)^{-(\lambda_k+1)} \exp\left(-\frac{\Theta_k}{\delta_{jk}^2}\right) \\ &= (\delta_{jk}^2)^{-(\frac{n_k}{2}+\lambda_k)-1} \exp\left(-\frac{\Theta_k + \frac{1}{2}\sum_l (Z_{jkl} - \gamma_{jk})^2}{\delta_{jk}^2}\right). \end{aligned}$$

Die hergeleitete Formel ist nun die Dichte einer invers gammaverteilten Größe mit Formparameter $\frac{n_k}{2} + \lambda_k$ und Skalenparameter $\Theta_k + \frac{1}{2}\sum_l (Z_{jkl} - \gamma_{jk})^2$. Der a-posteriori Erwartungswertet lautet

$$E(\delta_{jk}^2 \mid Z_{jkl}, \gamma_{jk}) = \frac{\Theta_k + \frac{1}{2}\sum_l (Z_{jkl} - \gamma_{jk})^2}{\frac{n_k}{2} + \lambda_k - 1}$$

□

Lemma 3. *Es gelten die Beziehungen* (3.32) *und* (3.33)*, d.h es gilt:*

$$\begin{aligned} \bar{\lambda}_k &= \frac{\bar{V}_k + 2\bar{S}_k^2}{\bar{S}_k^2} \\ \bar{\Theta}_k &= \frac{\bar{V}_k^3 + \bar{V}_k \bar{S}_k^2}{\bar{S}_k^2}. \end{aligned}$$

Beweis. Setzt man die empirischen Momente $\bar{V}_k$ und $\bar{S}_k^2$ den theoretischen Momenten einer inversen Gammaverteilung $\Gamma^{-1}(\lambda_k, \Theta_k)$ gleich, erhält man folgendes Gleichungssystem:

$$\begin{aligned} \bar{V}_k &= \frac{\Theta_k}{\lambda_k - 1} \\ \bar{S}_k^2 &= \frac{\Theta_k^2}{(\lambda_k - 1)^2(\lambda_k - 2)}. \end{aligned}$$

Löst man die erste Gleichung nach Θ_k auf, erhält man

$$\Theta_k = \bar{V}_k(\lambda_k - 1).$$

Setzt man diese Beziehung in die zweite Gleichung, vereinfacht sich diese zu

$$\bar{S}_k^2 = \frac{\bar{V}_k^2}{\lambda_k - 2}.$$

Nach λ_k aufgelöst, erhält man den Schätzer:

$$\bar{\lambda}_k = \frac{\bar{V}_k + 2\bar{S}_k^2}{\bar{S}_k^2}.$$

Setzt man dies in die Gleichung für Θ_k, bekommt man den entsprechenden Schätzer (3.33). □

Lemma 4. *Sei X eine binomialverteilte Zufallsgröße mit Erfolgswahrscheinlichkeit $\Theta \in (0,1)$. In n Versuchen werden k Erfolge beobachtet. Wird für Θ eine $\mathrm{B}(\alpha,\beta)$-Verteilung als a-priori Verteilung verwendet, so ist die a-posteriori Verteilung eine $\mathrm{B}(\alpha+k, \beta+n-k)$-Verteilung.*

Beweis. Es gilt nach dem Satz von Bayes:

$$\begin{aligned}
P(\Theta|n,k,\alpha,\beta) &= \frac{P(k|n,\Theta)P(\Theta|n,\alpha,\beta)}{P(k|n,\alpha,\beta} \\
&\sim P(k|n,\Theta)P(\Theta|n,\alpha,\beta) \\
&= P(k|n,\Theta)P(\Theta|\alpha,\beta) \\
&= \frac{n!}{k!(n-k)!}\Theta^k(1-\Theta)^{n-k} \times \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\Theta^{\alpha-1}(1-\Theta)^{\beta-1} \\
&\sim \Theta^k(1-\Theta)^{n-k} \times \Theta^{\alpha-1}(1-\Theta)^{\beta-1} \\
&= \Theta^{k+\alpha-1}(1-\Theta)^{n-k+\beta-1}.
\end{aligned}$$

In der letzten Zeile erkennen wir den Kern einer Dichte von betaverteilten Zufallsvariablen mit den Parametern $k+\alpha$ und $n-k+\beta$. □

Lemma 5. *Es gilt die Gleichung 3.42. Dies bedeutet, dass für die Varianz des Durchschnitts von B gleichverteilten Zufallsvariablen X_i, welche mit Korrelationskoeffizienten ρ paarweise korreliert sind die Gleichung*

$$Var(\frac{1}{B}\sum_{i=1}^{B} X_i) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

gilt, wobei σ^2 die Varianz von X_i für $i = 1, ..., B$ ist.

Beweis. Es gilt für die linke Seite:

$$\begin{aligned} Var(\frac{1}{B}\sum_{i=1}^{B} X_i) &= \frac{1}{B^2} Var(\sum_{i=1}^{B} X_i) = \frac{1}{B^2}\sum_{i=1}^{B} Var(X_i) + \frac{2}{B^2}\sum_{1\le i<j\le B}\sum Cov(X_i, X_j) \\ &= \frac{1}{B^2} B\sigma^2 + \frac{2}{B^2}\sum_{1\le i<j\le B}\sum \rho\sigma^2 = \frac{\sigma^2}{n} + \frac{2\rho\sigma^2}{B^2}\sum_{1\le i<j\le B}\sum 1 \\ &= \frac{\sigma^2}{B} + \frac{2\rho\sigma^2}{B^2}\frac{B(B-1)}{2} = \frac{\sigma^2}{B} + \rho\sigma^2\frac{B-1}{B} = \frac{\sigma^2}{B} + \rho\sigma^2 - \frac{\rho\sigma^2}{B} \\ &= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \end{aligned}$$

Somit ist die Behauptung bewiesen. □

6.2 Tabellen

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Studie
UKN1_sat_c_VPA_1000μ M_D12_25NS_141.CEL	D12_25NS	VPA	VPA 1 mM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_1000μ M_D12_26NS_142.CEL	D12_26NS	VPA	VPA 1 mM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_1000μ M_D12_27NS_143.CEL	D12_27NS	VPA	VPA 1 mM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_800μ M_D12_25NS_144.CEL	D12_25NS	VPA	VPA 800 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_800μ M_D12_26NS_145.CEL	D12_26NS	VPA	VPA 800 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_800μ M_D12_27NS_146.CEL	D12_27NS	VPA	VPA 800 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_650μ M_D12_25NS_147.CEL	D12_25NS	VPA	VPA 650 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_650μ M_D12_26NS_148.CEL	D12_26NS	VPA	VPA 650 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_650μ M_D12_27NS_149.CEL	D12_27NS	VPA	VPA 650 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_550μ M_D12_25NS_150.CEL	D12_25NS	VPA	VPA 550 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_550μ M_D12_26NS_151.CEL	D12_26NS	VPA	VPA 550 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_550μ M_D12_27NS_152.CEL	D12_27NS	VPA	VPA 550 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_450μ M_D12_25NS_153.CEL	D12_25NS	VPA	VPA 450 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_450μ M_D12_26NS_154.CEL	D12_26NS	VPA	VPA 450 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_450μ M_D12_27NS_155.CEL	D12_27NS	VPA	VPA 450 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_350μ M_D12_25NS_156.CEL	D12_25NS	VPA	VPA 350 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_350μ M_D12_26NS_157.CEL	D12_26NS	VPA	VPA 350 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_350μ M_D12_27NS_158.CEL	D12_27NS	VPA	VPA 350 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_150μ M_D12_25NS_159.CEL	D12_25NS	VPA	VPA 150 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_150μ M_D12_26NS_160.CEL	D12_26NS	VPA	VPA 150 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_150μ M_D12_27NS_161.CEL	D12_27NS	VPA	VPA 150 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_25μ M_D12_25NS_162.CEL	D12_25NS	VPA	VPA 25 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_25μ M_D12_26NS_163.CEL	D12_26NS	VPA	VPA 25 μM	DoD0-6	VPA sat conc
UKN1_sat_c_VPA_25μ M_D12_27NS_164.CEL	D12_27NS	VPA	VPA 25 μM	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_25NS_165.CEL	D12_25NS	untreated	untreated	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_26NS_166.CEL	D12_26NS	untreated	untreated	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_27NS_167.CEL	D12_27NS	untreated	untreated	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_25NS_168.CEL	D12_25NS	untreated	untreated	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_26NS_169.CEL	D12_26NS	untreated	untreated	DoD0-6	VPA sat conc
UKN1_sat_c_untr_D12_27NS_170.CEL	D12_27NS	untreated	untreated	DoD0-6	VPA sat conc

Tabelle 19: VPA Chronic Konzentrationsstudie

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Studie
UKN1_sat_a_VPA_5mM_D12_31NS_195.CEL	D12_31NS	VPA	VPA 5 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_4mM_D12_31NS_196.CEL	D12_31NS	VPA	VPA 4 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_2mM_D12_31NS_197.CEL	D12_31NS	VPA	VPA 2 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_1mM_D12_31NS_198.CEL	D12_31NS	VPA	VPA 1 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_0,6mM_D12_31NS_199.CEL	D12_31NS	VPA	VPA 0.6 mM	DoD4-6	VPA sat acute
UKN1_sat_a_untr_D12_31NS_200.CEL	D12_31NS	untreated	untreated	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_5mM_D12_32NS_201.CEL	D12_32NS	VPA	VPA 5 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_4mM_D12_32NS_202.CEL	D12_32NS	VPA	VPA 4 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_2mM_D12_32NS_203.CEL	D12_32NS	VPA	VPA 2 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_1mM_D12_32NS_204.CEL	D12_32NS	VPA	VPA 1 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_0,6mM_D12_32NS_205.CEL	D12_32NS	VPA	VPA 0.6 mM	DoD4-6	VPA sat acute
UKN1_sat_a_untr_D12_32NS_206.CEL	D12_32NS	untreated	untreated	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_5mM_D12_TW_207.CEL	D12_TW	VPA	VPA 5 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_4mM_D12_TW_208.CEL	D12_TW	VPA	VPA 4 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_2mM_D12_TW_209.CEL	D12_TW	VPA	VPA 2 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_1mM_D12_TW_210.CEL	D12_TW	VPA	VPA 1 mM	DoD4-6	VPA sat acute
UKN1_sat_a_VPA_0,6mM_D12_TW_211.CEL	D12_TW	VPA	VPA 0.6 mM	DoD4-6	VPA sat acute
UKN1_sat_a_untr_D12_TW_212.CEL	D12_TW	untreated	untreated	DoD4-6	VPA sat acute

Tabelle 20: VPA Acute Konzentrationsstudie

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Studie
UKN1_sat_c___MeHg_3uM_D12_32NS_171.CEL	D12_32NS	MeHg	MeHg 3 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_2uM_D12_32NS_172.CEL	D12_32NS	MeHg	MeHg 2 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,8uM_D12_32NS_173.CEL	D12_32NS	MeHg	MeHg 1.8 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,6uM_D12_32NS_174.CEL	D12_32NS	MeHg	MeHg 1.6 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,4uM_D12_32NS_175.CEL	D12_32NS	MeHg	MeHg 1.4 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_0,25uM_D12_32NS_176.CEL	D12_32NS	MeHg	MeHg 0.25 μM	DoD0-6	MeHg sat conc
UKN1_sat_c_EtOH(MeHg solvent)_D12_32NS_177.CEL	D12_32NS	EtOH	EtOH	DoD0-6	MeHg sat conc
UKN1_sat_c_untr_D12_32NS_178.CEL	D12_32NS	untreated	untreated	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_3uM_D12_TW5_179.CEL	D12_TW5	MeHg	MeHg 3 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_2uM_D12_TW5_180.CEL	D12_TW5	MeHg	MeHg 2 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,8uM_D12_TW5_181.CEL	D12_TW5	MeHg	MeHg 1.8 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,6uM_D12_TW5_182.CEL	D12_TW5	MeHg	MeHg 1.6 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,4uM_D12_TW5_183.CEL	D12_TW5	MeHg	MeHg 1.4 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_0,25uM_D12_TW5_184.CEL	D12_TW5	MeHg	MeHg 0.25 μM	DoD0-6	MeHg sat conc
UKN1_sat_c_EtOH(MeHg solvent)_D12_TW5_185.CEL	D12_TW5	EtOH	EtOH	DoD0-6	MeHg sat conc
UKN1_sat_c_untr_D12_TW5_186.CEL	D12_TW5	untreated	untreated	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_3uM_D12_TW6_187.CEL	D12_TW6	MeHg	MeHg 3 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_2uM_D12_TW6_188.CEL	D12_TW6	MeHg	MeHg 2 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,8uM_D12_TW6_189.CEL	D12_TW6	MeHg	MeHg 1.8 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,6uM_D12_TW6_190.CEL	D12_TW6	MeHg	MeHg 1.6 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_1,4uM_D12_TW6_191.CEL	D12_TW6	MeHg	MeHg 1.4 μM	DoD0-6	MeHg sat conc
UKN1_sat_c___MeHg_0,25uM_D12_TW6_192.CEL	D12_TW6	MeHg	MeHg 0.25 μM	DoD0-6	MeHg sat conc
UKN1_sat_c_EtOH(MeHg solvent)_D12_TW6_193.CEL	D12_TW6	EtOH	EtOH	DoD0-6	MeHg sat conc
UKN1_sat_c_untr_D12_TW6_194.CEL	D12_TW6	untreated	untreated	DoD0-6	MeHg sat conc

Tabelle 21: MeHg Chronic Konzentrationsstudie

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Studie
UKN1_sat_a_MeHg_40uM_D12_TW_213.CEL	D12_TW	MeHg	MeHg 40 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_20uM_D12_TW_214.CEL	D12_TW	MeHg	MeHg 20 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_15uM_D12_TW_215.CEL	D12_TW	MeHg	MeHg 15 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_10uM_D12_TW_216.CEL	D12_TW	MeHg	MeHg 10 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_1,5uM_D12_TW_217.CEL	D12_TW	MeHg	MeHg 1.5 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_untr_D12_TW_218.CEL	D12_TW	untreated	untreated	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_40uM_D12_NS_219.CEL	D12_NS	MeHg	MeHg 40 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_20uM_D12_NS_220.CEL	D12_NS	MeHg	MeHg 20 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_15uM_D12_NS_221.CEL	D12_NS	MeHg	MeHg 15 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_10uM_D12_NS_222.CEL	D12_NS	MeHg	MeHg 10 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_1,5uM_D12_NS_223.CEL	D12_NS	MeHg	MeHg 1.5 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_untr_D12_NS_224.CEL	D12_NS	untreated	untreated	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_20uM_D12_32NS_225.CEL	D12_32NS	MeHg	MeHg 20 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_15uM_D12_32NS_226.CEL	D12_32NS	MeHg	MeHg 15 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_10uM_D12_32NS_227.CEL	D12_32NS	MeHg	MeHg 10 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_MeHg_1,5uM_D12_32NS_228.CEL	D12_32NS	MeHg	MeHg 1.5 μM	DoD4-6	MeHg sat acute
UKN1_sat_a_untr_D12_32NS_229.CEL	D12_32NS	untreated	untreated	DoD4-6	MeHg sat acute

Tabelle 22: MeHg Acute Konzentrationsstudie

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Typ
UKN1_1_biom_A_MeHg_D11_21.CEL	D11_21	MeHg	MeHg 1,5 μM	DoD0-6	Hg+
UKN1_2_biom_A_MeHg_D11_22.CEL	D11_20 = 21	MeHg	MeHg 1,5 μM	DoD0-6	Hg+
UKN1_3_biom_A_MeHg_D11_23.CEL	D11_23	MeHg	MeHg 1,5 μM	DoD0-6	Hg+
UKN1_4_biom_A_MeHg_D11_24.CEL	D11_24	MeHg	MeHg 1,5 μM	DoD0-6	Hg+
UKN1_5_biom_A_MeHg_D11_25.CEL	D11_25	MeHg	MeHg 1,5 μM	DoD0-6	Hg+
UKN1_11_biom_A_EtOH(MeHg solvent)_D11_21.CEL	D11_21	EtOH	EtOH ctr	DoD0-6	Solvent
UKN1_12_biom_A_EtOH(MeHg solvent)_D11_22.CEL	D11_20 = 21	EtOH	EtOH ctr	DoD0-6	Solvent
UKN1_13_biom_A_EtOH(MeHg solvent)_D11_23.CEL	D11_23	EtOH	EtOH ctr	DoD0-6	Solvent
UKN1_14_biom_A_EtOH(MeHg solvent)_D11_24.CEL	D11_24	EtOH	EtOH ctr	DoD0-6	Solvent
UKN1_15_biom_A_EtOH(MeHg solvent)_D11_25.CEL	D11_25	EtOH	EtOH ctr	DoD0-6	Solvent
UKN1_16_biom_A_TSA_D11_21.CEL	D11_21	TSA	TSA 10 nM	DoD0-6	HDACi
UKN1_17_biom_A_TSA_D11_22.CEL	D11_22	TSA	TSA 10 nM	DoD0-6	HDACi
UKN1_18_biom_A_TSA_D11_23.CEL	D11_23	TSA	TSA 10 nM	DoD0-6	HDACi
UKN1_19_biom_A_TSA_D11_24.CEL	D11_24	TSA	TSA 10 nM	DoD0-6	HDACi
UKN1_20_biom_A_TSA_D11_25.CEL	D11_25	TSA	TSA 10 nM	DoD0-6	HDACi
UKN1_26_biom_A_DMSO(TSA solvent)_D11_21.CEL	D11_21	DMSO	DMSO ctr	DoD0-6	Solvent
UKN1_27_biom_A_DMSO(TSA solvent)_D11_22.CEL	D11_22	DMSO	DMSO ctr	DoD0-6	Solvent
UKN1_28_biom_A_DMSO(TSA solvent)_D11_23.CEL	D11_23	DMSO	DMSO ctr	DoD0-6	Solvent
UKN1_29_biom_A_DMSO(TSA solvent)_D11_24.CEL	D11_24	DMSO	DMSO ctr	DoD0-6	Solvent
UKN1_30_biom_A_DMSO(TSA solvent)_D11_25.CEL	D11_25	DMSO	DMSO ctr	DoD0-6	Solvent
UKN1_41_biom_A_untr_D11_21.CEL	D11_21	untreated	untreated	untreated	Untreated
UKN1_42_biom_A_untr_D11_22.CEL	D11_22	untreated	untreated	untreated	Untreated
UKN1_43_biom_A_untr_D11_23.CEL	D11_23	untreated	untreated	untreated	Untreated
UKN1_44_biom_A_untr_D11_24.CEL	D11_24	untreated	untreated	untreated	Untreated
UKN1_45_biom_A_untr_D11_25.CEL	D11_25	untreated	untreated	untreated	Untreated
UKN1_46_biom_A_untr_D12_07.CEL	D12_07	untreated	untreated	untreated	Untreated
UKN1_47_biom_A_untr_D12_08.CEL	D12_08	untreated	untreated	untreated	Untreated
UKN1_48_biom_A_untr_D12_09.CEL	D12_09	untreated	untreated	untreated	Untreated
UKN1_49_biom_A_untr_D12_10.CEL	D12_10	untreated	untreated	untreated	Untreated

Tabelle 23: UKN1 Klassifikationsstudie

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Typ
UKN1_50_biom_A_VPA_D12_07.CEL	D12_07	VPA	VPA 0.6 mM	DoD0-6	HDACi
UKN1_51_biom_A_VPA_D12_08.CEL	D12_08	VPA	VPA 0.6 mM	DoD0-6	HDACi
UKN1_52_biom_A_VPA_D12_09.CEL	D12_09	VPA	VPA 0.6 mM	DoD0-6	HDACi
UKN1_53_biom_A_VPA_D12_10.CEL	D12_10	VPA	VPA 0.6 mM	DoD0-6	HDACi
UKN1_58_biom_A_untr_D12_18.CEL	D12_18	untreated	untreated	untreated	Untreated
UKN1_59_biom_A_untr_D12_19.CEL	D12_19	untreated	untreated	untreated	Untreated
UKN1_60_biom_A_untr_D12_20.CEL	D12_20	untreated	untreated	untreated	Untreated
UKN1_61_biom_A_untr_D12_21.CEL	D12_21	untreated	untreated	untreated	Untreated
UKN1_62_biom_A_DMSO(SAHA solvent)_D12_18.CEL	D12_18	DMSO	DMSO 36.4 μM	DoD0-6	Solvent
UKN1_63_biom_A_DMSO(SAHA solvent)_D12_19.CEL	D12_19	DMSO	DMSO 36.4 μM	DoD0-6	Solvent
UKN1_64_biom_A_DMSO(SAHA solvent)_D12_20.CEL	D12_20	DMSO	DMSO 36.4 μM	DoD0-6	Solvent
UKN1_65_biom_A_DMSO(SAHA solvent)_D12_21.CEL	D12_21	DMSO	DMSO 36.4 μM	DoD0-6	Solvent
UKN1_66_biom_A_SAHA_D12_18.CEL	D12_18	SAHA	SAHA 0.13 μM	DoD0-6	HDACi
UKN1_67_biom_A_SAHA_D12_19.CEL	D12_19	SAHA	SAHA 0.13 μM	DoD0-6	HDACi
UKN1_68_biom_A_SAHA_D12_20.CEL	D12_20	SAHA	SAHA 0.13 μM	DoD0-6	HDACi
UKN1_69_biom_A_SAHA_D12_21.CEL	D12_21	SAHA	SAHA 0.13 μM	DoD0-6	HDACi
UKN1_70_biom_A_Thim_D12_18.CEL	D12_18	Thimerosal	Thimerosal 0.6 μM	DoD0-6	Hg+
UKN1_71_biom_A_Thim_D12_19.CEL	D12_19	Thimerosal	Thimerosal 0.6 μM	DoD0-6	Hg+
UKN1_72_biom_A_Thim_D12_20.CEL	D12_20	Thimerosal	Thimerosal 0.6 μM	DoD0-6	Hg+
UKN1_73_biom_A_Thim_D12_21.CEL	D12_21	Thimerosal	Thimerosal 0.6 μM	DoD0-6	Hg+
UKN1_74_biom_A_HgCl2_D12_18.CEL	D12_18	HgCl2	HgCl2 0.1 μM	DoD0-6	Hg+
UKN1_75_biom_A_HgCl2_D12_19.CEL	D12_19	HgCl2	HgCl2 0.1 μM	DoD0-6	Hg+
UKN1_76_biom_A_HgCl2_D12_20.CEL	D12_20	HgCl2	HgCl2 0.1 μM	DoD0-6	Hg+
UKN1_77_biom_A_HgCl2_D12_21.CEL	D12_21	HgCl2	HgCl2 0.1 μM	DoD0-6	Hg+

Tabelle 24: UKN1 Klassifikationsstudie (Fortsetzung)

CEL.name	Code	Substanz	Behandlung	Behandlungsdauer	Typ
Tanja_Waldmann_1.1_.HG.U133_Plus_2..CEL	1	Belinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_1.2_.HG.U133_Plus_2..CEL	2	Belinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_1.3_.HG.U133_Plus_2..CEL	3	Belinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_1.4_.HG.U133_Plus_2..CEL	4	Belinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_2.1_.HG.U133_Plus_2..CEL	1	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_2.2_.HG.U133_Plus_2..CEL	2	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_2.3_.HG.U133_Plus_2..CEL	3	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_2.4_.HG.U133_Plus_2..CEL	4	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_3.1_.HG.U133_Plus_2..CEL	1	Panobinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_3.2_.HG.U133_Plus_2..CEL	2	Panobinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_3.3_.HG.U133_Plus_2..CEL	3	Panobinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_3.4_.HG.U133_Plus_2..CEL	4	Panobinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_4.1_.HG.U133_Plus_2..CEL	1	HgBr2	unknown	DoD0-6	Hg+
Tanja_Waldmann_4.2_.HG.U133_Plus_2..CEL	2	HgBr2	unknown	DoD0-6	Hg+
Tanja_Waldmann_4.3_.HG.U133_Plus_2..CEL	3	HgBr2	unknown	DoD0-6	Hg+
Tanja_Waldmann_4.4_.HG.U133_Plus_2..CEL	4	HgBr2	unknown	DoD0-6	Hg+
Tanja_Waldmann_5.1_.HG.U133_Plus_2..CEL	1	Entinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_5.2_.HG.U133_Plus_2..CEL	2	Entinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_5.3_.HG.U133_Plus_2..CEL	3	Entinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_5.4_.HG.U133_Plus_2..CEL	4	Entinostat	unknown	DoD0-6	HDACi
Tanja_Waldmann_6.1_.HG.U133_Plus_2..CEL	1	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_6.2_.HG.U133_Plus_2..CEL	2	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_6.3_.HG.U133_Plus_2..CEL	3	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_6.4_.HG.U133_Plus_2..CEL	4	untreated	unknown	DoD0-6	Untreated
Tanja_Waldmann_7.1_.HG.U133_Plus_2..CEL	1	PMA	unknown	DoD0-6	Hg+
Tanja_Waldmann_7.2_.HG.U133_Plus_2..CEL	2	PMA	unknown	DoD0-6	Hg+
Tanja_Waldmann_7.3_.HG.U133_Plus_2..CEL	3	PMA	unknown	DoD0-6	Hg+
Tanja_Waldmann_7.4_.HG.U133_Plus_2..CEL	4	PMA	unknown	DoD0-6	Hg+
Tanja_Waldmann_8.1_.HG.U133_Plus_2..CEL	1	PCMB	unknown	DoD0-6	Hg+
Tanja_Waldmann_8.2_.HG.U133_Plus_2..CEL	2	PCMB	unknown	DoD0-6	Hg+
Tanja_Waldmann_8.3_.HG.U133_Plus_2..CEL	3	PCMB	unknown	DoD0-6	Hg+
Tanja_Waldmann_8.4_.HG.U133_Plus_2..CEL	4	PCMB	unknown	DoD0-6	Hg+

Tabelle 25: UKN1 Klassifikationsstudie (Fortsetzung)

	Names	Compound	Klasse	Batch
1	UKK_Solvent_Control_1	Control	Control	First
2	UKK_Solvent_Control_2	Control	Control	First
3	UKK_Solvent_Control_3	Control	Control	First
4	UKK_Solvent_Control_4	Control	Control	First
5	UKK_Solvent_Control_5	Control	Control	First
6	UKK_MeHg_IC10_1	MeHg high	M	First
7	UKK_MeHg_IC10_2	MeHg high	M	First
8	UKK_MeHg_IC10_3	MeHg high	M	First
9	UKK_MeHg_IC10_4	MeHg high	M	First
10	UKK_MeHg_IC10_5	MeHg high	M	First
11	UKK_MeHg_IC2.5_1	MeHg low	M	First
12	UKK_MeHg_IC2.5_2	MeHg low	M	First
13	UKK_MeHg_IC2.5_3	MeHg low	M	First
14	UKK_MeHg_IC2.5_4	MeHg low	M	First
15	UKK_MeHg_IC2.5_5	MeHg low	M	First
16	UKK_VPA_IC10_1	VPA high	HDAC	First
17	UKK_VPA_IC10_2	VPA high	HDAC	First
18	UKK_VPA_IC10_3	VPA high	HDAC	First
19	UKK_VPA_IC10_4	VPA high	HDAC	First
20	UKK_VPA_IC10_5	VPA high	HDAC	First
21	UKK_VPA_IC2.5_1	VPA low	HDAC	First
22	UKK_VPA_IC2.5_2	VPA low	HDAC	First
23	UKK_VPA_IC2.5_3	VPA low	HDAC	First
24	UKK_VPA_IC2.5_4	VPA low	HDAC	First
25	UKK_VPA_IC2.5_5	VPA low	HDAC	First
26	UKK_II_DMSO Control_A_9_(HG-U133_Plus_2).CEL	DMSO	Control	Second
27	UKK_II_DMSO Control_B_10_(HG-U133_Plus_2).CEL	DMSO	Control	Second
28	UKK_II_DMSO Control_C_11_(HG-U133_Plus_2).CEL	DMSO	Control	Second
29	UKK_II_DMSO Control_D_12_(HG-U133_Plus_2).CEL	DMSO	Control	Second

Tabelle 26: UKK Klassifikationsstudie

	Names	Compound	Klasse	Batch
30	UKK_II_ES_A_1_(HG-U133_Plus_2).CEL	ES	Control	Second
31	UKK_II_ES_B_2_(HG-U133_Plus_2).CEL	ES	Control	Second
32	UKK_II_ES_C_3_(HG-U133_Plus_2).CEL	ES	Control	Second
33	UKK_II_ES_D_4_(HG-U133_Plus_2).CEL	ES	Control	Second
34	UKK_II_Hgcl 0.7uM_A_25_(HG-U133_Plus_2).CEL	HgCl2	M	Second
35	UKK_II_Hgcl 0.7uM_B_26_(HG-U133_Plus_2).CEL	HgCl2	M	Second
36	UKK_II_Hgcl 0.7uM_C_27_(HG-U133_Plus_2).CEL	HgCl2	M	Second
37	UKK_II_Hgcl 0.7uM_D_28_(HG-U133_Plus_2).CEL	HgCl2	M	Second
38	UKK_II_SAHA 0.5uM_A_17_(HG-U133_Plus_2).CEL	SAHA	HDAC	Second
39	UKK_II_SAHA 0.5uM_B_18_(HG-U133_Plus_2).CEL	SAHA	HDAC	Second
40	UKK_II_SAHA 0.5uM_C_19_(HG-U133_Plus_2).CEL	SAHA	HDAC	Second
41	UKK_II_SAHA 0.5uM_D_20_(HG-U133_Plus_2).CEL	SAHA	HDAC	Second
42	UKK_II_Thiomersal 0.4uM_A_21_(HG-U133_Plus_2).CEL	Thiomersal	M	Second
43	UKK_II_Thiomersal 0.4uM_B_22_(HG-U133_Plus_2).CEL	Thiomersal	M	Second
44	UKK_II_Thiomersal 0.4uM_C_23_(HG-U133_Plus_2).CEL	Thiomersal	M	Second
45	UKK_II_Thiomersal 0.4uM_D_24_(HG-U133_Plus_2).CEL	Thiomersal	M	Second
46	UKK_II_TSA 0.02uM_A_13_(HG-U133_Plus_2).CEL	TSA	HDAC	Second
47	UKK_II_TSA 0.02uM_B_14_(HG-U133_Plus_2).CEL	TSA	HDAC	Second
48	UKK_II_TSA 0.02uM_C_15_(HG-U133_Plus_2).CEL	TSA	HDAC	Second
49	UKK_II_TSA 0.02uM_D_16_(HG-U133_Plus_2).CEL	TSA	HDAC	Second
50	UKK_II_Untreated Control_A_5_(HG-U133_Plus_2).CEL	Control	Control	Second
51	UKK_II_Untreated Control_B_6_(HG-U133_Plus_2).CEL	Control	Control	Second
52	UKK_II_Untreated Control_C_7_(HG-U133_Plus_2).CEL	Control	Control	Second
53	UKK_II_Untreated Control_D_8_(HG-U133_Plus_2).CEL	Control	Control	Second

Tabelle 27: UKK Klassifikationsstudie (Fortsetzung)

	Names	Compound	Klasse	Batch
54	UKK_BS_01_.HG.U133_Plus_2..CEL	Control	Control	Third
55	UKK_BS_02_.HG.U133_Plus_2..CEL	Panobinostat	HDAC	Third
56	UKK_BS_03_.HG.U133_Plus_2..CEL	HgBr2	M	Third
57	UKK_BS_04_.HG.U133_Plus_2..CEL	Control	Control	Third
58	UKK_BS_05_.HG.U133_Plus_2..CEL	PMA	M	Third
59	UKK_BS_06_.HG.U133_Plus_2..CEL	PCMB	M	Third
60	UKK_BS_07_.HG.U133_Plus_2..CEL	Control	Control	Third
61	UKK_BS_08_.HG.U133_Plus_2..CEL	Control	Control	Third
62	UKK_BS_09_.HG.U133_Plus_2..CEL	Panobinostat	HDAC	Third
63	UKK_BS_10_.HG.U133_Plus_2..CEL	HgBr2	M	Third
64	UKK_BS_11_.HG.U133_Plus_2..CEL	Control	Control	Third
65	UKK_BS_12_.HG.U133_Plus_2..CEL	PMA	M	Third
66	UKK_BS_13_.HG.U133_Plus_2..CEL	PCMB	M	Third
67	UKK_BS_14_.HG.U133_Plus_2..CEL	Control	Control	Third
68	UKK_BS_15_.HG.U133_Plus_2..CEL	Control	Control	Third
69	UKK_BS_16_.HG.U133_Plus_2..CEL	Panobinostat	HDAC	Third
70	UKK_BS_17_.HG.U133_Plus_2..CEL	HgBr2	M	Third
71	UKK_BS_18_.HG.U133_Plus_2..CEL	Control	Control	Third
72	UKK_BS_19_.HG.U133_Plus_2..CEL	PMA	M	Third
73	UKK_BS_20_.HG.U133_Plus_2..CEL	PCMB	M	Third
74	UKK_BS_21_.HG.U133_Plus_2..CEL	Control	Control	Third
75	UKK_BS_22_.HG.U133_Plus_2..CEL	Control	Control	Third
76	UKK_BS_23.HG.U133_Plus_2..CEL	Panobinostat	HDAC	Third
77	UKK_BS_24_.HG.U133_Plus_2..CEL	HgBr2	M	Third
78	UKK_BS_25_.HG.U133_Plus_2..CEL	Control	Control	Third
79	UKK_BS_26_.HG.U133_Plus_2..CEL	PMA	M	Third
80	UKK_BS_27_.HG.U133_Plus_2..CEL	PCMB	M	Third
81	UKK_BS_28_.HG.U133_Plus_2..CEL	Control	Control	Third

Tabelle 28: UKK Klassifikationsstudie (Fortsetzung)

	Names	Compound	Klasse	Batch
82	UKK_BS_I_1_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
83	UKK_BS_I_2_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
84	UKK_BS_I_5_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
85	UKK_BS_I_6_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
86	UKK_BS_I_7_(HG-U133_Plus_2).CEL	Control	Control	Fourth
87	UKK_BS_II_1_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
88	UKK_BS_II_2_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
89	UKK_BS_II_6_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
90	UKK_BS_II_7_(HG-U133_Plus_2).CEL	Control	Control	Fourth
91	UKK_BS_III_1_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
92	UKK_BS_III_2_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
93	UKK_BS_III_5_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
94	UKK_BS_III_6_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
95	UKK_BS_III_7_(HG-U133_Plus_2).CEL	Control	Control	Fourth
96	UKK_BS_IV_1_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
97	UKK_BS_IV_2_(HG-U133_Plus_2).CEL	Belinostat	HDAC	Fourth
98	UKK_BS_IV_5_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
99	UKK_BS_IV_6_(HG-U133_Plus_2).CEL	Entinostat	HDAC	Fourth
100	UKK_BS_IV_7_(HG-U133_Plus_2).CEL	Control	Control	Fourth

Tabelle 29: UKK Klassifikationsstudie (Fortsetzung)

6.3 Abbildungen

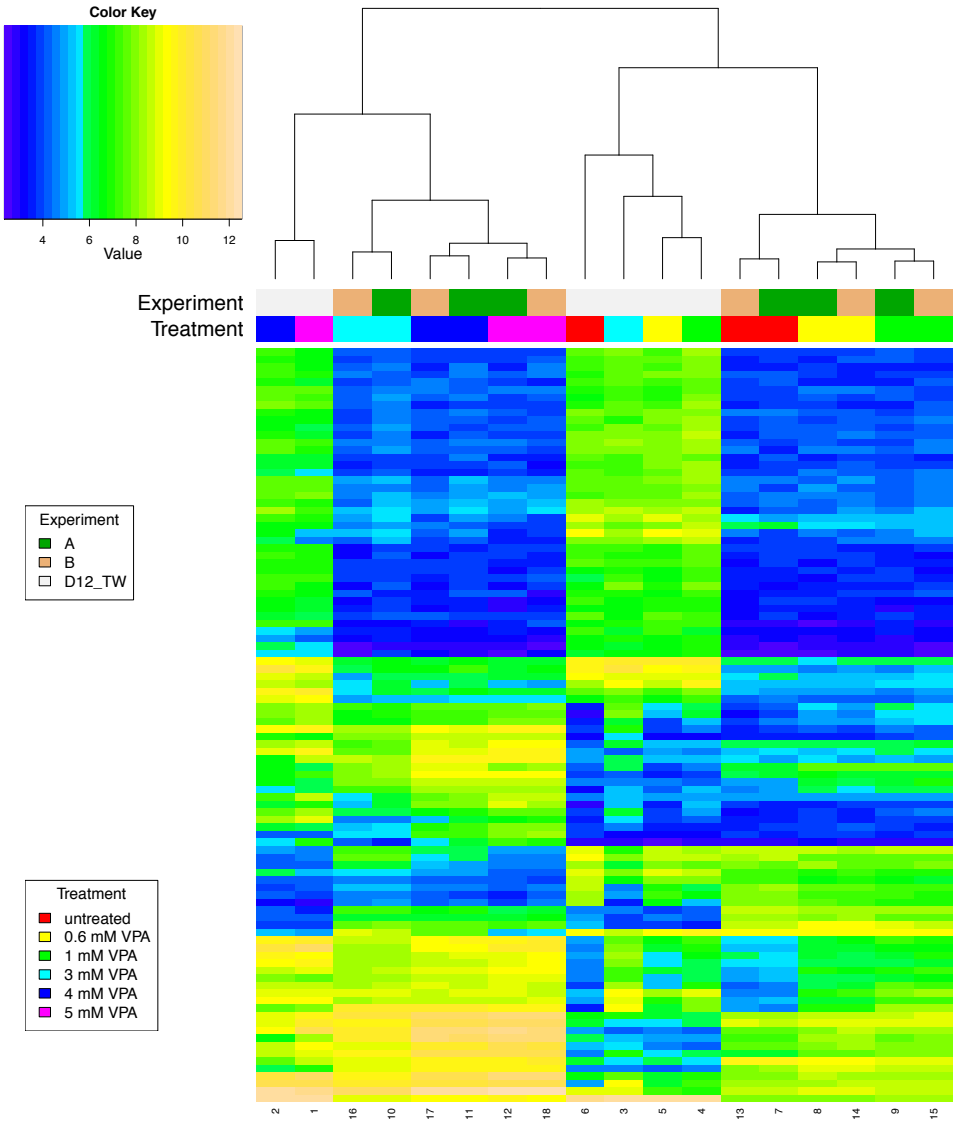

Abbildung 43: Heatmap der VPA Acute Konzentrationsstudie

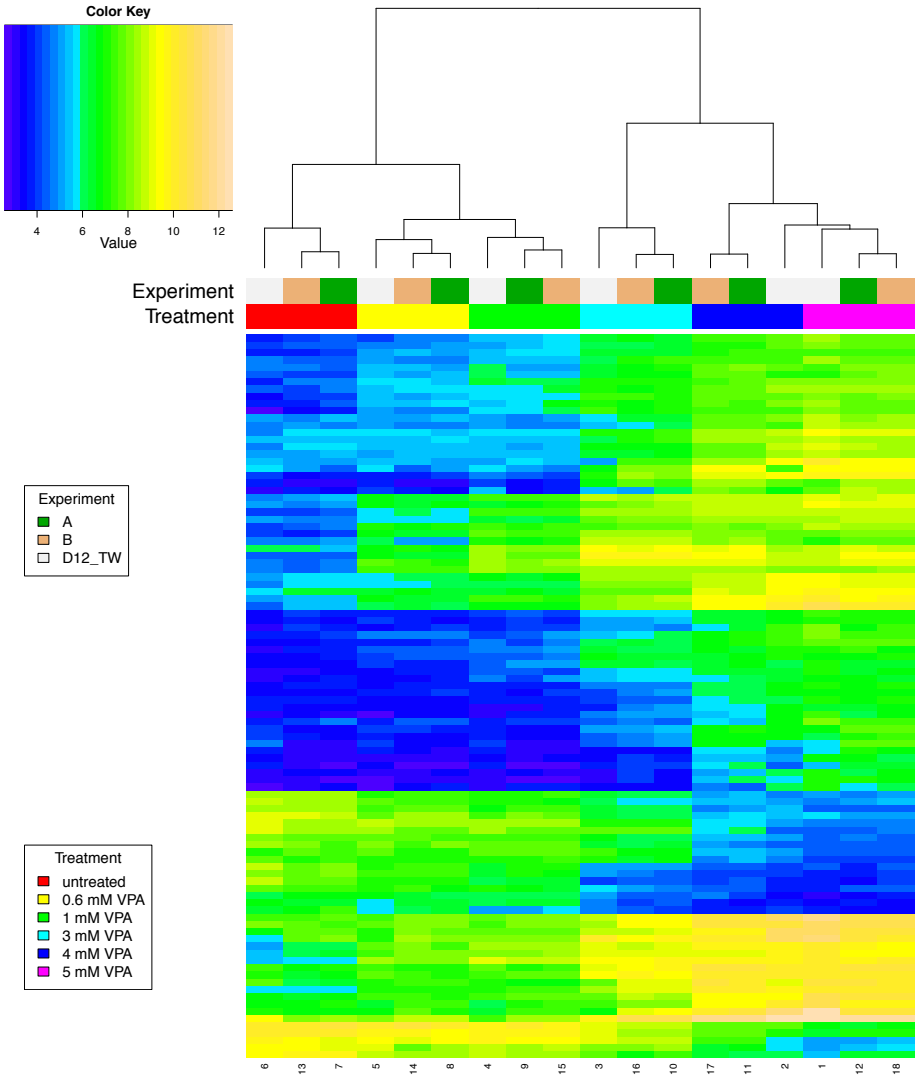

Abbildung 44: Heatmap der VPA Acute Konzentrationsstudie nach Transformation mit **ComBat**. Die farbigen Leisten unterhalb des Dendrogramms symbolisieren die Aufteilung der Proben nach Zugehörigkeiten zu verschiedenen Behandlungen bzw. Abfertigungsschüben.

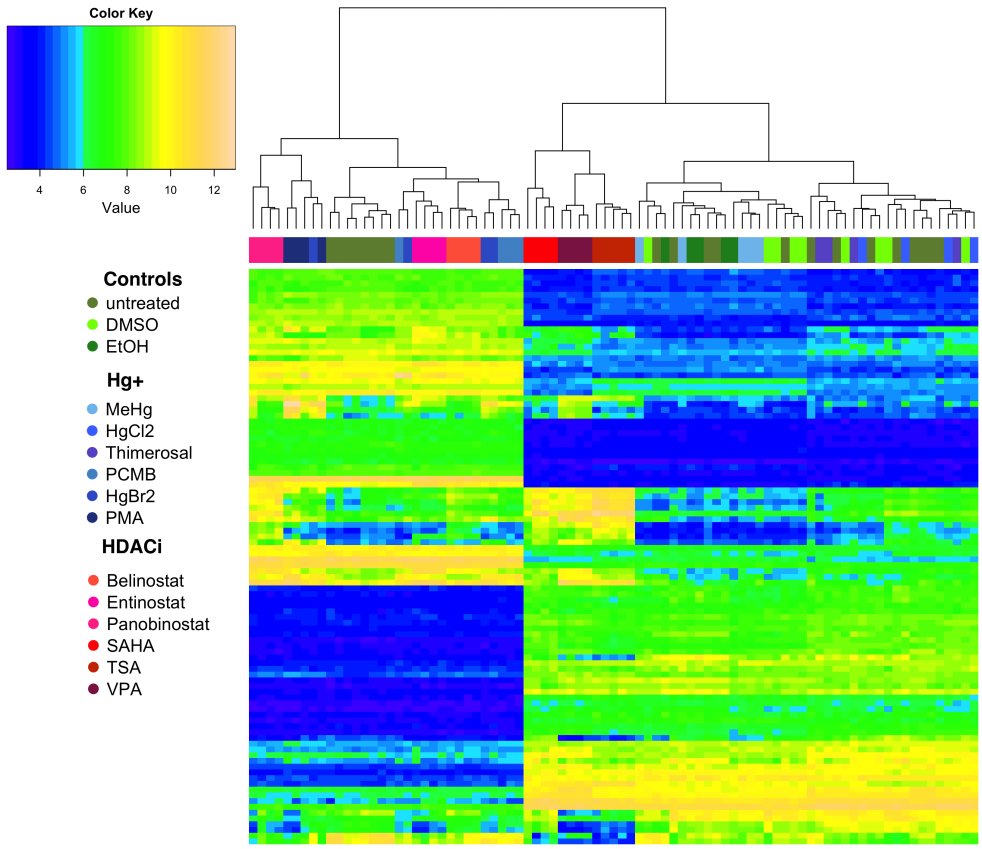

Abbildung 45: Heatmap der UKN1 Klassifikationsstudie. Die farbigen Leisten unterhalb des Dendrogramms symbolisieren die Aufteilung der Proben nach Zugehörigkeiten zu verschiedenen Behandlungen.

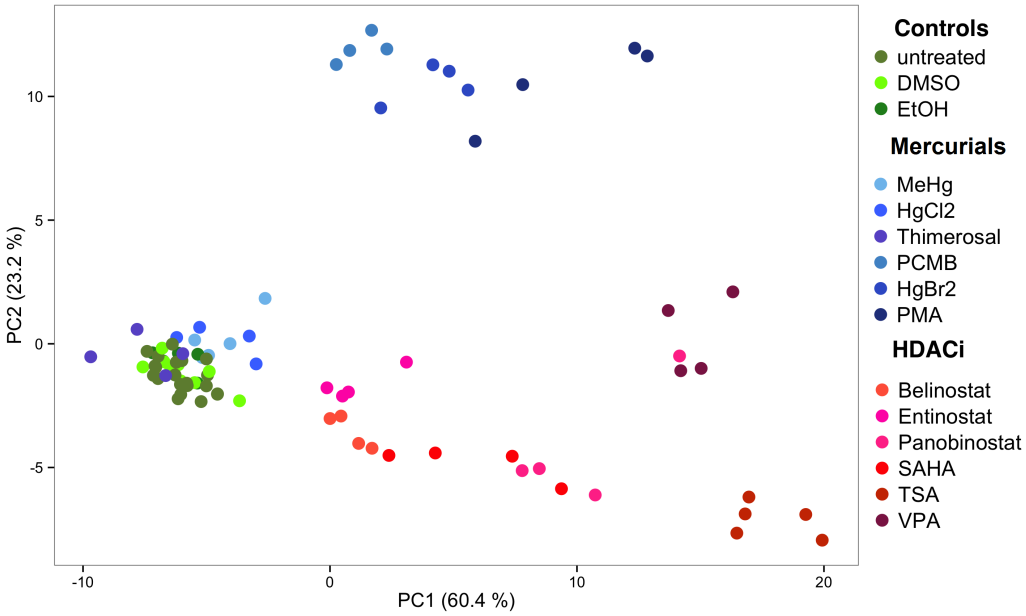

Abbildung 46: Hauptkomponentenplot der ComBat-transformierten Daten der UKN1 Klassifikationsstudie

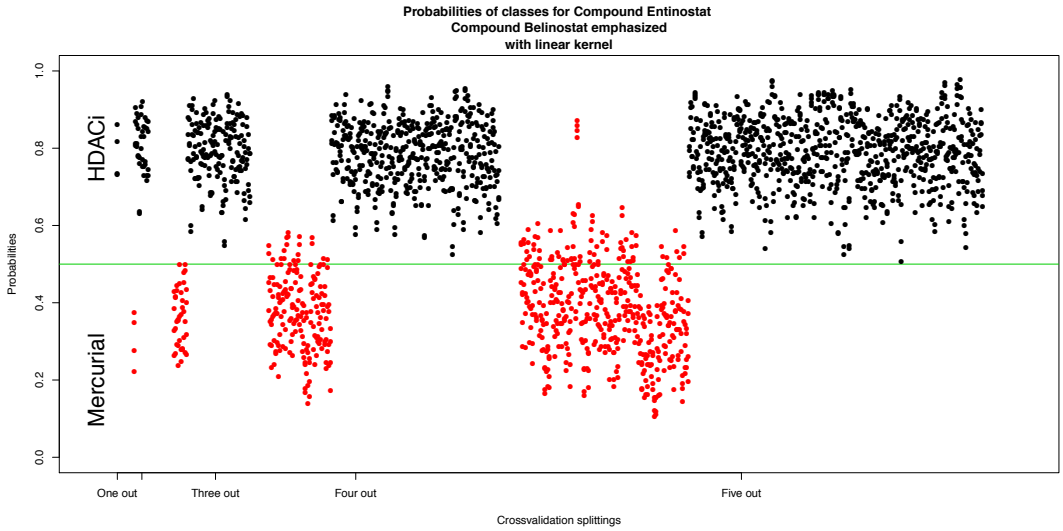

Abbildung 47: Wahrscheinlichkeiten-Graphik für die Vorhersage von Entinostat. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz Entinostat stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von Entinostat der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \geq 0.5$ und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben Entinostat auch Belinostat enthielten, sind durch rote Farbe hervorgehoben.

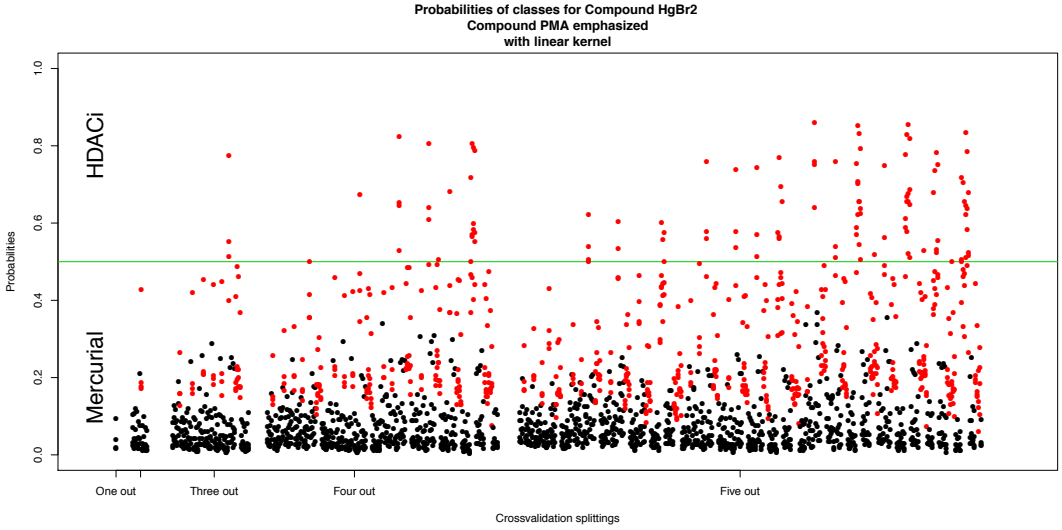

Abbildung 48: Wahrscheinlichkeiten-Graphik für die Vorhersage von HgBr2. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz HgBr2 stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von HgBr2 der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \leq 0.5$ und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben HgBr2 auch PMA enthielten, sind durch rote Farbe hervorgehoben.

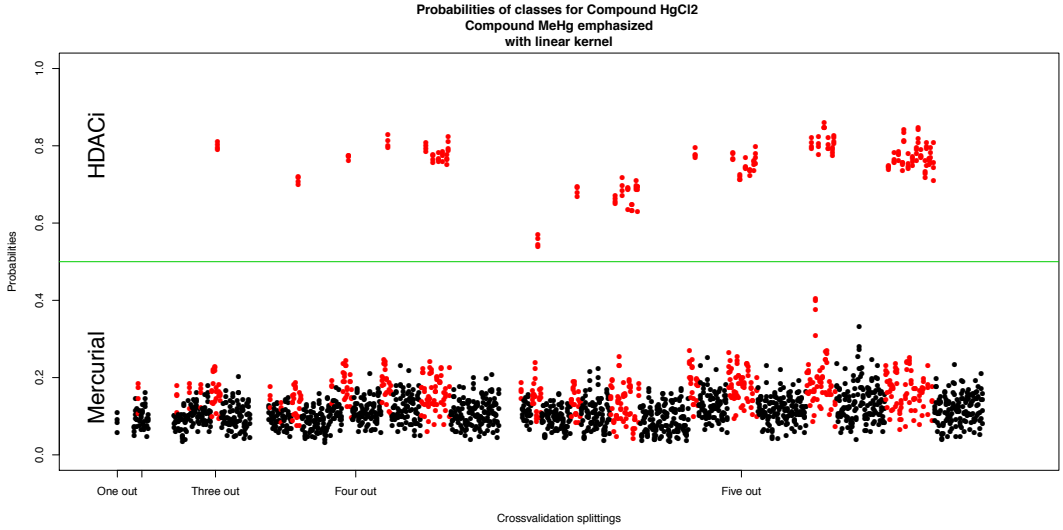

Abbildung 49: Wahrscheinlichkeiten-Graphik für die Vorhersage von HgCl2. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz HgCl2 stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von HgCl2 der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \leq 0.5$ und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben HgCl2 auch MeHg enthielten, sind durch rote Farbe hervorgehoben.

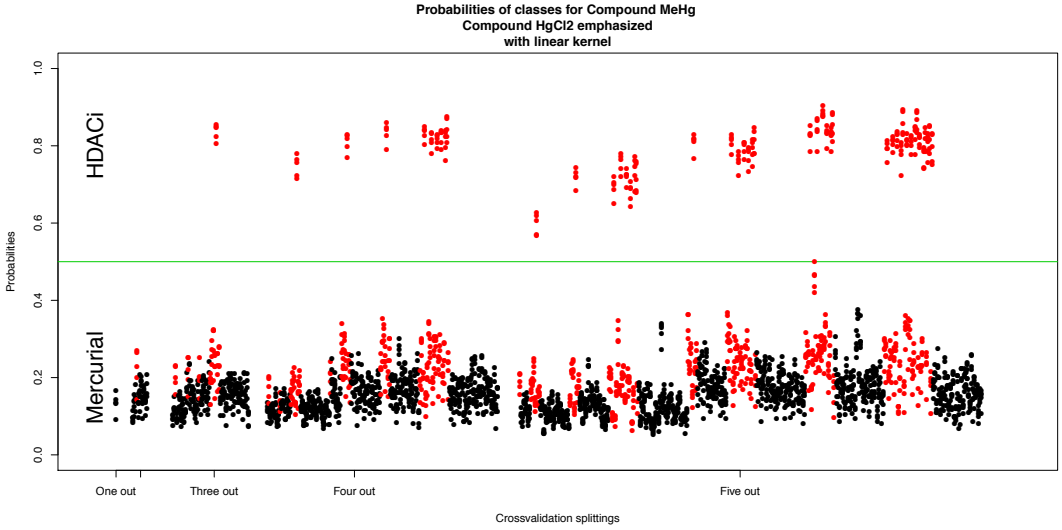

Abbildung 50: Wahrscheinlichkeiten-Graphik für die Vorhersage von MeHg. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz MeHg stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von MeHg der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \leq 0.5$ und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben MeHg auch HgCl2 enthielten, sind durch rote Farbe hervorgehoben.

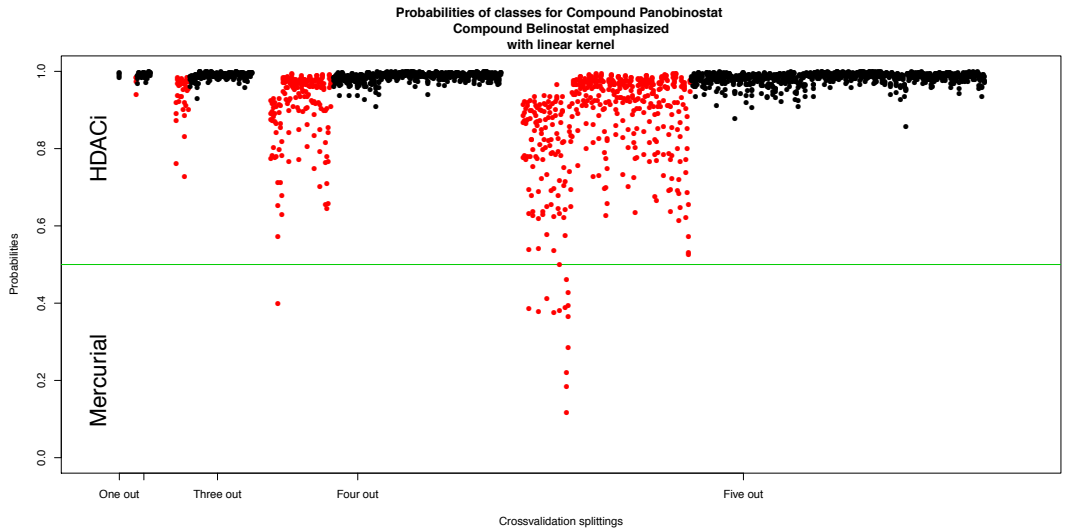

Abbildung 51: Wahrscheinlichkeiten-Graphik für die Vorhersage von Panobinostat. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz Panobinostat stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von Panobinostat der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \geq 0.5$ und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben Panobinostat auch Belinostat enthielten, sind durch rote Farbe hervorgehoben.

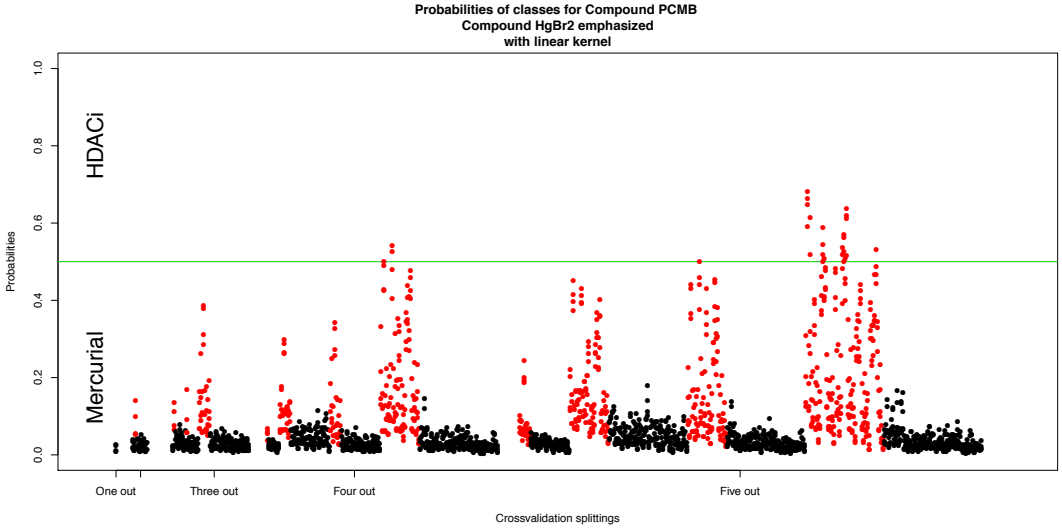

Abbildung 52: Wahrscheinlichkeiten-Graphik für die Vorhersage von PCMB. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz PCMB stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von PCMB der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \leq 0.5$ und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben PCMB auch HgBr2 enthielten, sind durch rote Farbe hervorgehoben.

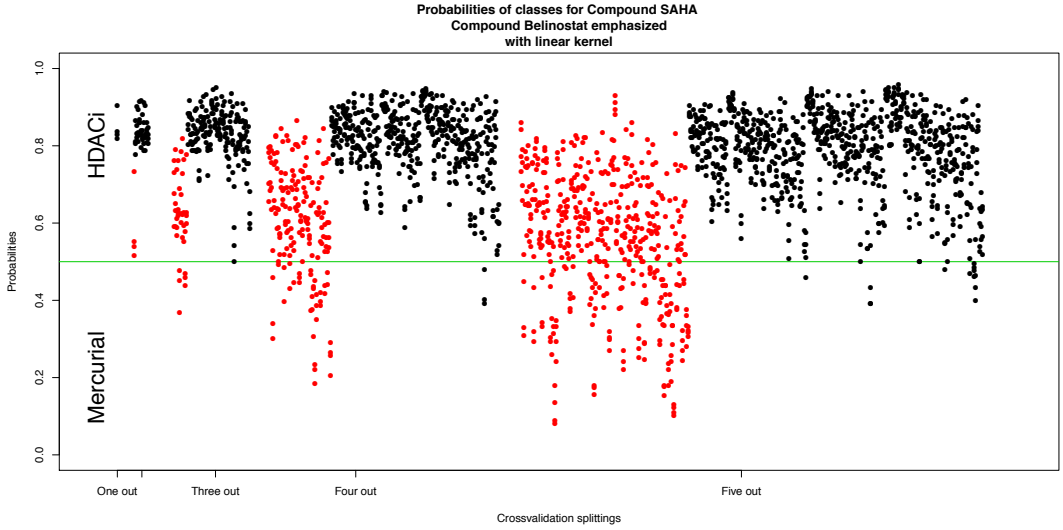

Abbildung 53: Wahrscheinlichkeiten-Graphik für die Vorhersage von SAHA. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz SAHA stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von SAHA der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \geq 0.5$ und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben SAHA auch Belinostat enthielten, sind durch rote Farbe hervorgehoben.

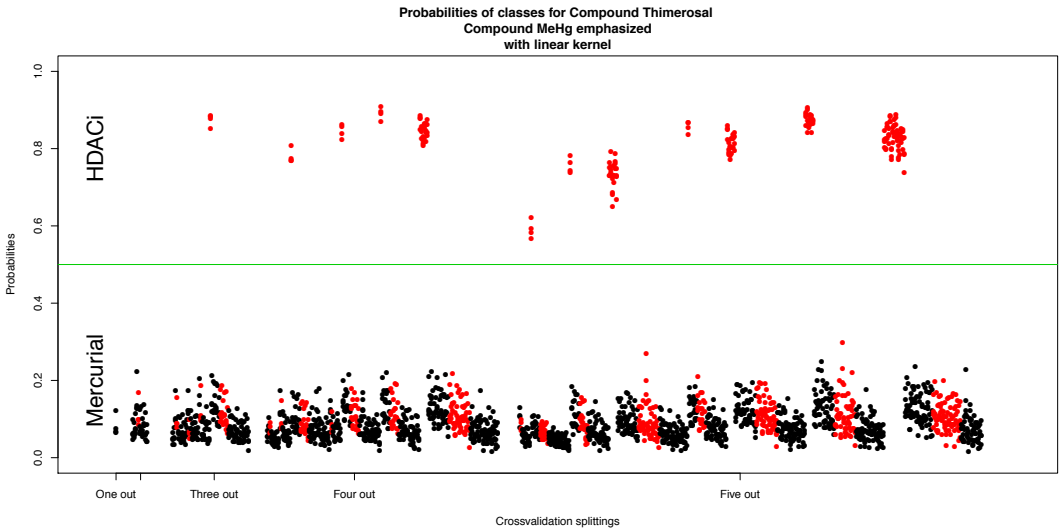

Abbildung 54: Wahrscheinlichkeiten-Graphik für die Vorhersage von Thimerosal. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz Thimerosal stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von Thimerosal der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \leq 0.5$ und liegt unterhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben Thimerosal auch MeHg enthielten, sind durch rote Farbe hervorgehoben.

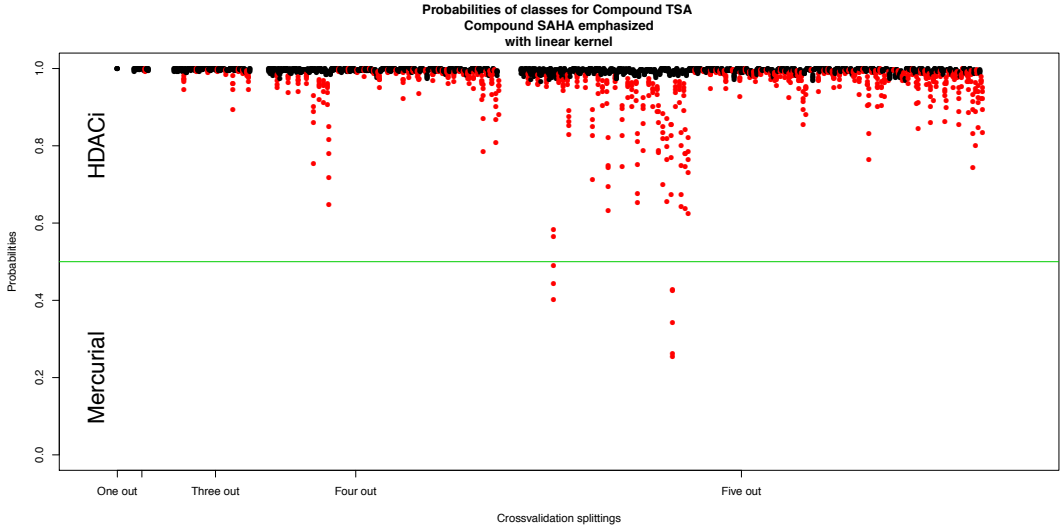

Abbildung 55: Wahrscheinlichkeiten-Graphik für die Vorhersage von TSA. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz TSA stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von TSA der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \geq 0.5$ und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben TSA auch SAHA enthielten, sind durch rote Farbe hervorgehoben.

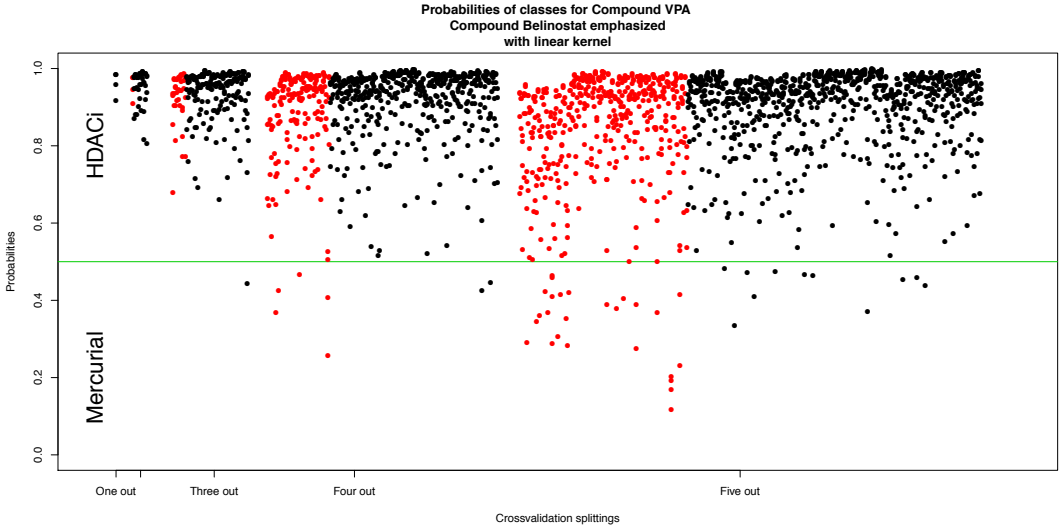

Abbildung 56: Wahrscheinlichkeiten-Graphik für die Vorhersage von VPA. Jedem Wert auf der x-Achse entspricht eine Aufteilung in die Training- und Testmenge. Dabei wurden nach einander eine bis fünf Substanzen in die Testmenge aufgenommen, wobei die Substanz VPA stets eine Testsubstanz bildete. Jeder Punkt entspricht der Wahrscheinlichkeit für ein Replikat von VPA der Klasse der HDAC-Inhibitoren zu gehören. Für richtige Vorhersagen weist der entsprechende Punkt den Wert $y \geq 0.5$ und liegt oberhalb der grünen Linie. Die Vorhersagewerte für die Testmengen, die neben VPA auch Belinostat enthielten, sind durch rote Farbe hervorgehoben.

6.3.1 Einfluss technischer Replikate auf UKN1 SVM

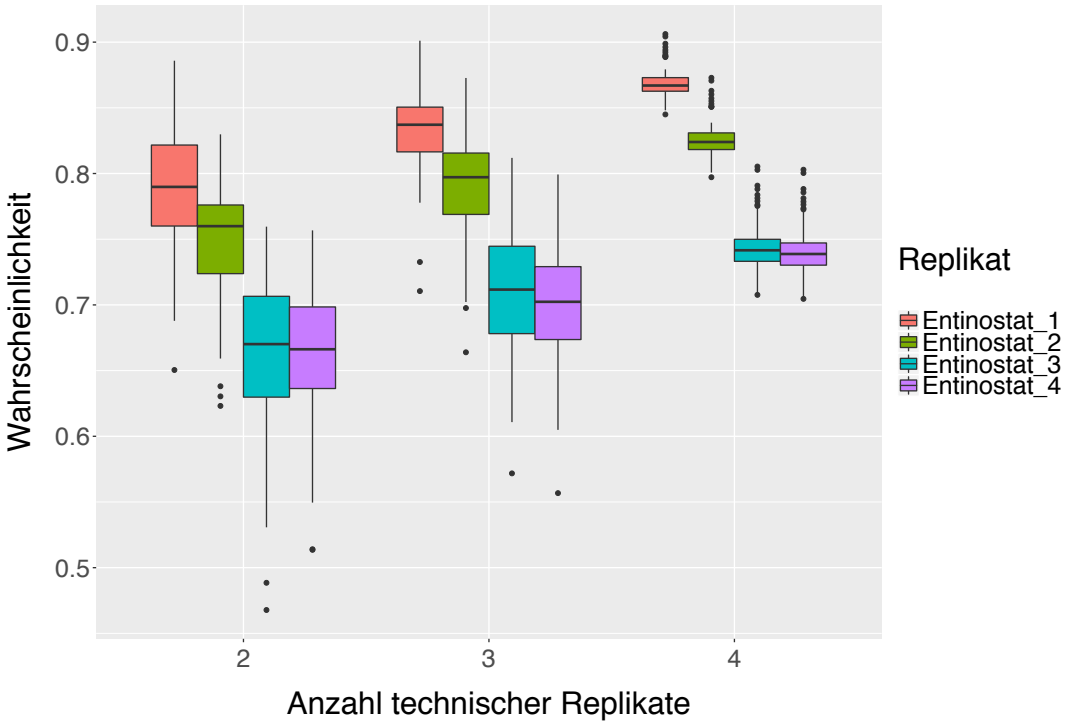

Abbildung 57: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

Abbildung 58: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

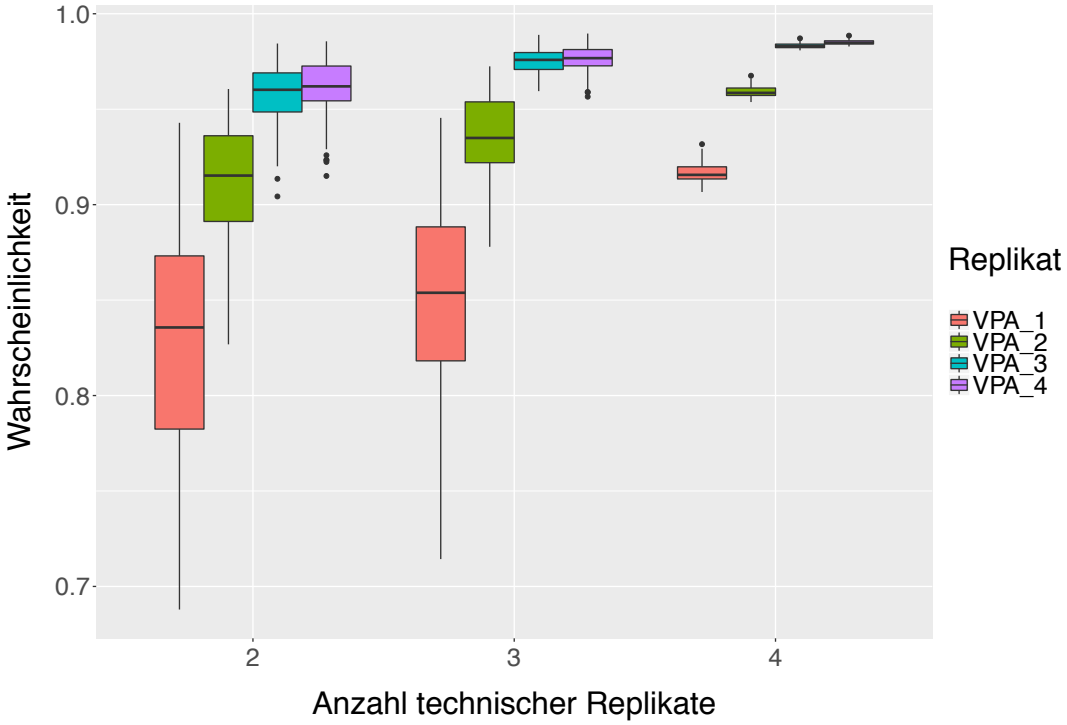

Abbildung 59: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

Abbildung 60: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

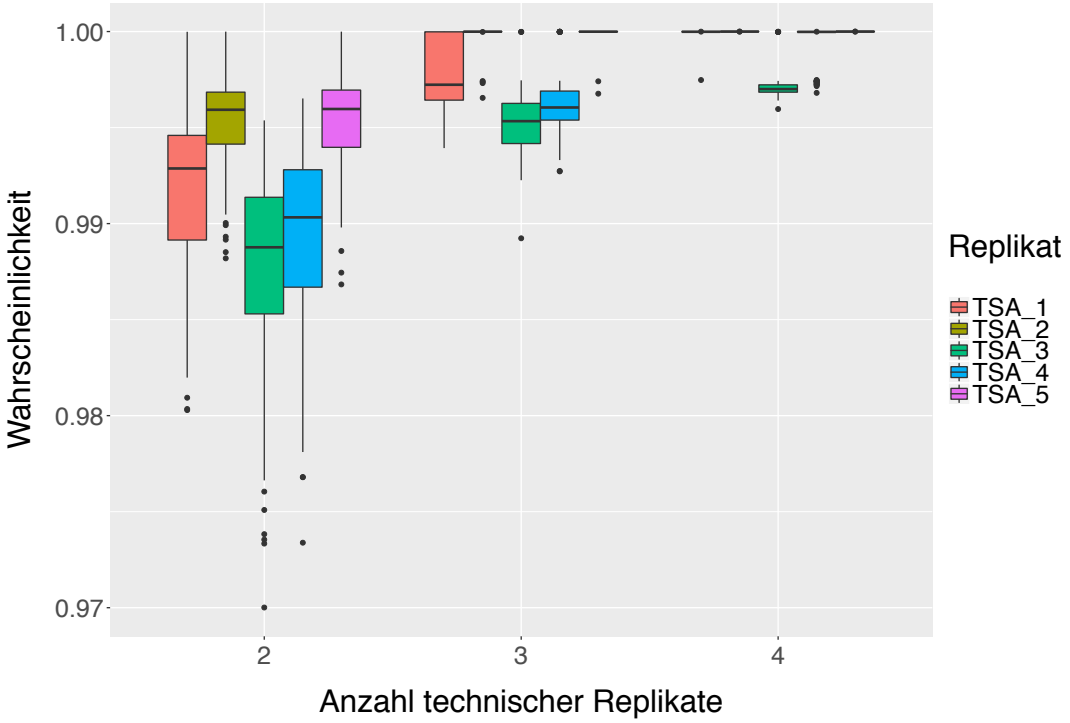

Abbildung 61: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

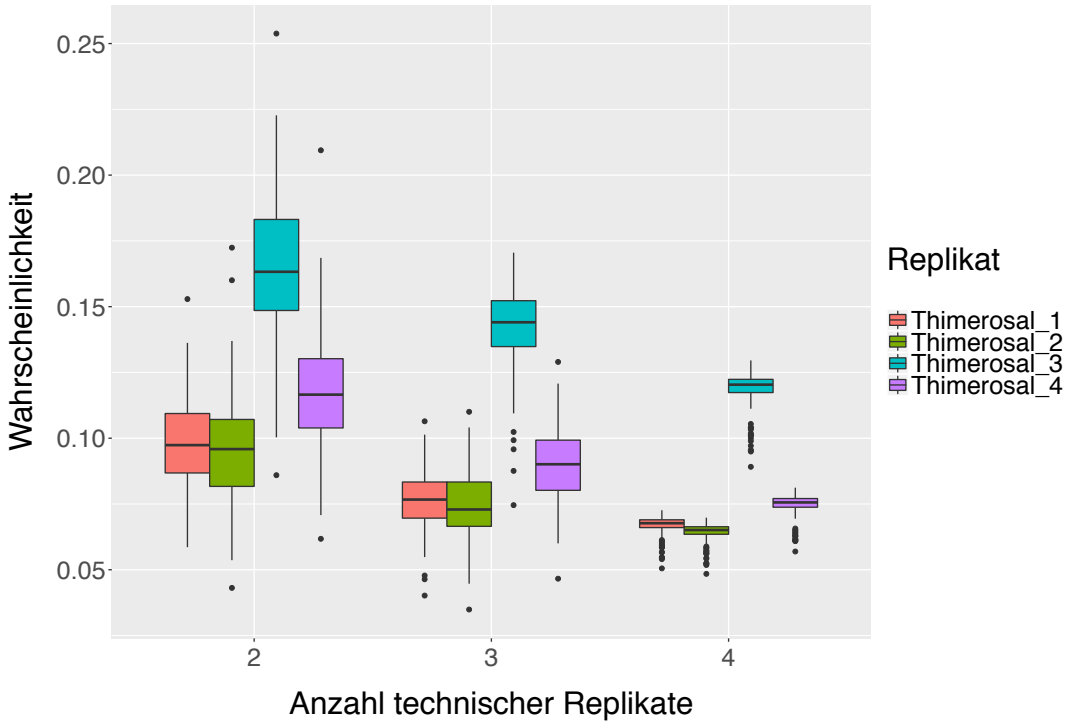

Abbildung 62: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thimerosal. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

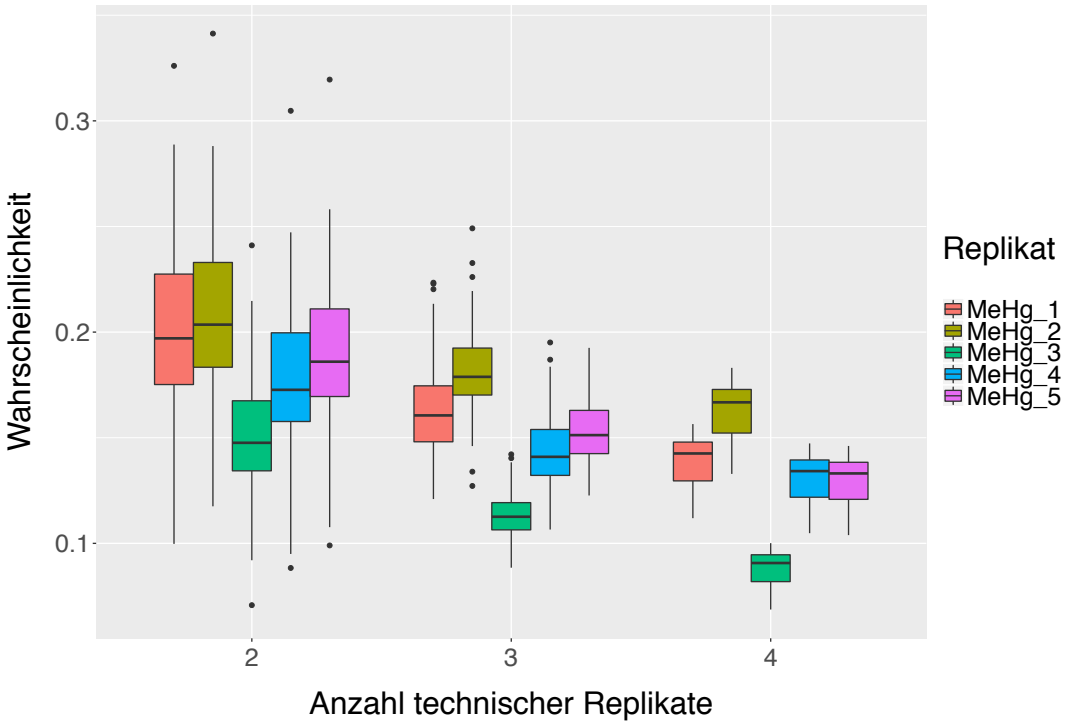

Abbildung 63: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

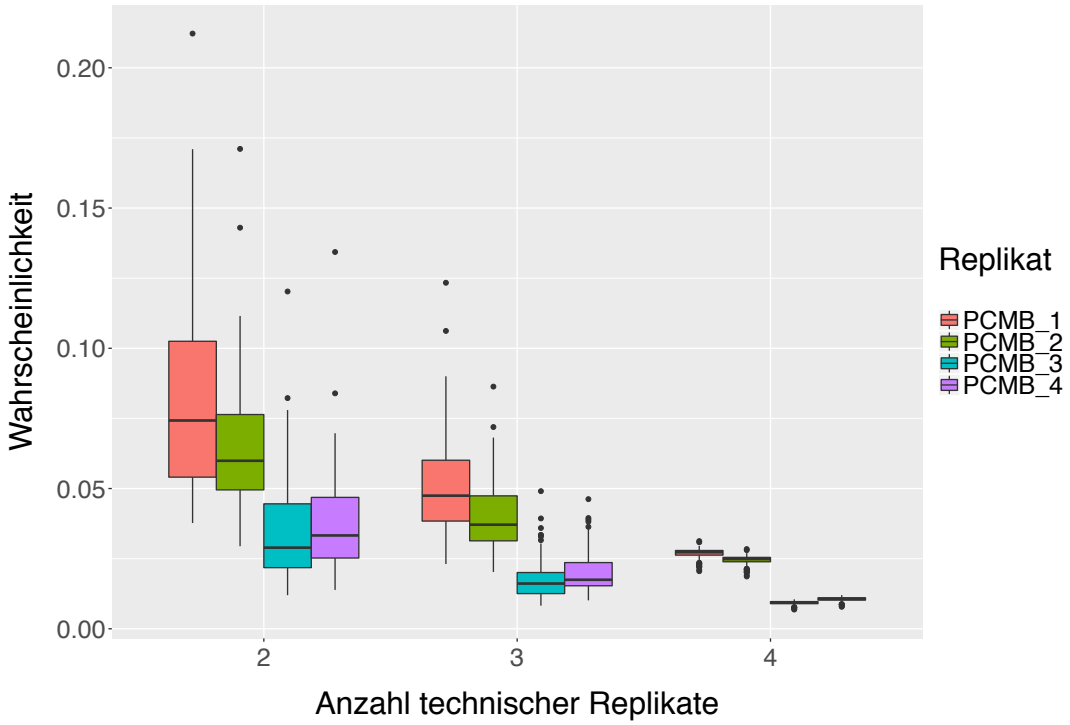

Abbildung 64: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

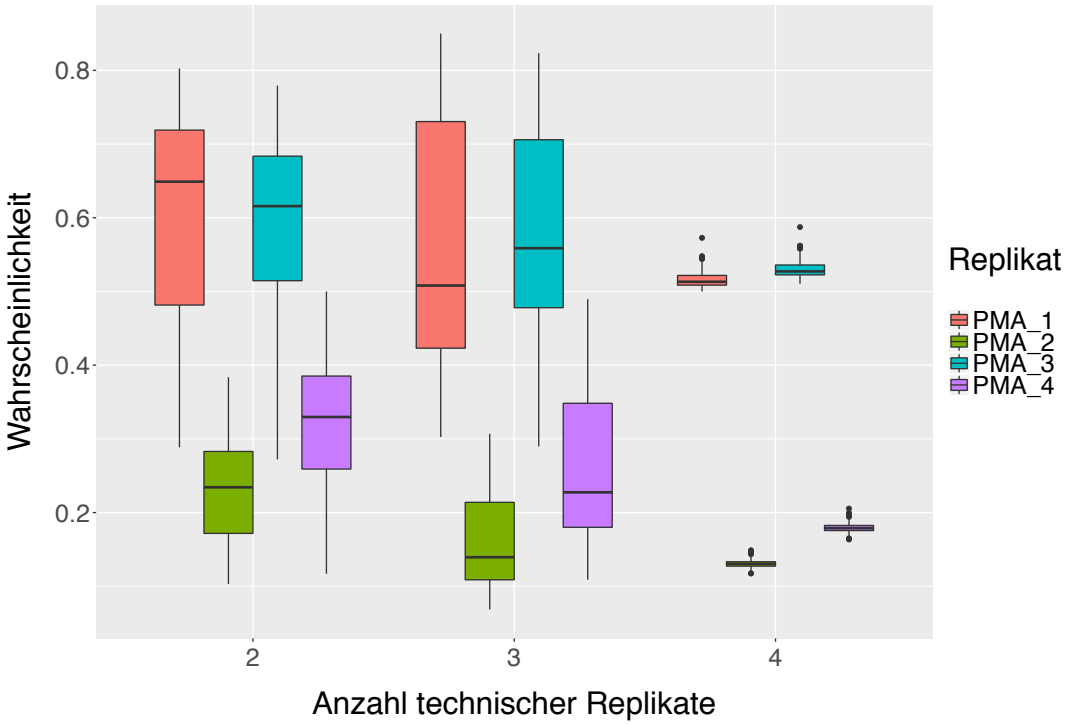

Abbildung 65: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

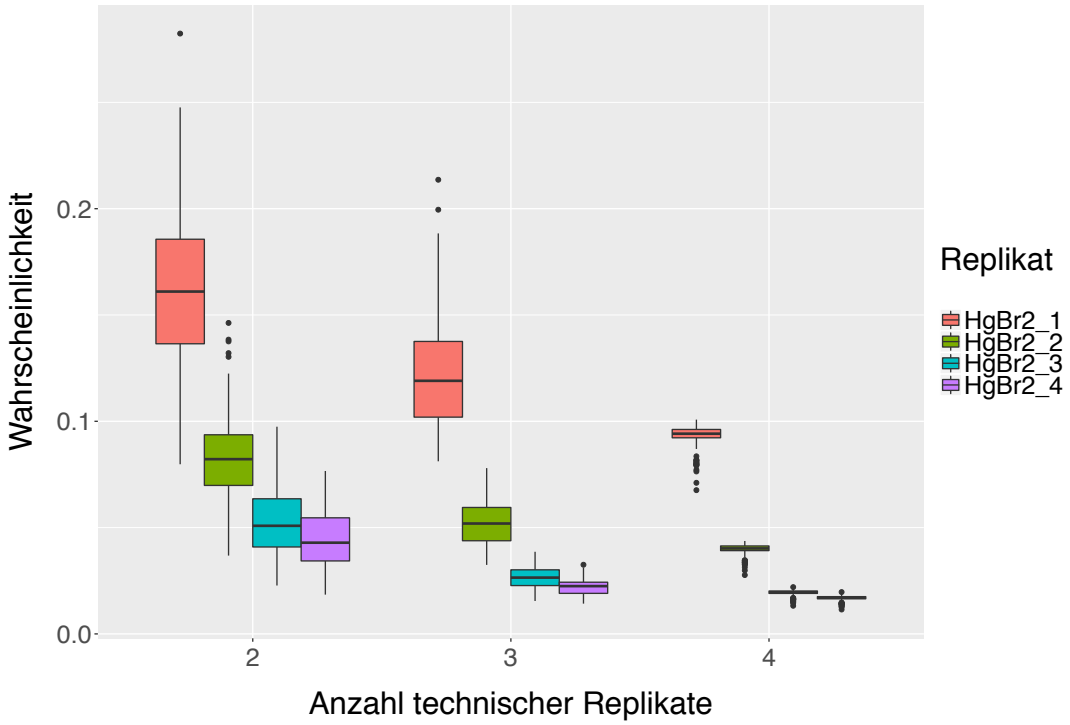

Abbildung 66: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgBr2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

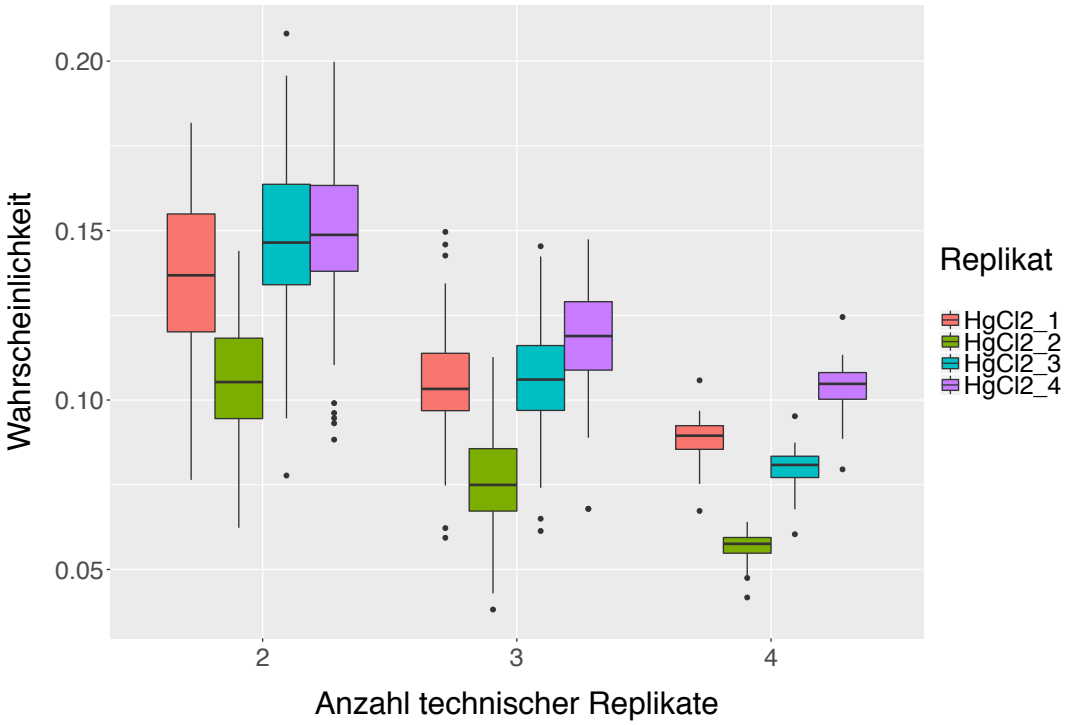

Abbildung 67: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgCl2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

6.3.2 Einfluss technischer Replikate auf UKN1 RF

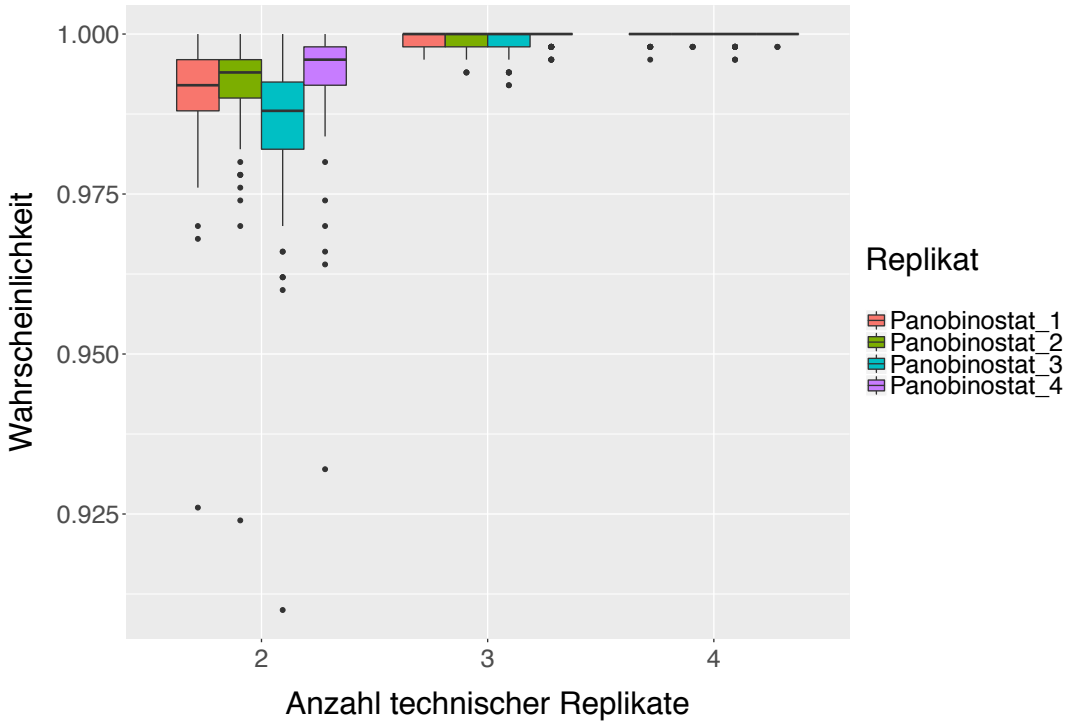

Abbildung 68: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

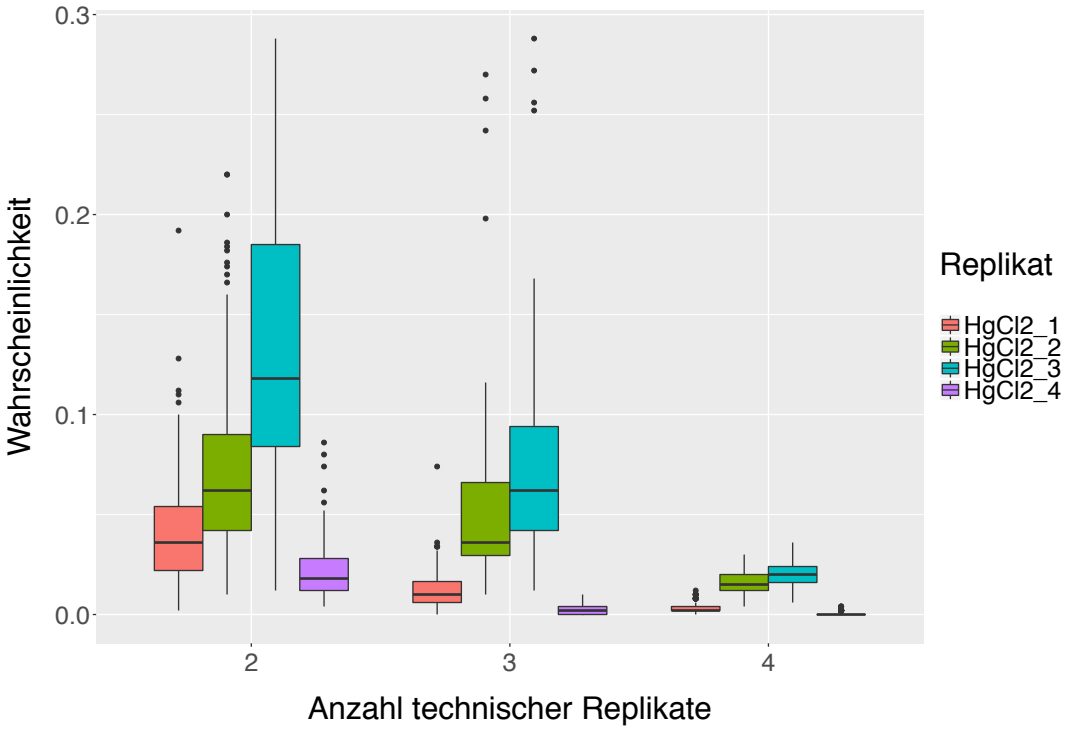

Abbildung 69: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgCl2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

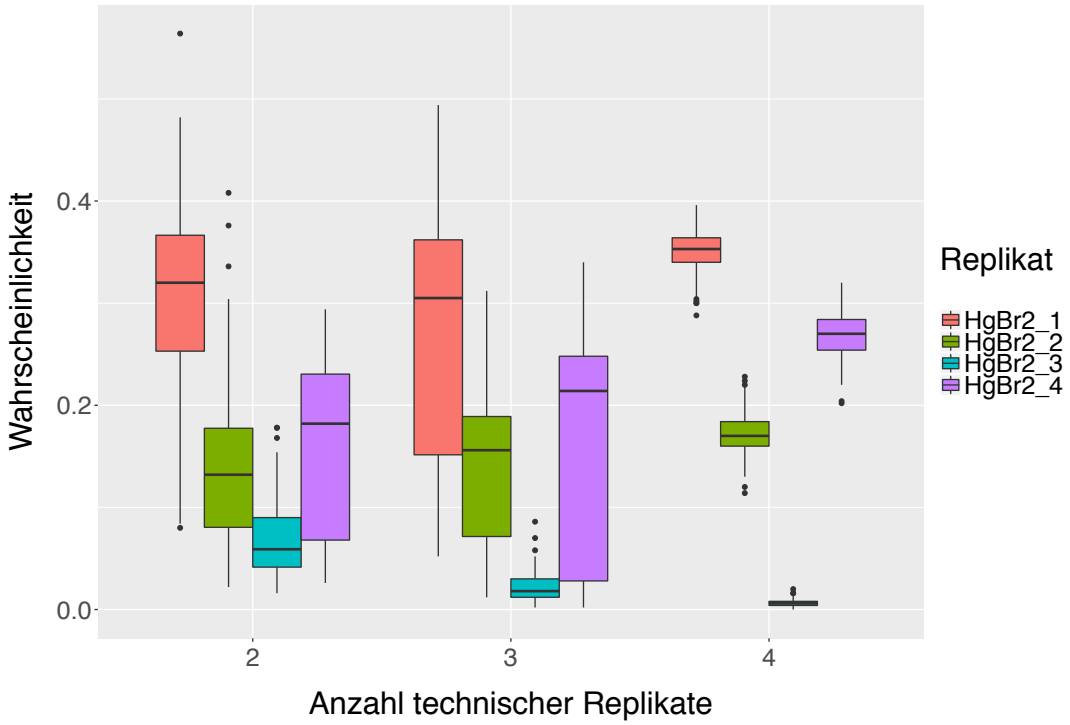

Abbildung 70: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgBr2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

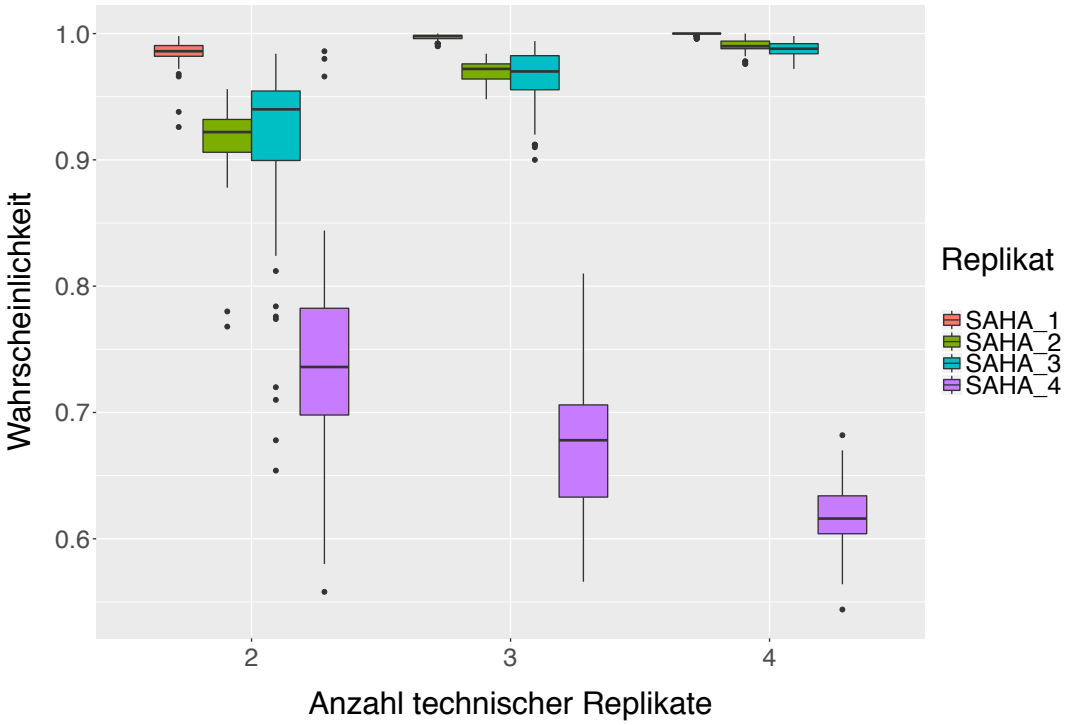

Abbildung 71: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

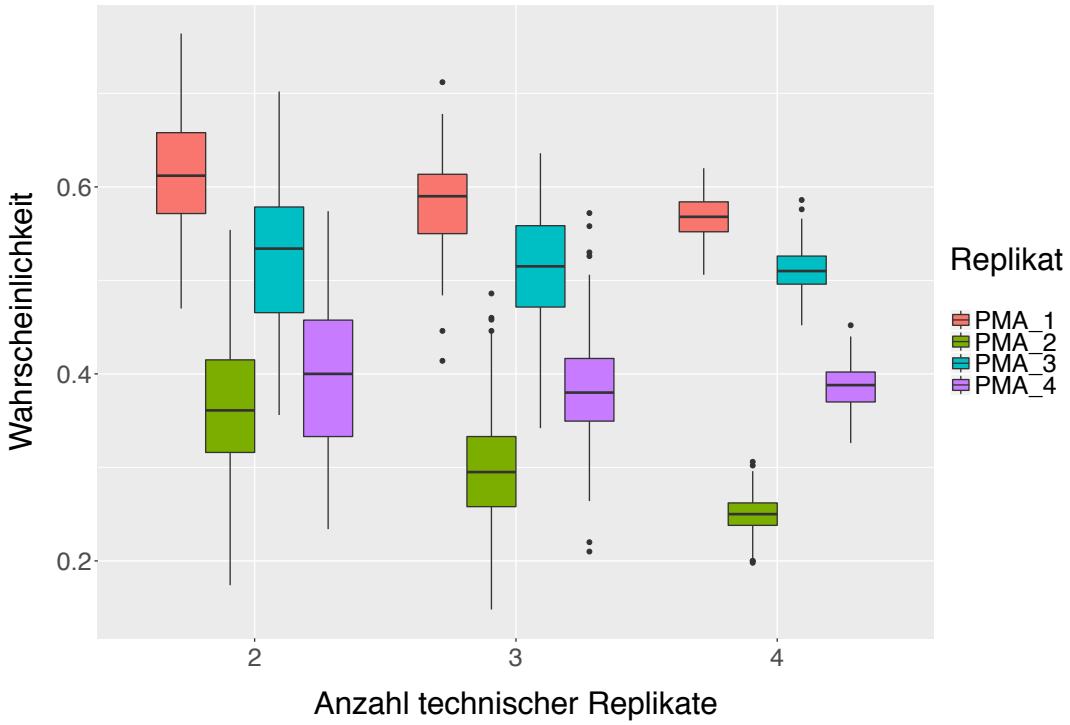

Abbildung 72: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

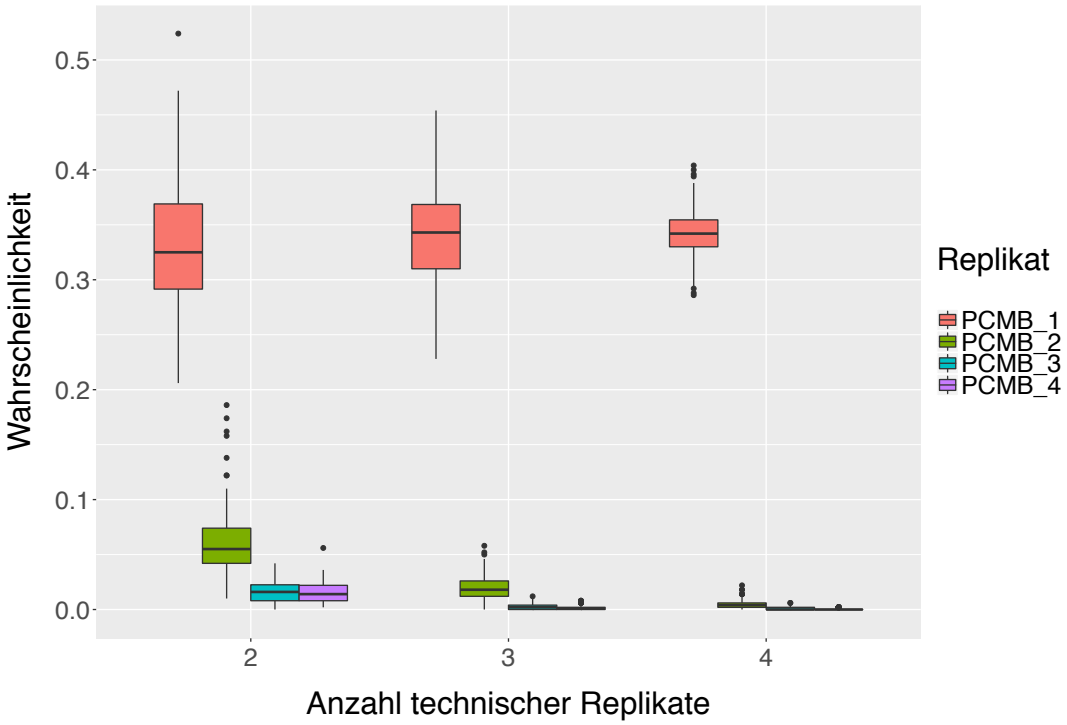

Abbildung 73: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

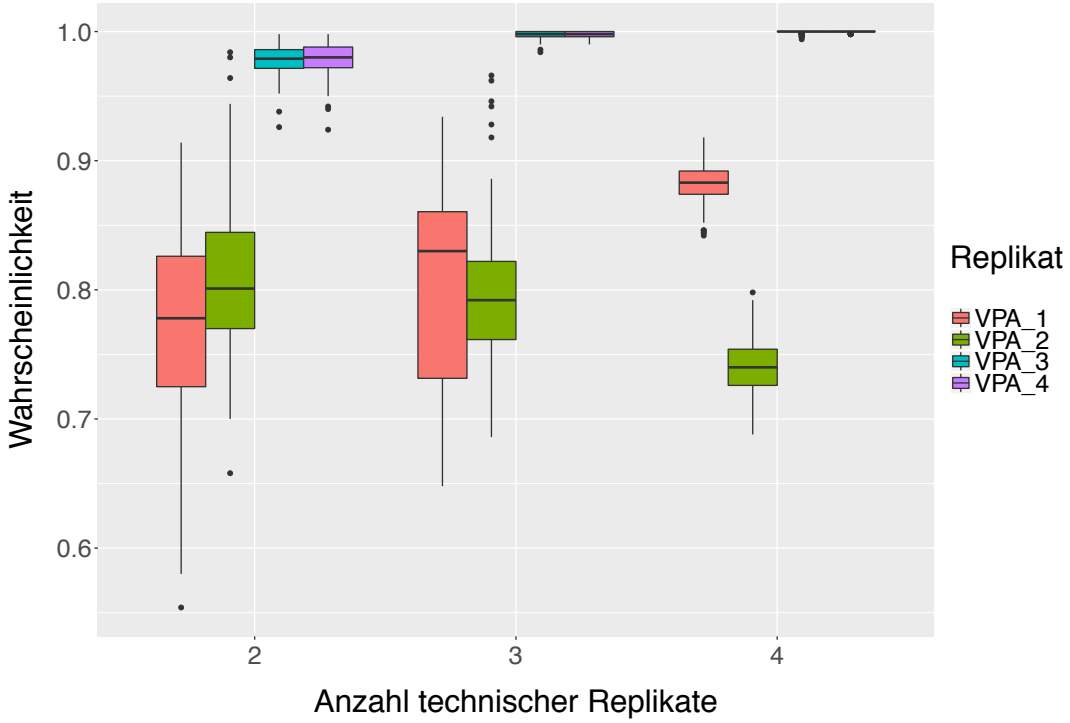

Abbildung 74: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

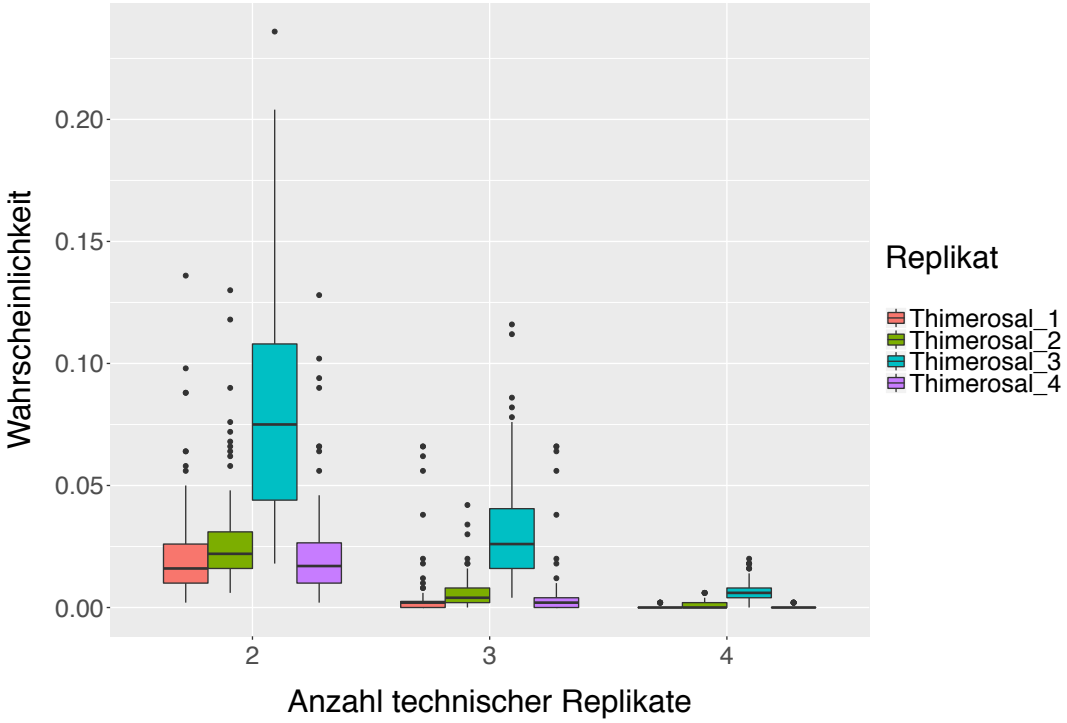

Abbildung 75: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thimerosal. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

6.3.3 Einfluss technischer Replikate auf UKK SVM

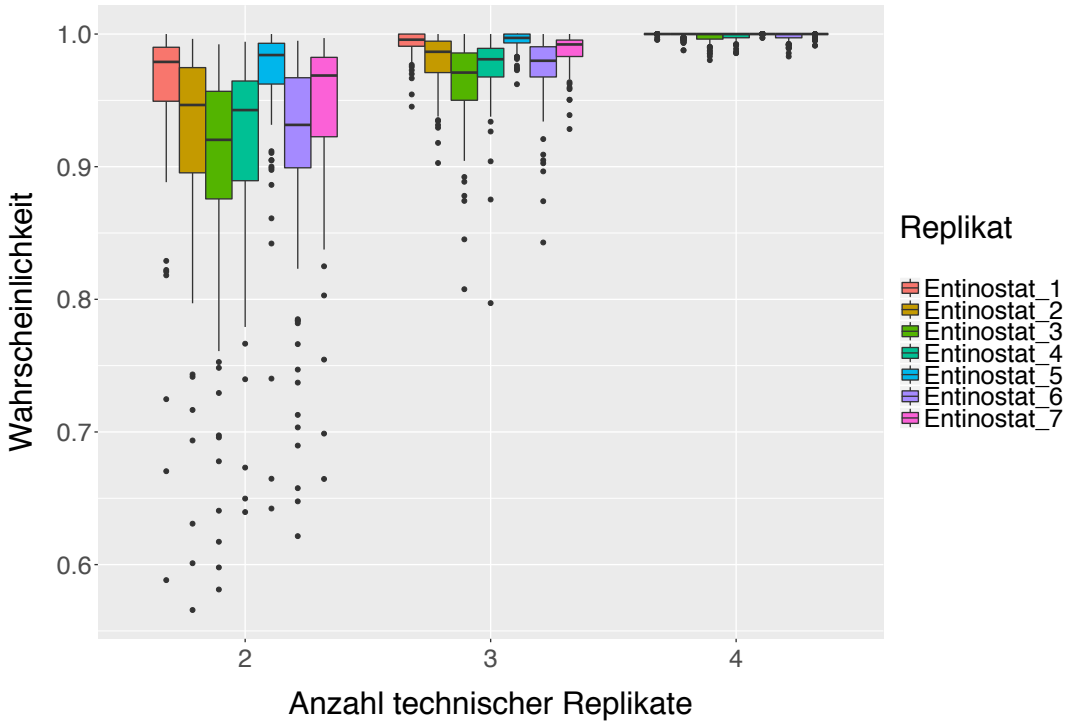

Abbildung 76: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

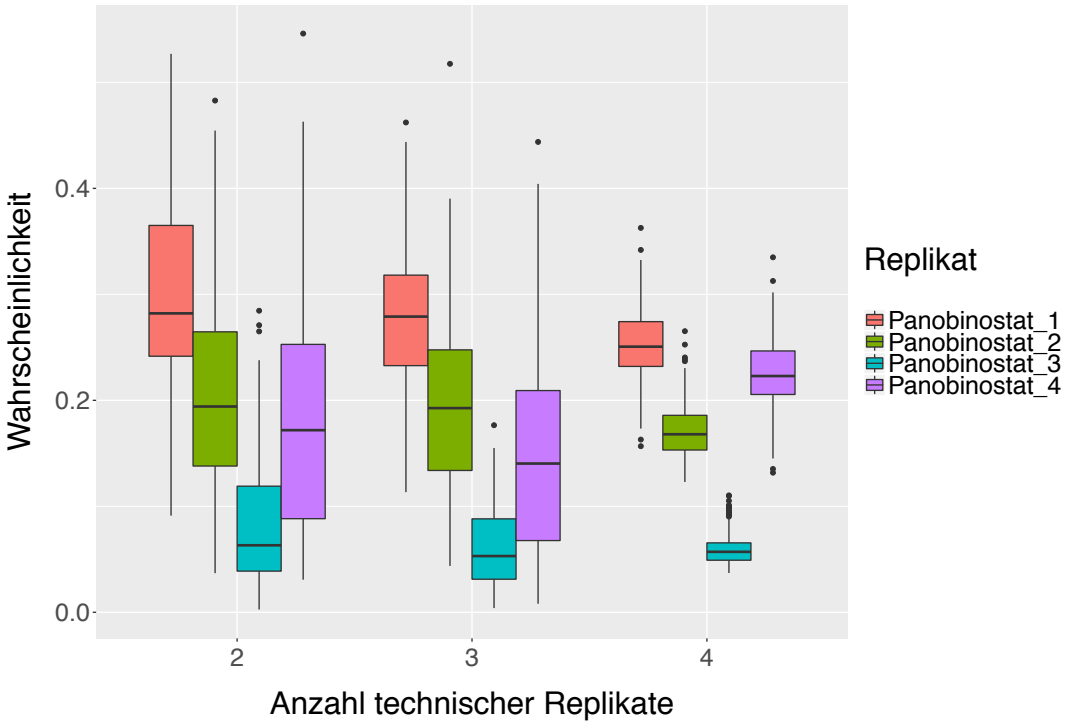

Abbildung 77: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

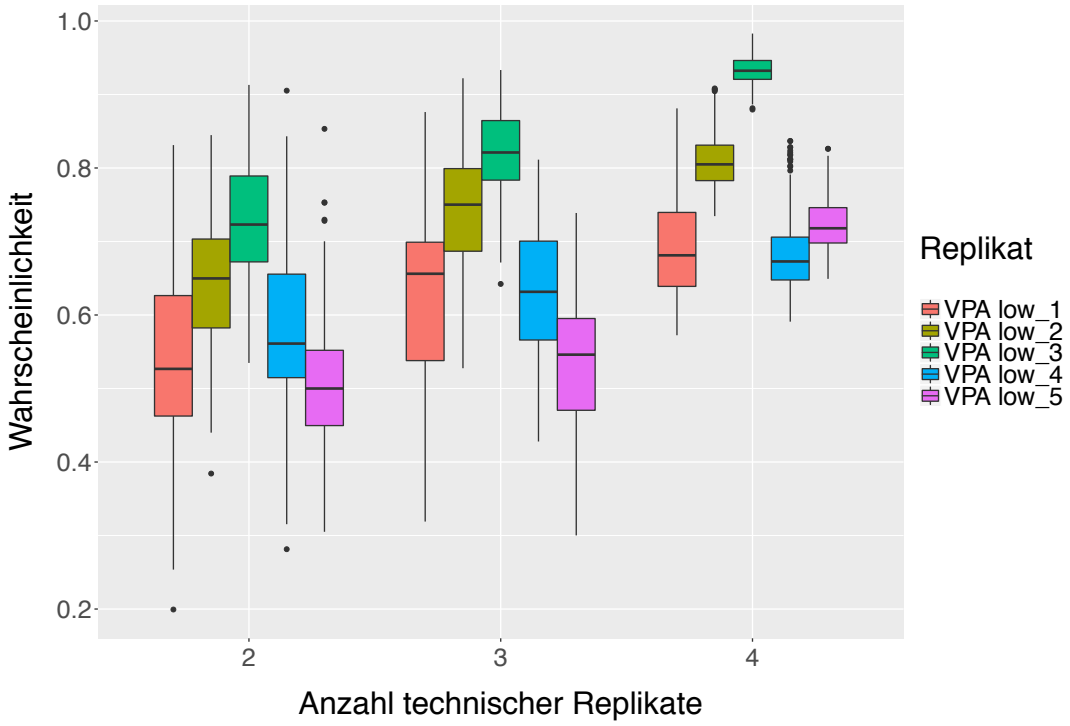

Abbildung 78: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA low. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

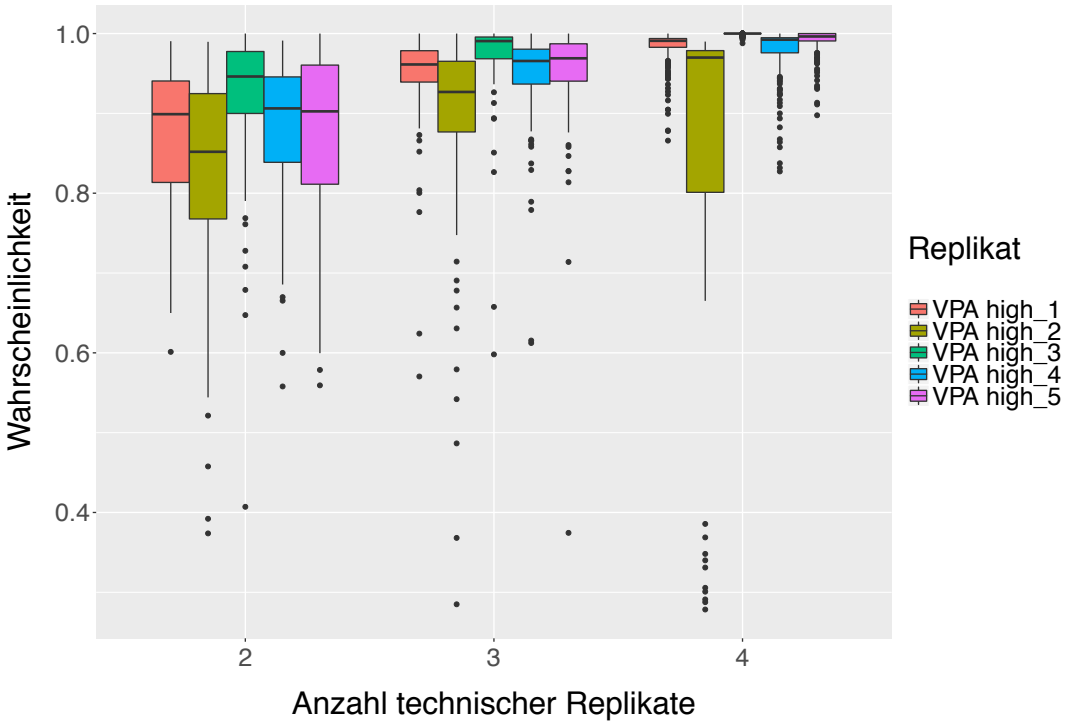

Abbildung 79: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA high. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

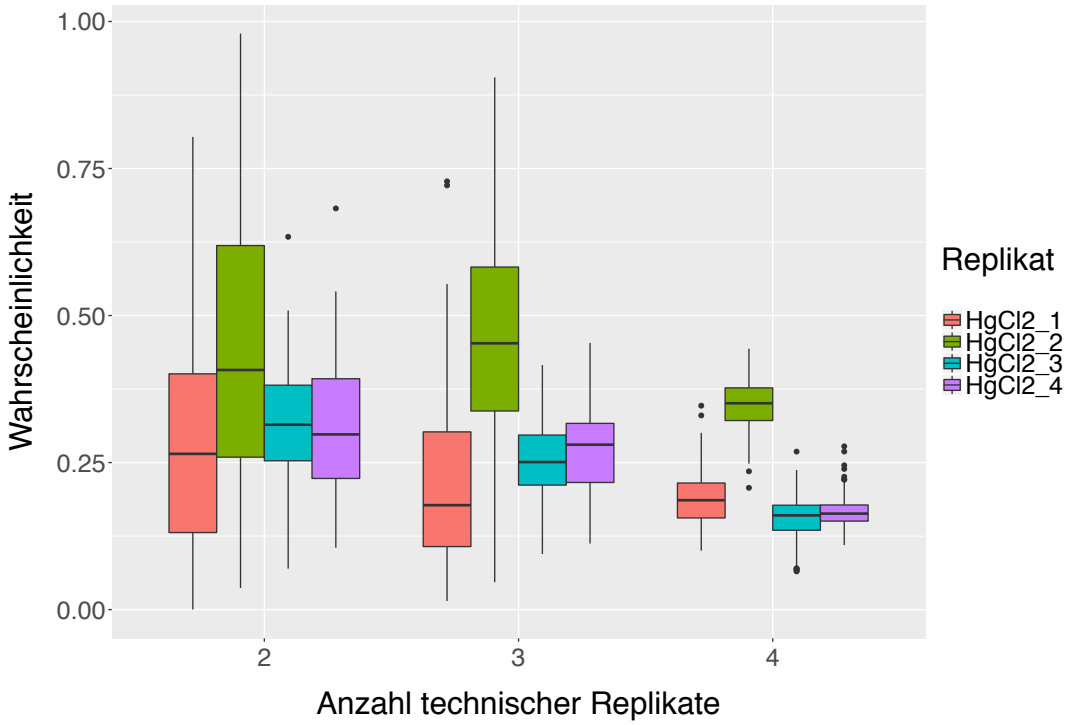

Abbildung 80: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgCl2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

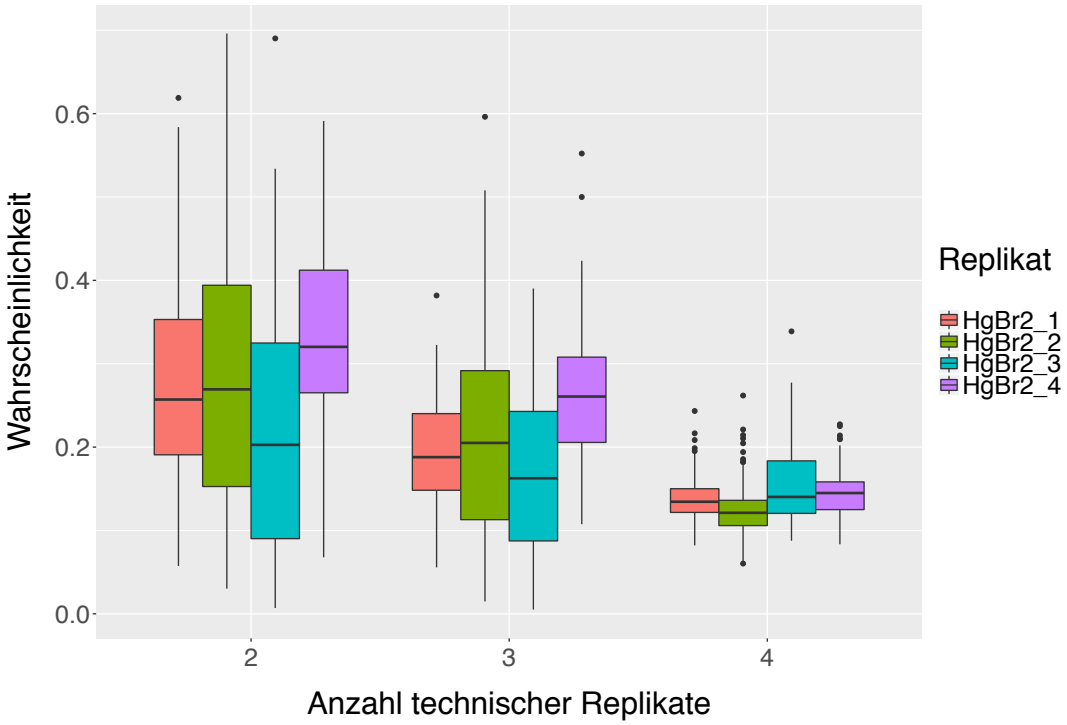

Abbildung 81: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgBr2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

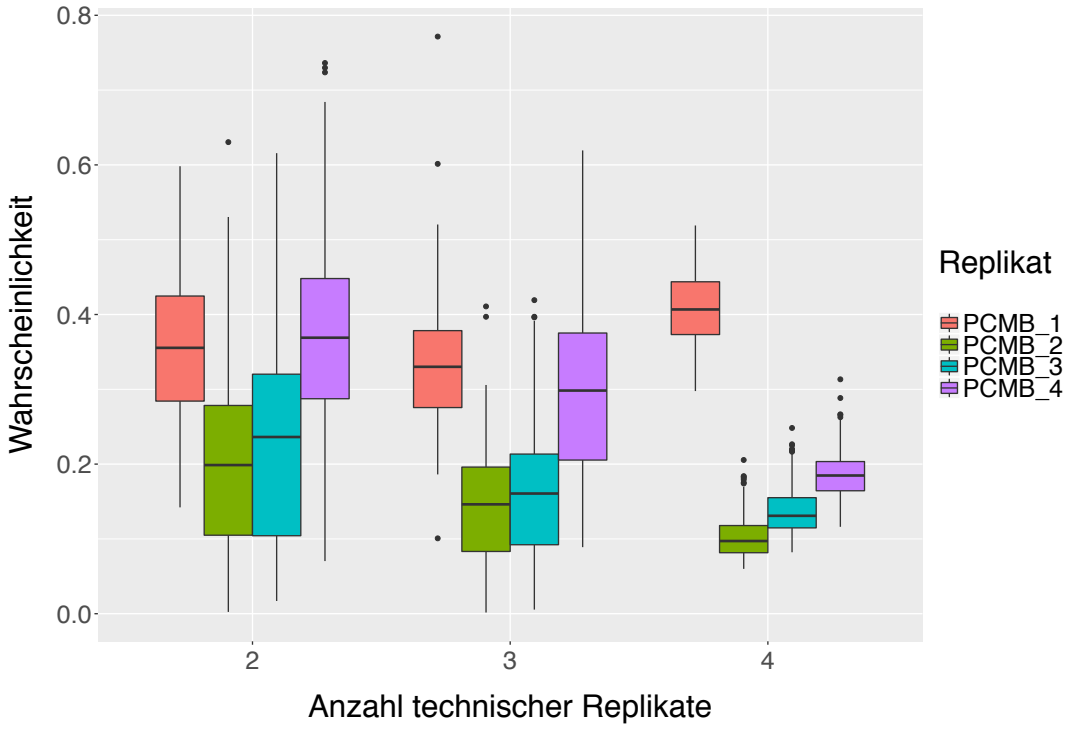

Abbildung 82: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

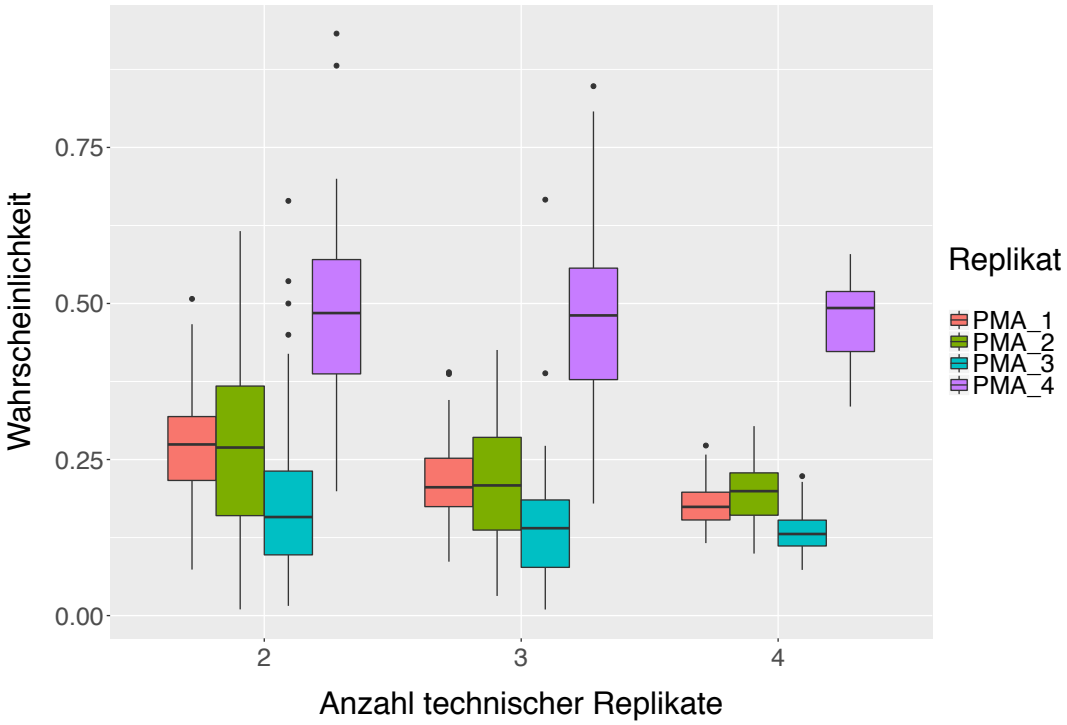

Abbildung 83: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

Abbildung 84: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

Abbildung 85: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

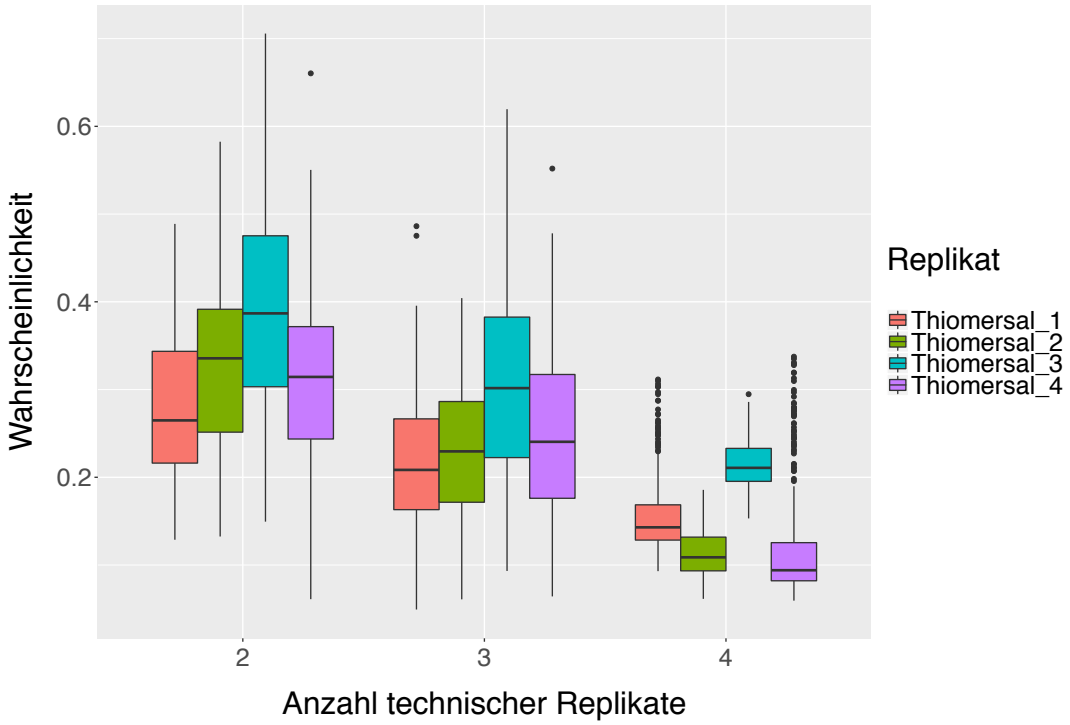

Abbildung 86: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

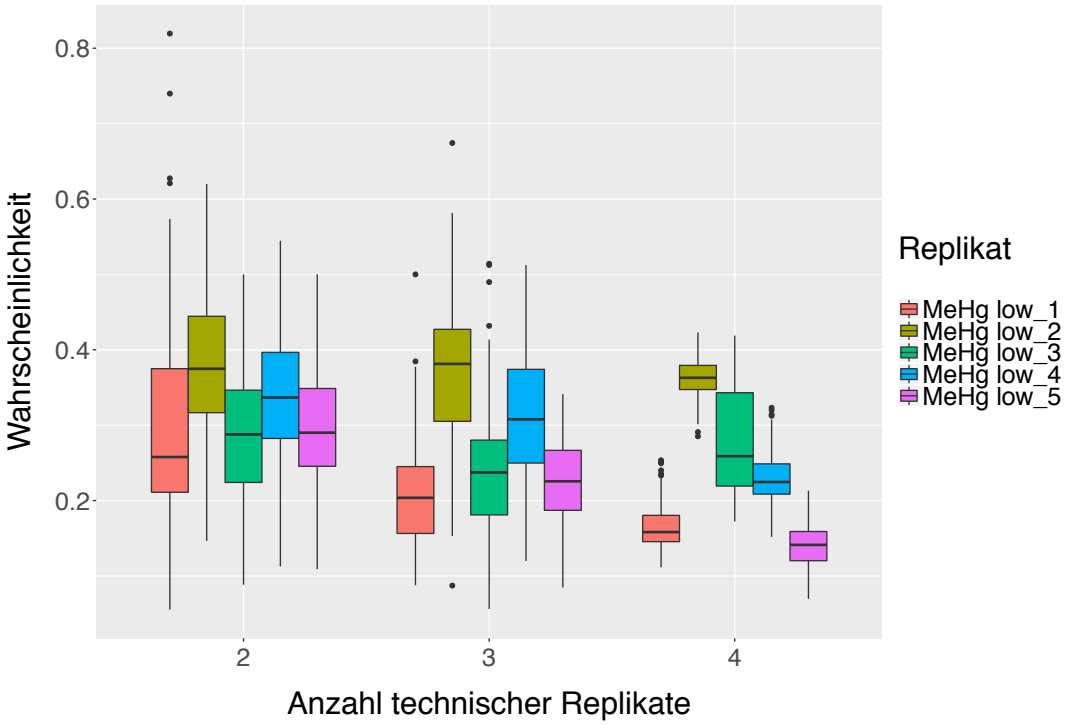

Abbildung 87: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg low. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

Abbildung 88: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg high. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

6.3.4 Einfluss technischer Replikate auf UKK RF

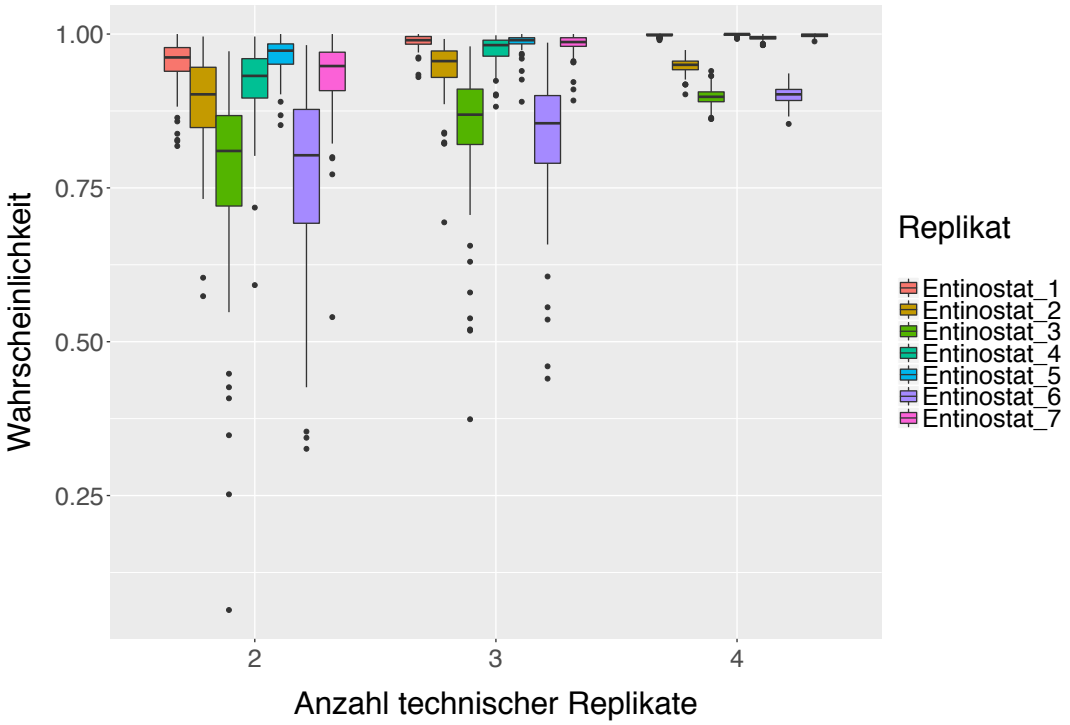

Abbildung 89: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

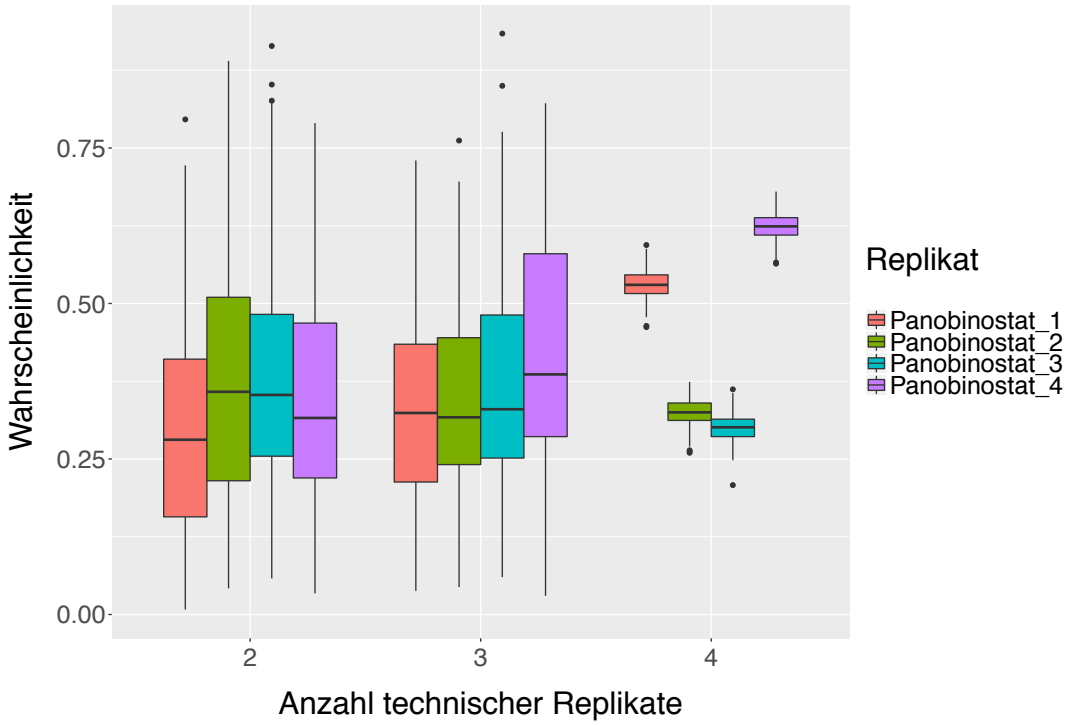

Abbildung 90: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

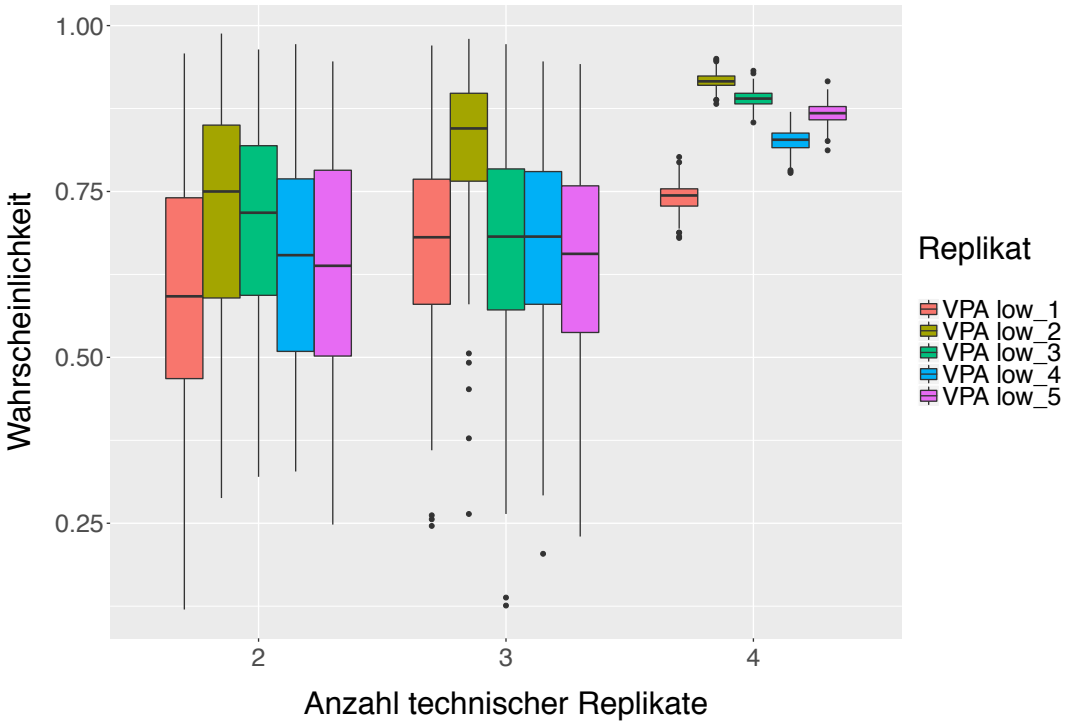

Abbildung 91: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA low. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

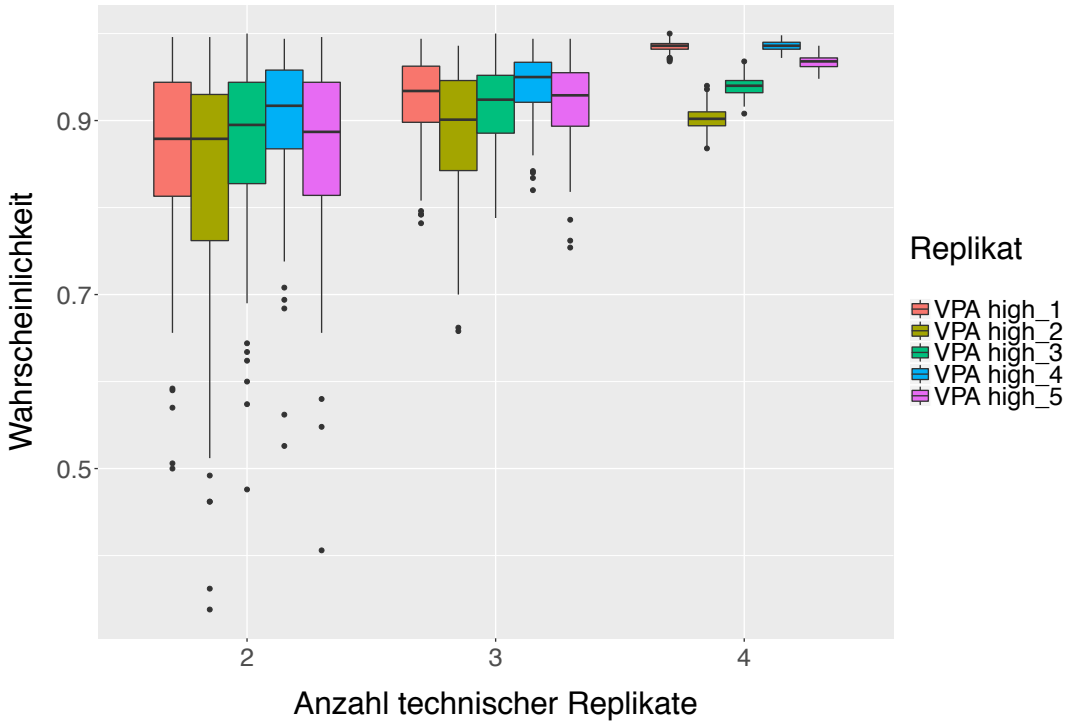

Abbildung 92: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA high. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

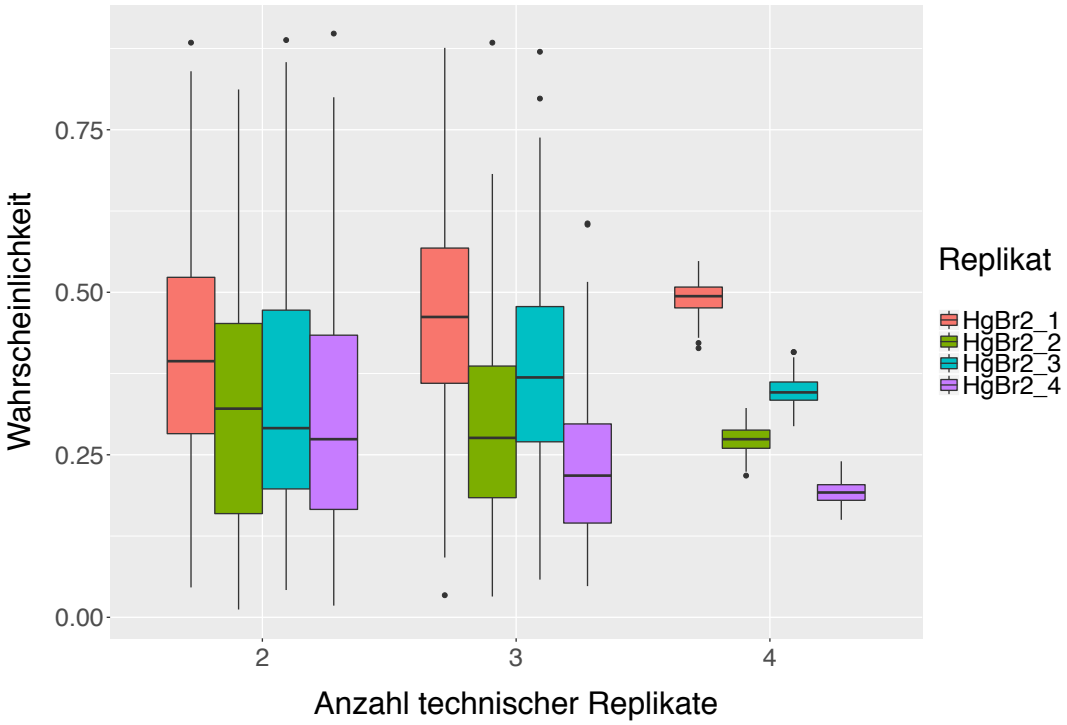

Abbildung 93: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von HgBr2. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

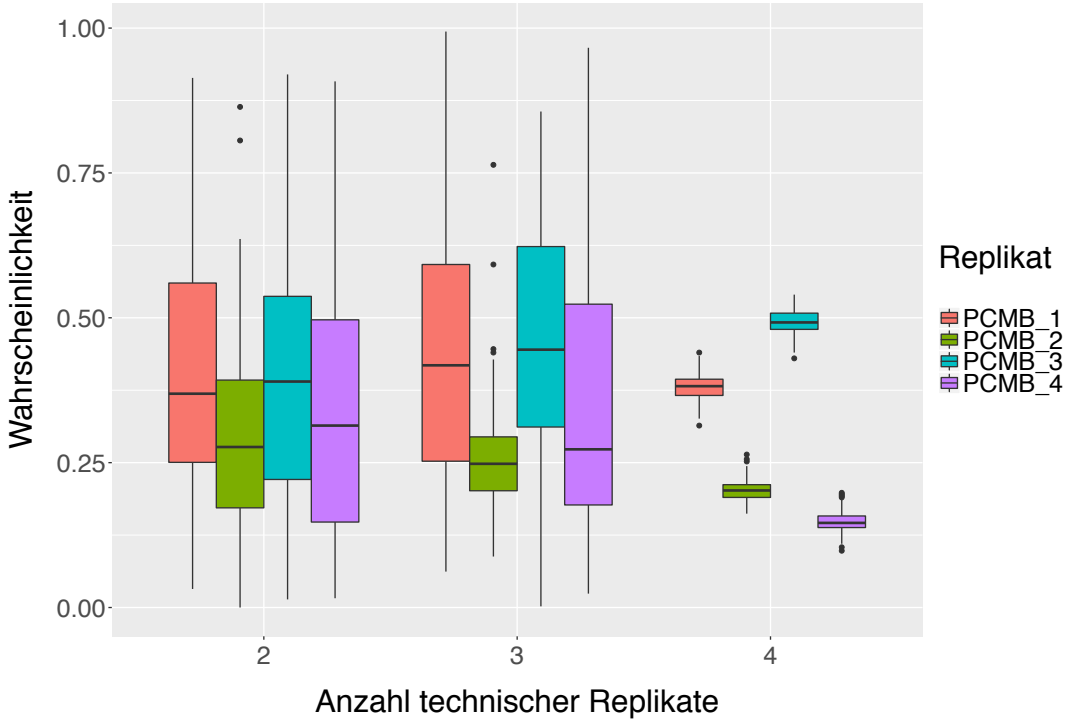

Abbildung 94: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

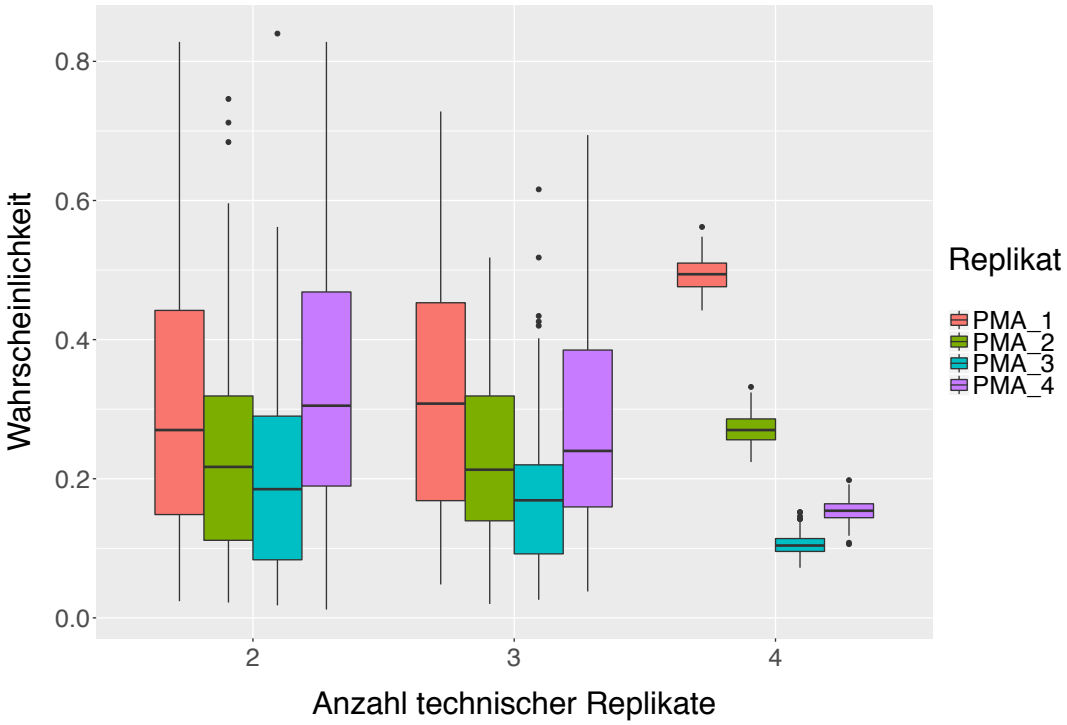

Abbildung 95: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

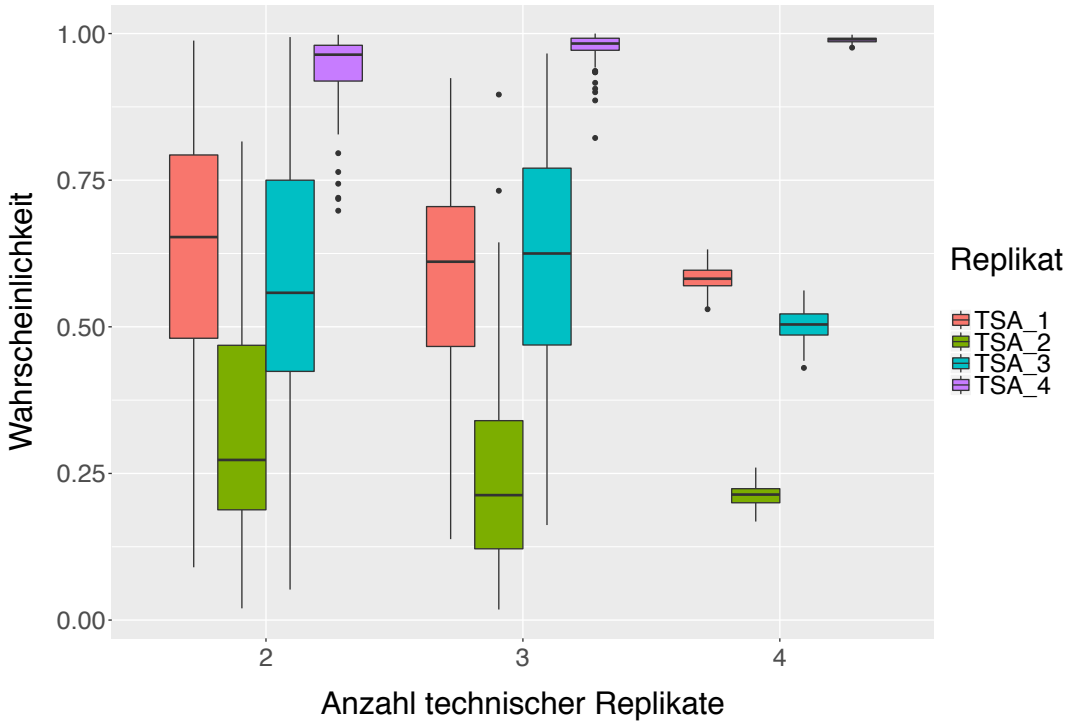

Abbildung 96: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

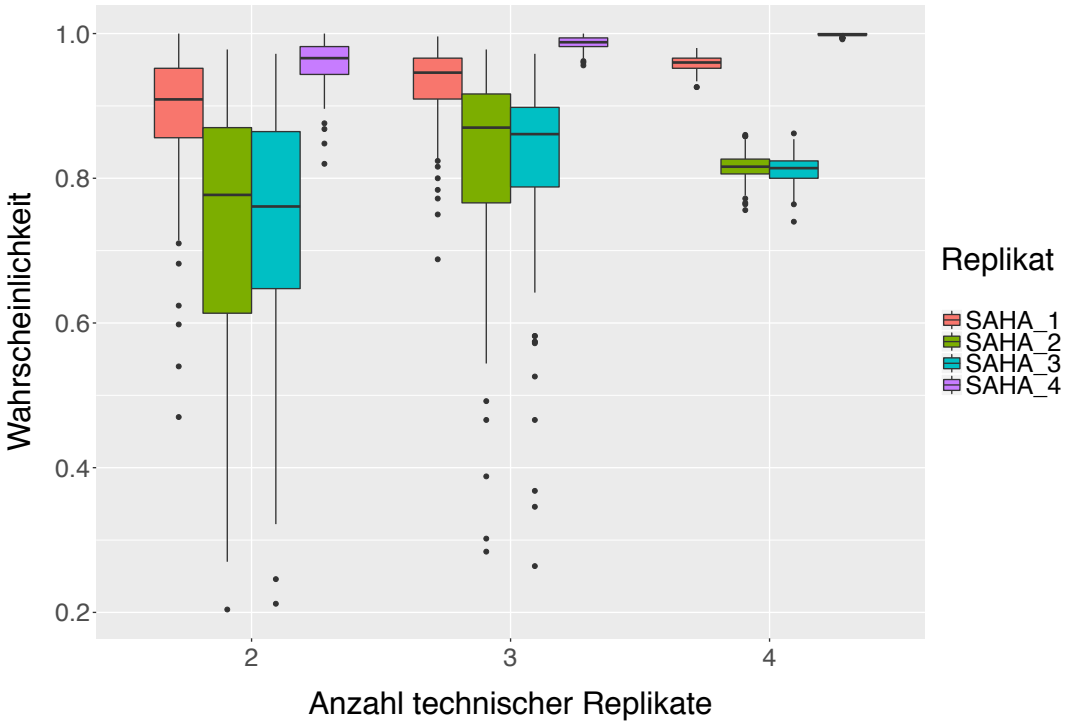

Abbildung 97: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

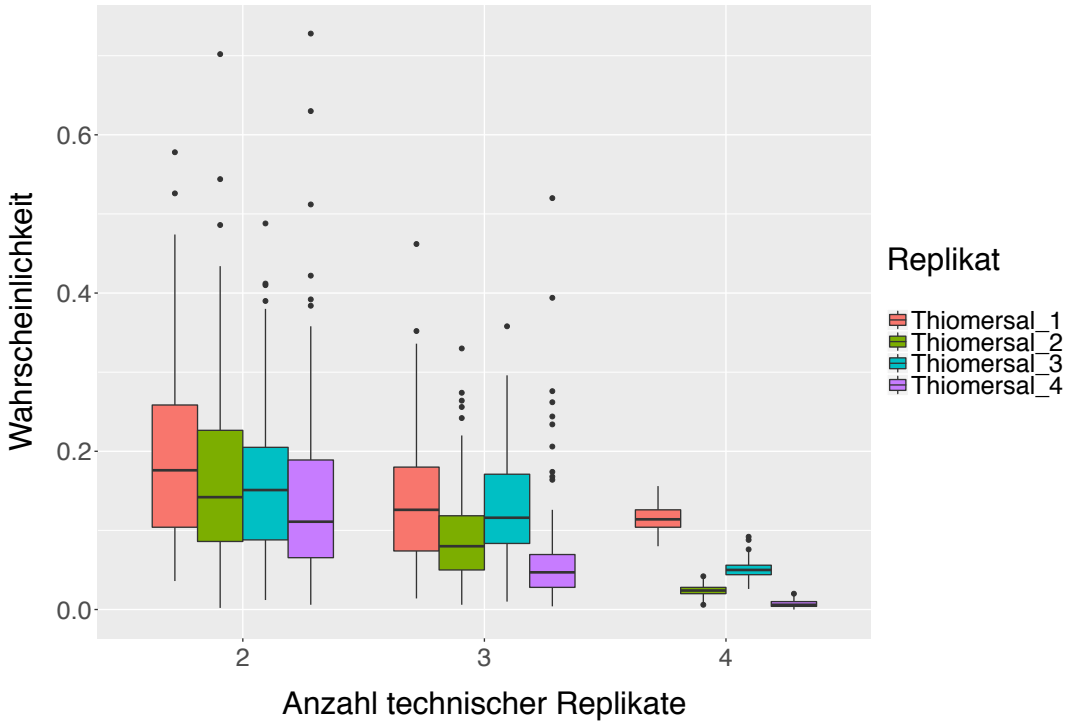

Abbildung 98: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

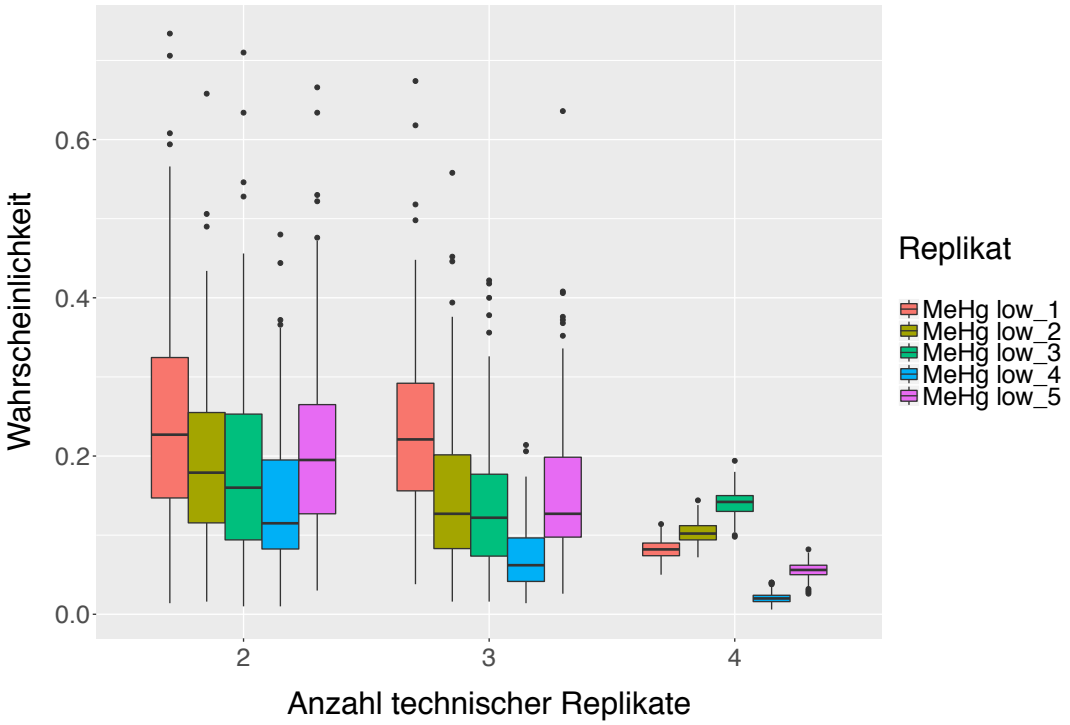

Abbildung 99: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg low. Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anzahl der verwendeter technischer Replikate in der Trainingsmenge, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen.

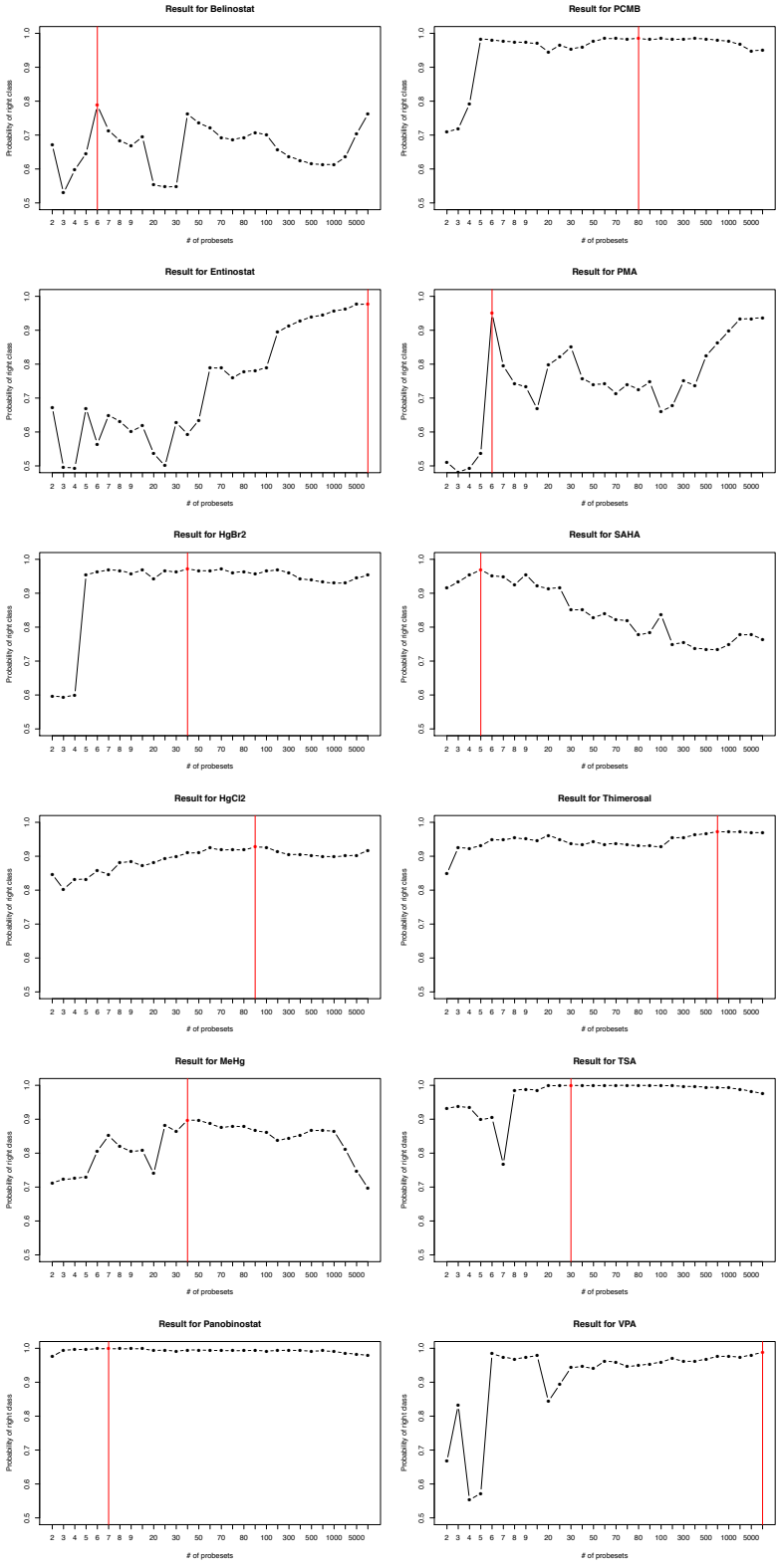

Abbildung 100: Genauigkeit der Vorhersage aller 12 Substanzen in Abhängigkeit von der Anzahl der in die Analyse aufgenommenen Probesets. Auf der x-Achse ist die Anzahl der Variablen (nicht maßstabsgetreu), auf der y-Achse die Vorhersagegenauigkeit abgetragen. Mit der roten Linie ist diejenige(n) Anzahl(en) gekennzeichnet, für welche(n) die maximale Genauigkeit erreicht wurde.

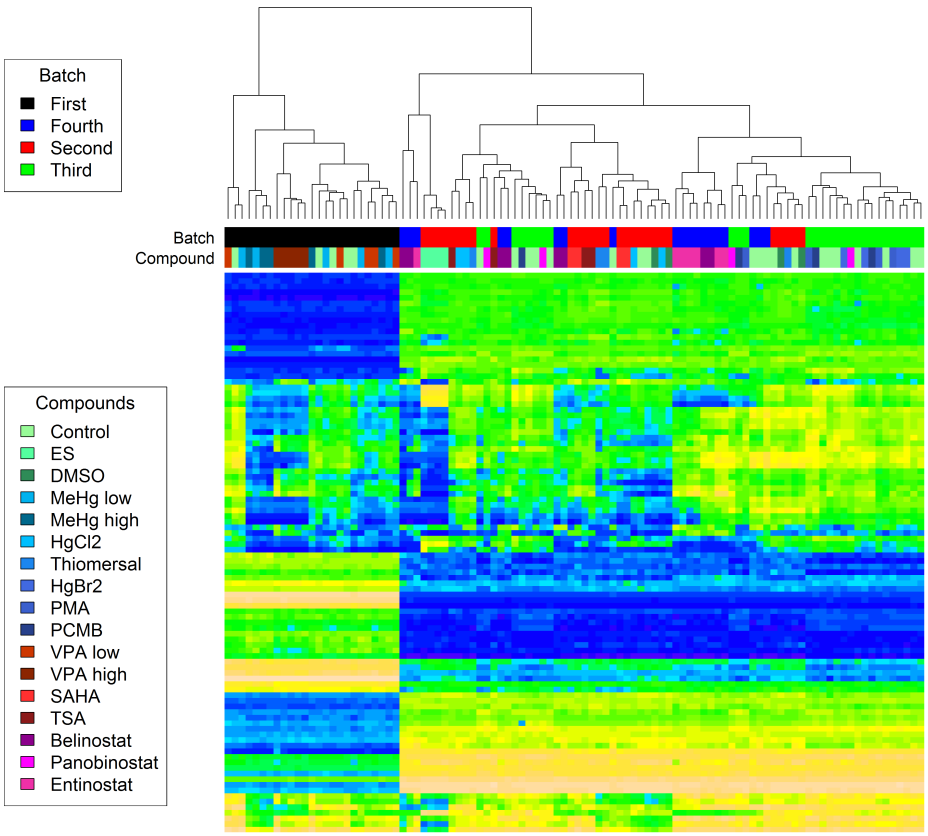

Abbildung 101: Heatmap der UKK Klassifikationsstudie

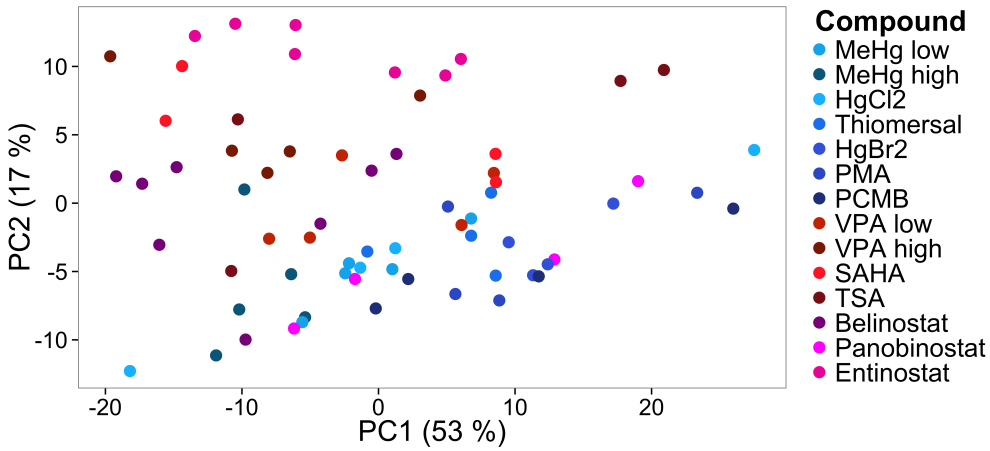

Abbildung 102: Hauptkomponentenplot der ComBat-transformierten Daten der UKN1 Klassifikationsstudie nach Abzug der Kontrollen

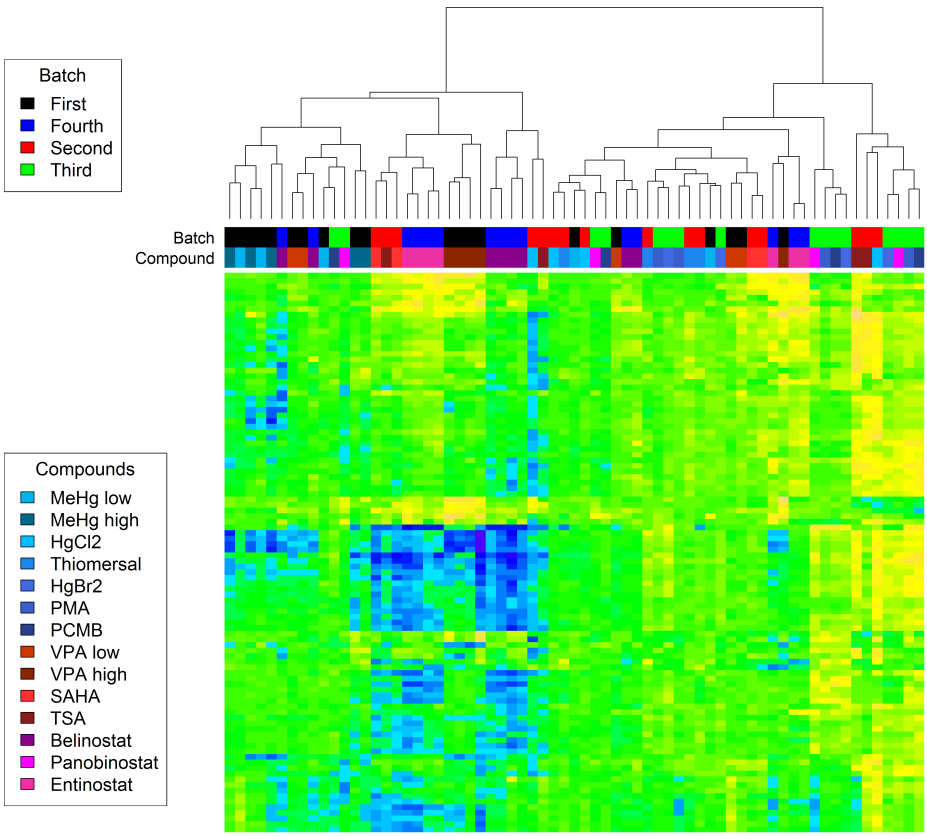

Abbildung 103: Heatmap der UKK Klassifikationsstudie nach Abzug der Kontrollen

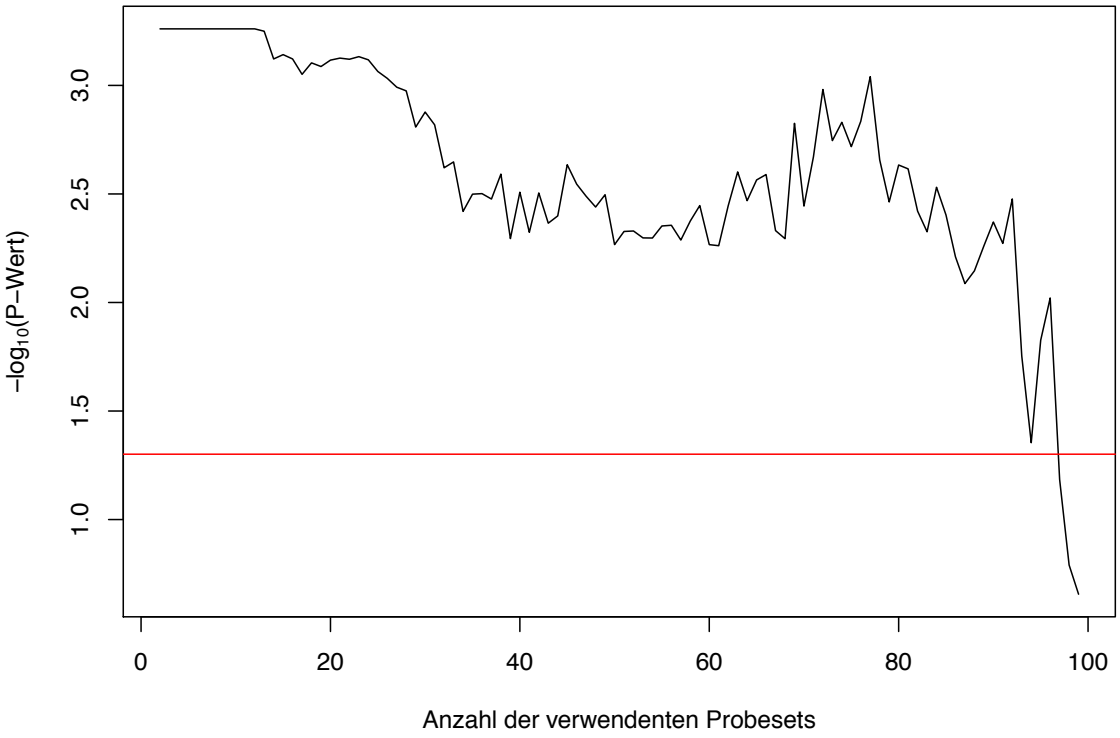

Abbildung 104: Log-transformierte P-Werte des gepaarten t-Tests für die verschiedenen Varianten des Random Forest Verfahrens. Auf der x-Achse ist die Anzahl der verwendetet Probesets abgetragen, auf der y-Achse sind die log-transformierten P-Werte abgetragen. Die rote horizontale Linie bezeichnet den Wert y=0.05.

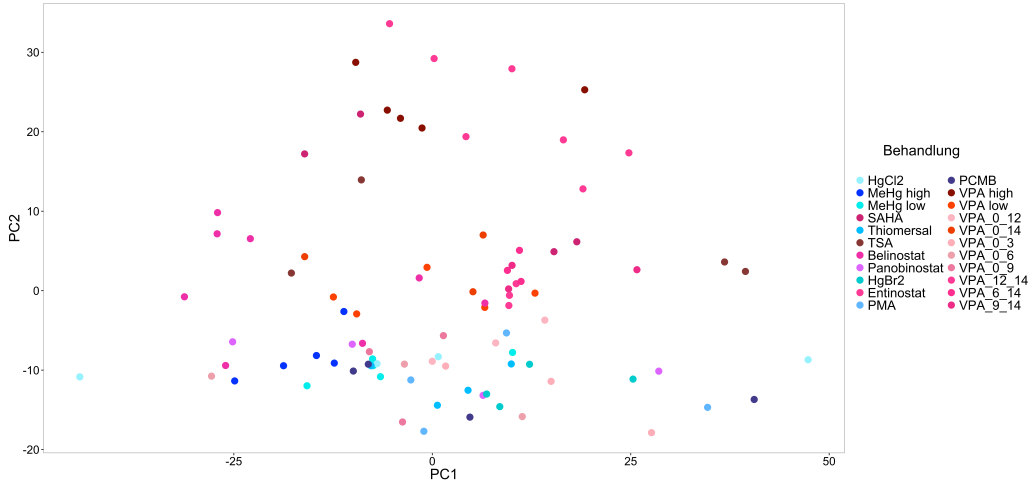

Abbildung 105: Der Hauptkomponentenplot der Daten aus der UKK Klassifikationsstudie. Die Daten aus der VPA Zeitfenserstudie sind mit Hilfe der Rotationsmatrix reinprojiziert.

6.3.5 Rauschplots für UKN1 SVM

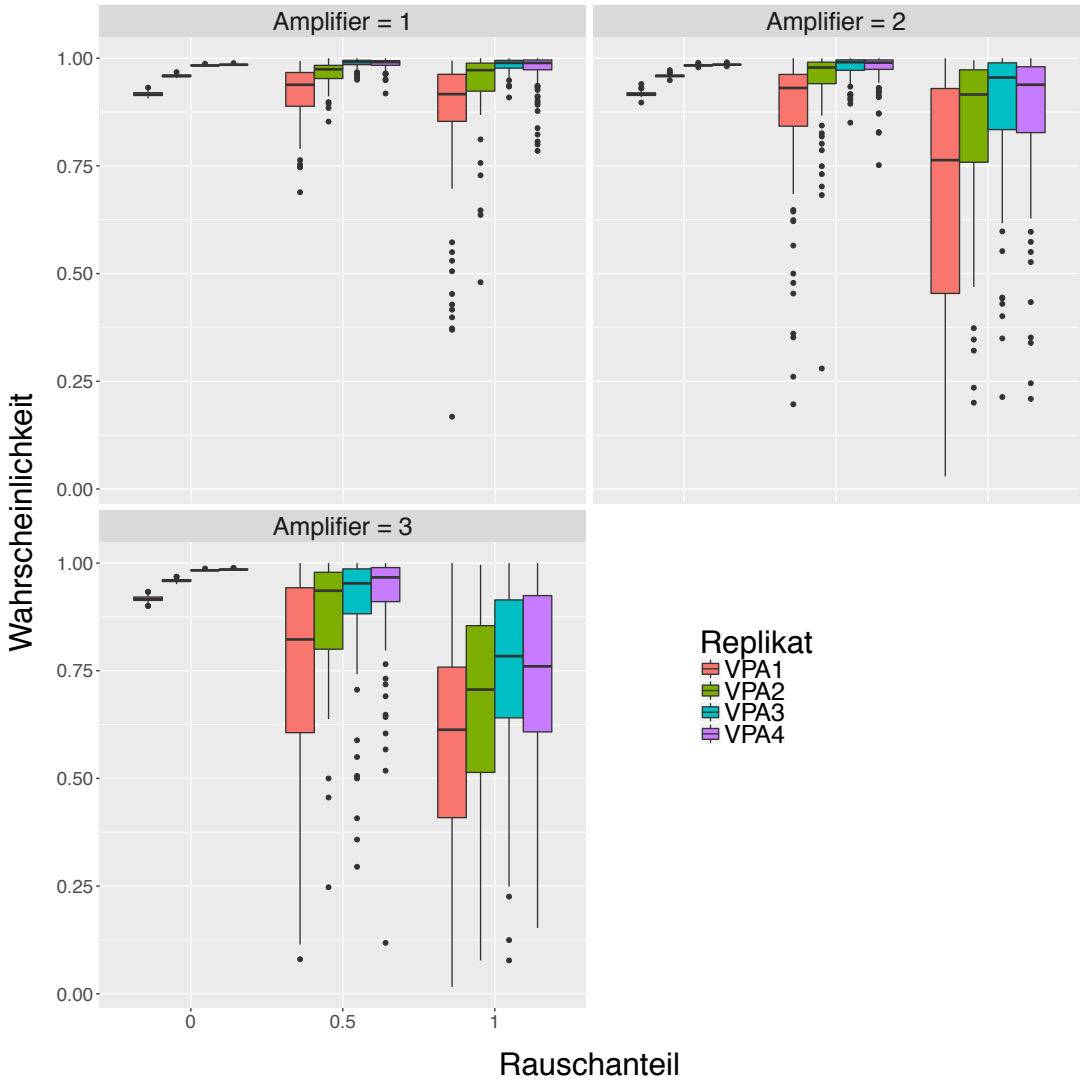

Abbildung 106: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

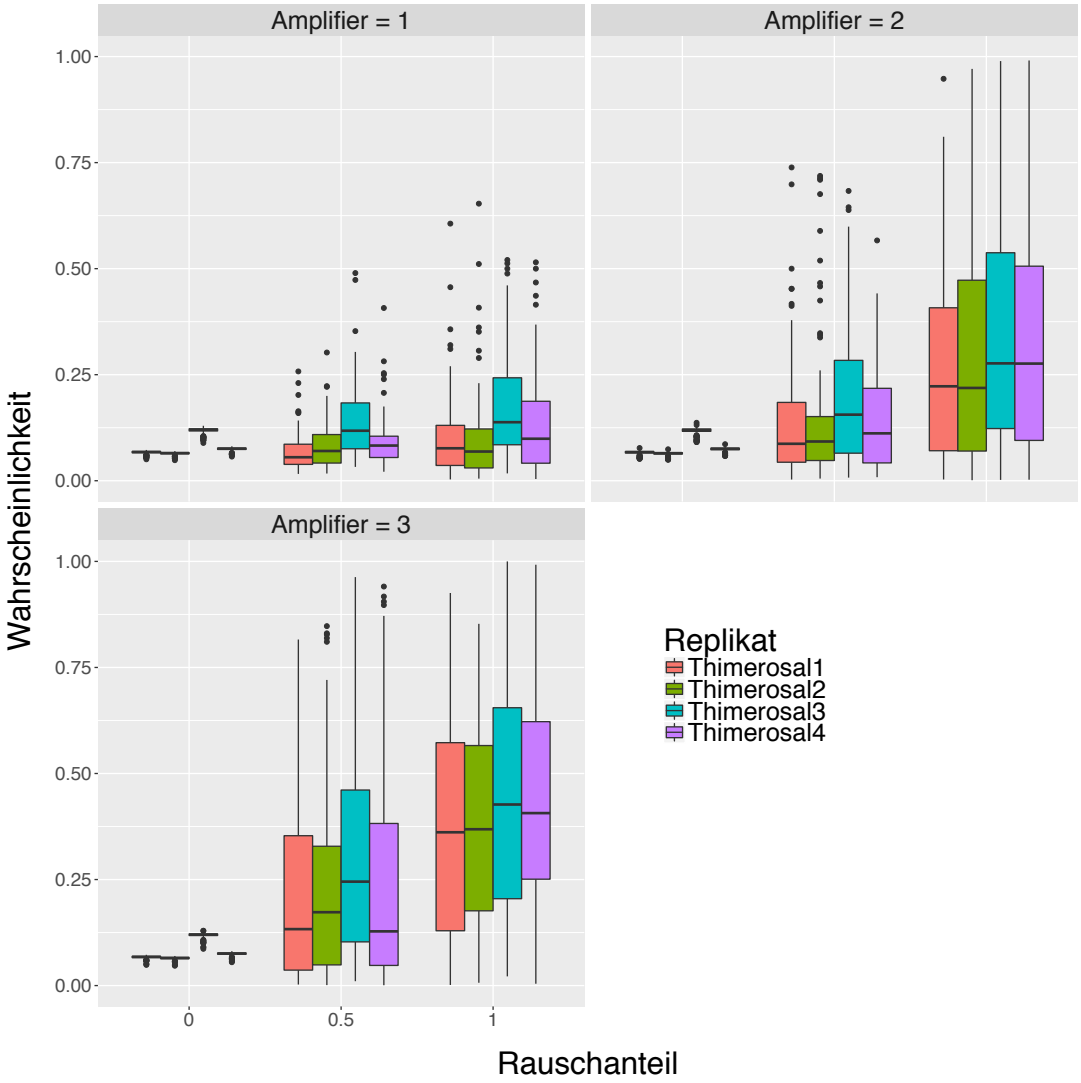

Abbildung 107: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

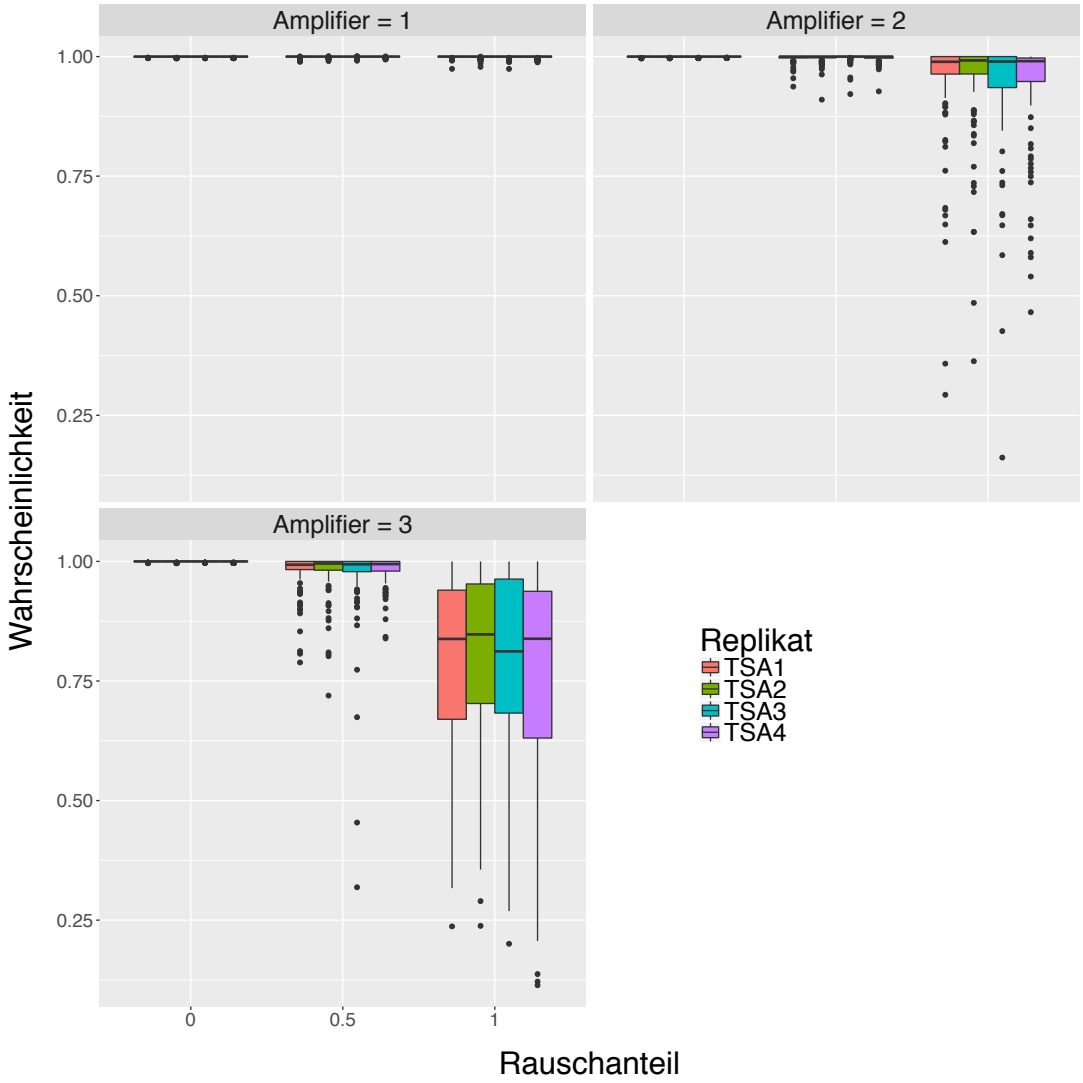

Abbildung 108: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

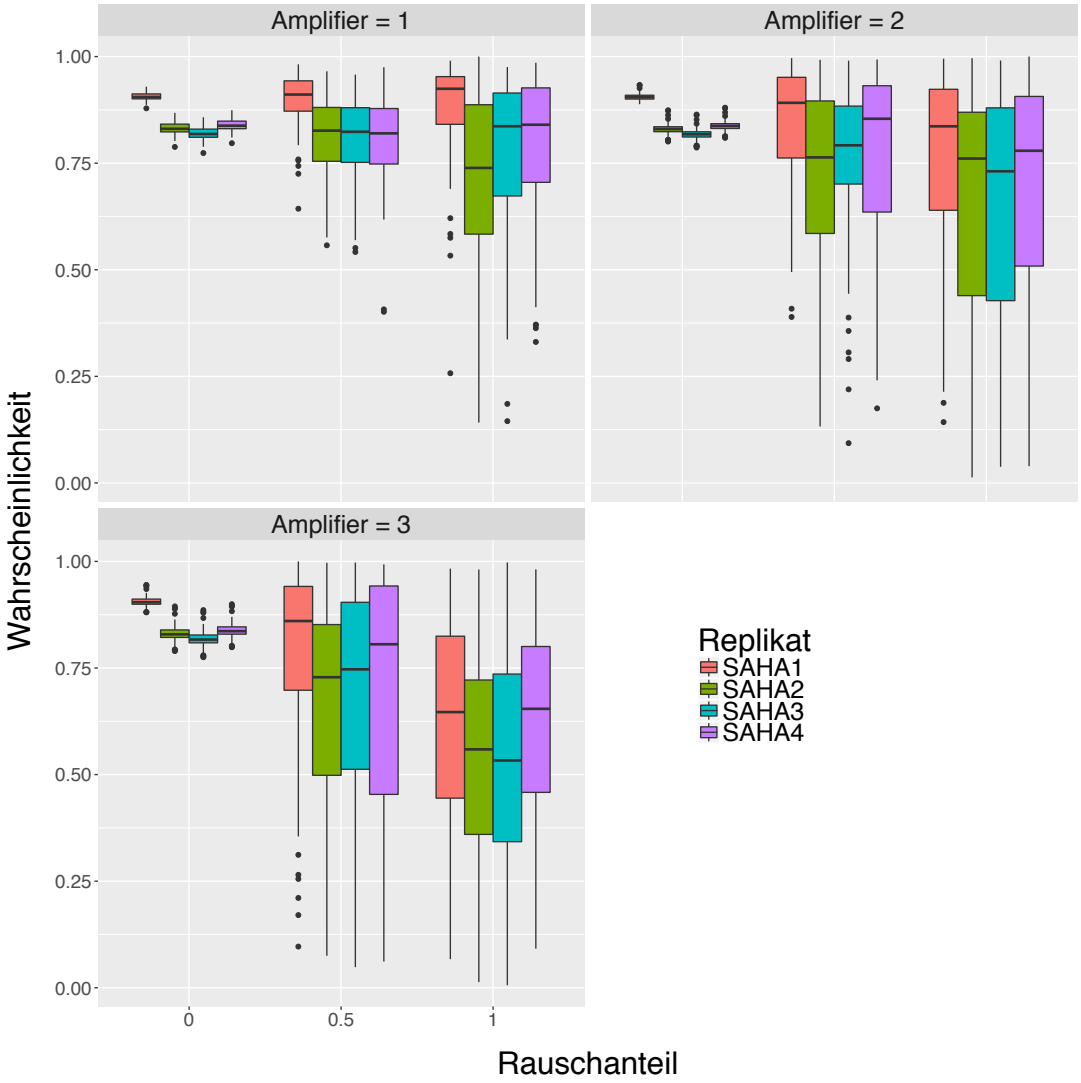

Abbildung 109: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

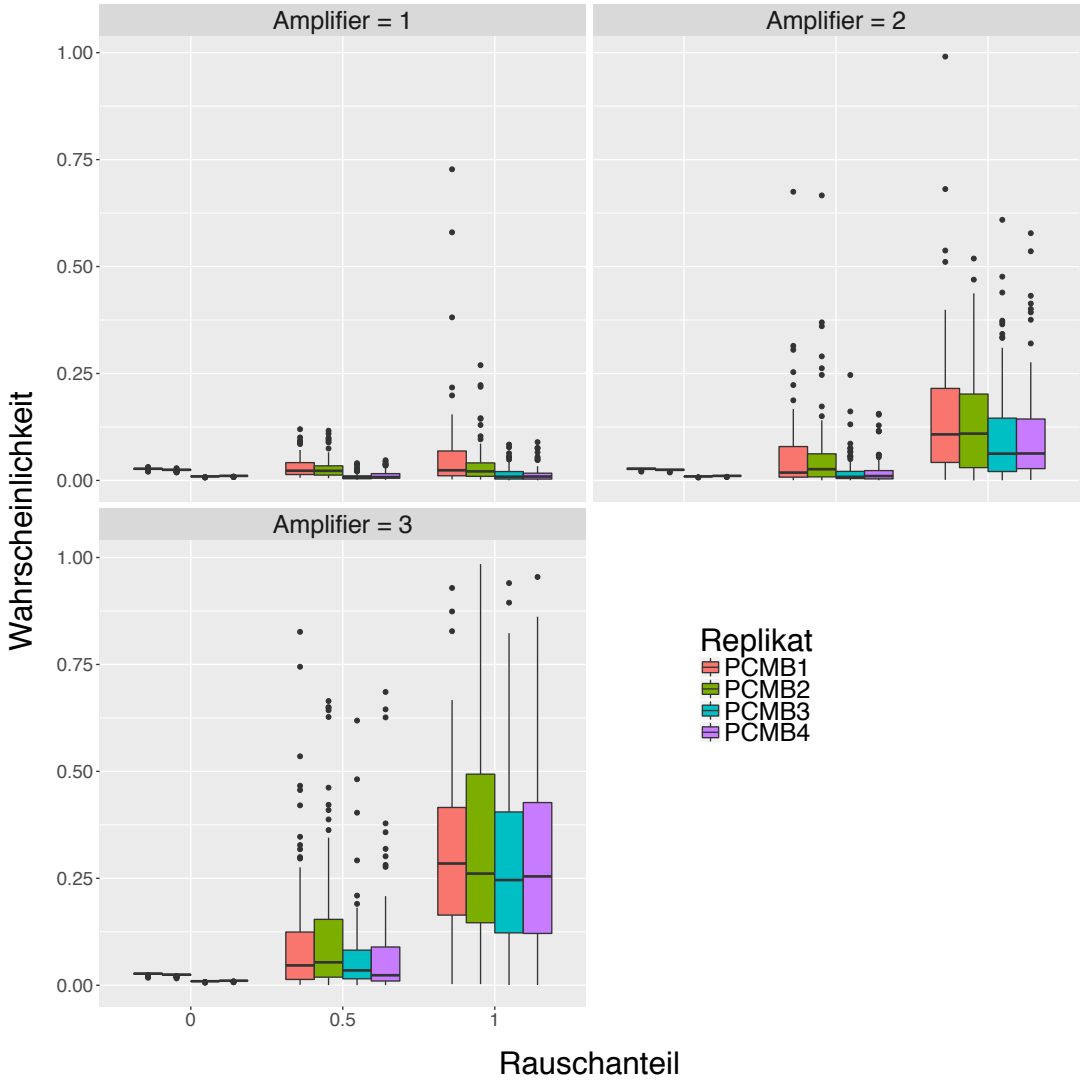

Abbildung 110: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

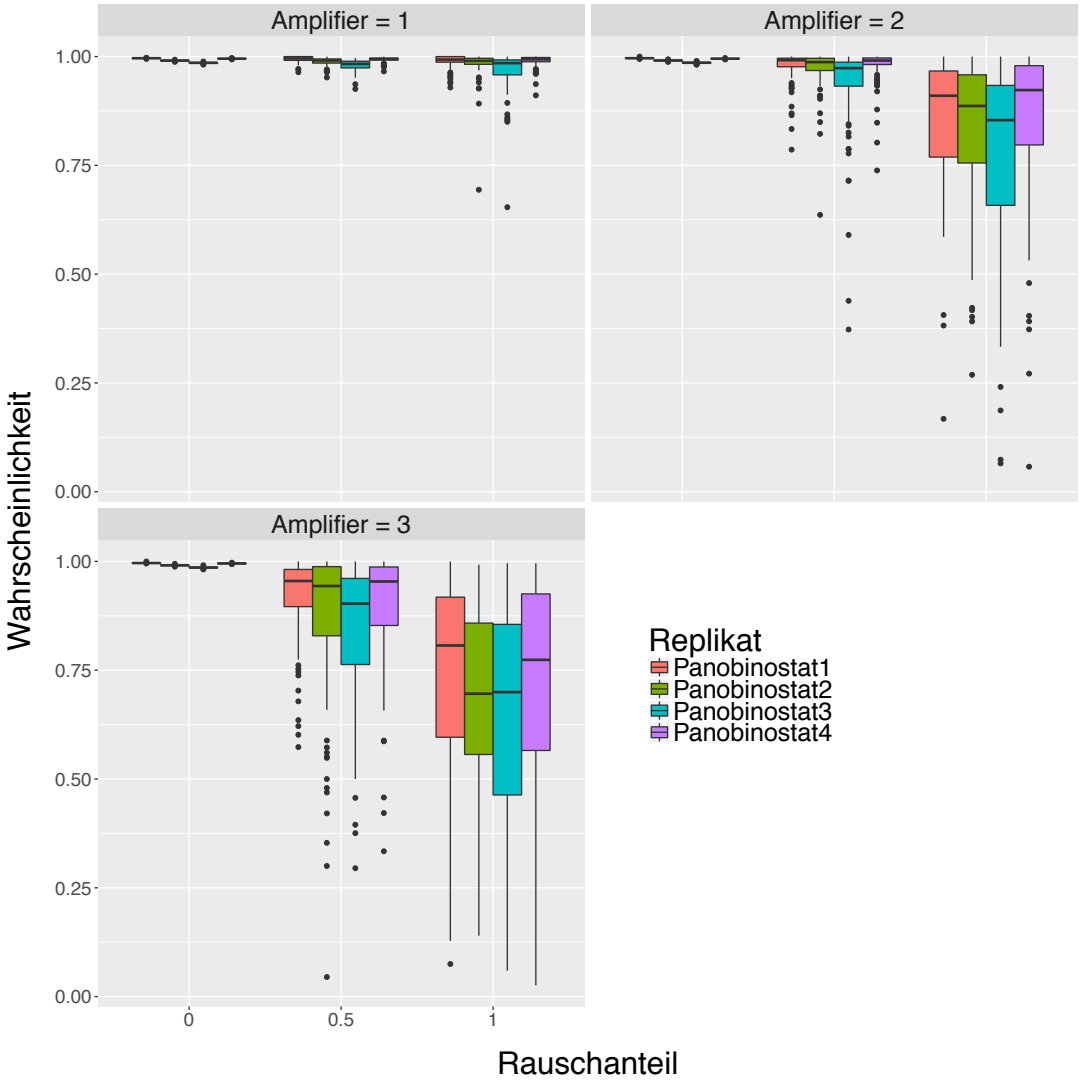

Abbildung 111: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

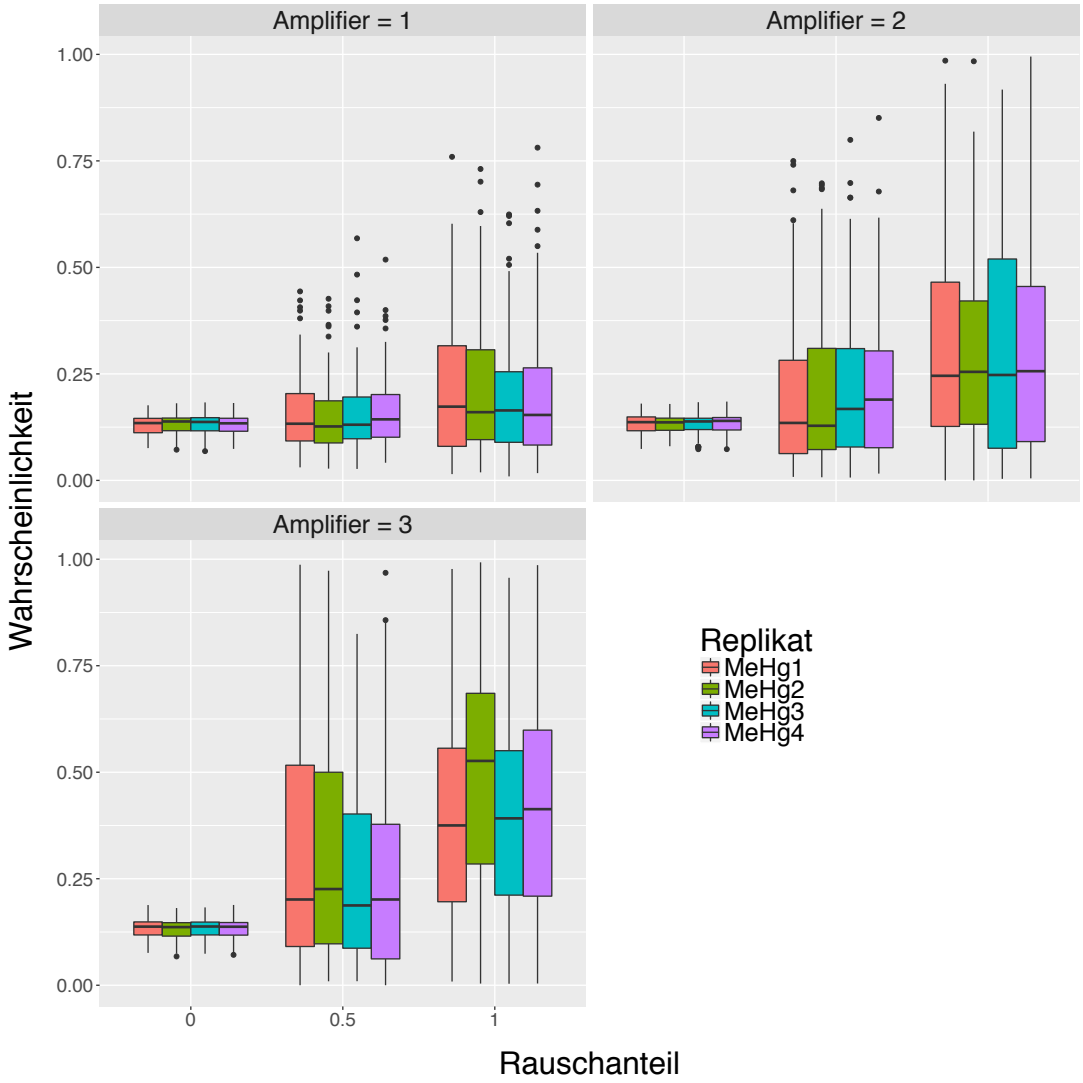

Abbildung 112: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

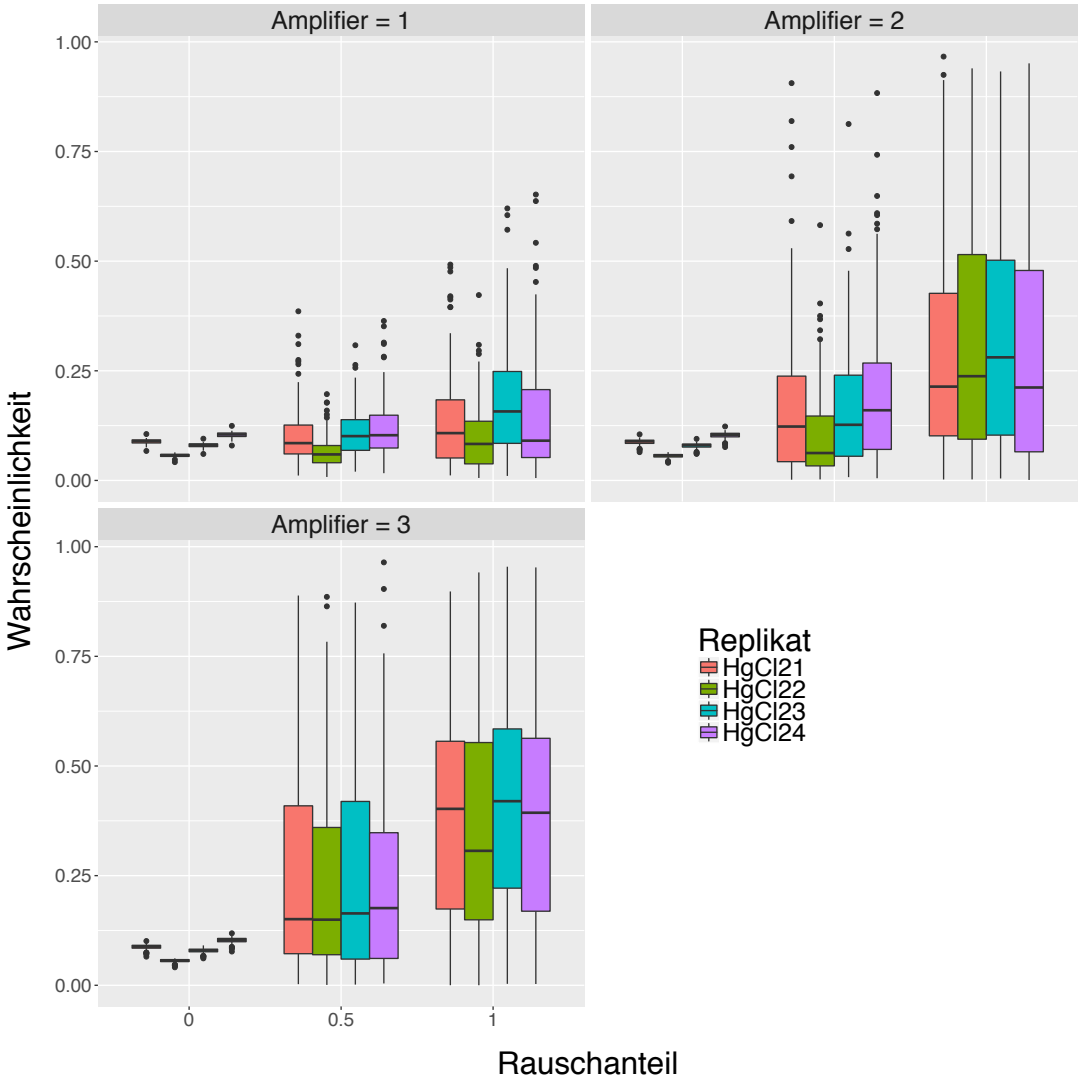

Abbildung 113: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgCl_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

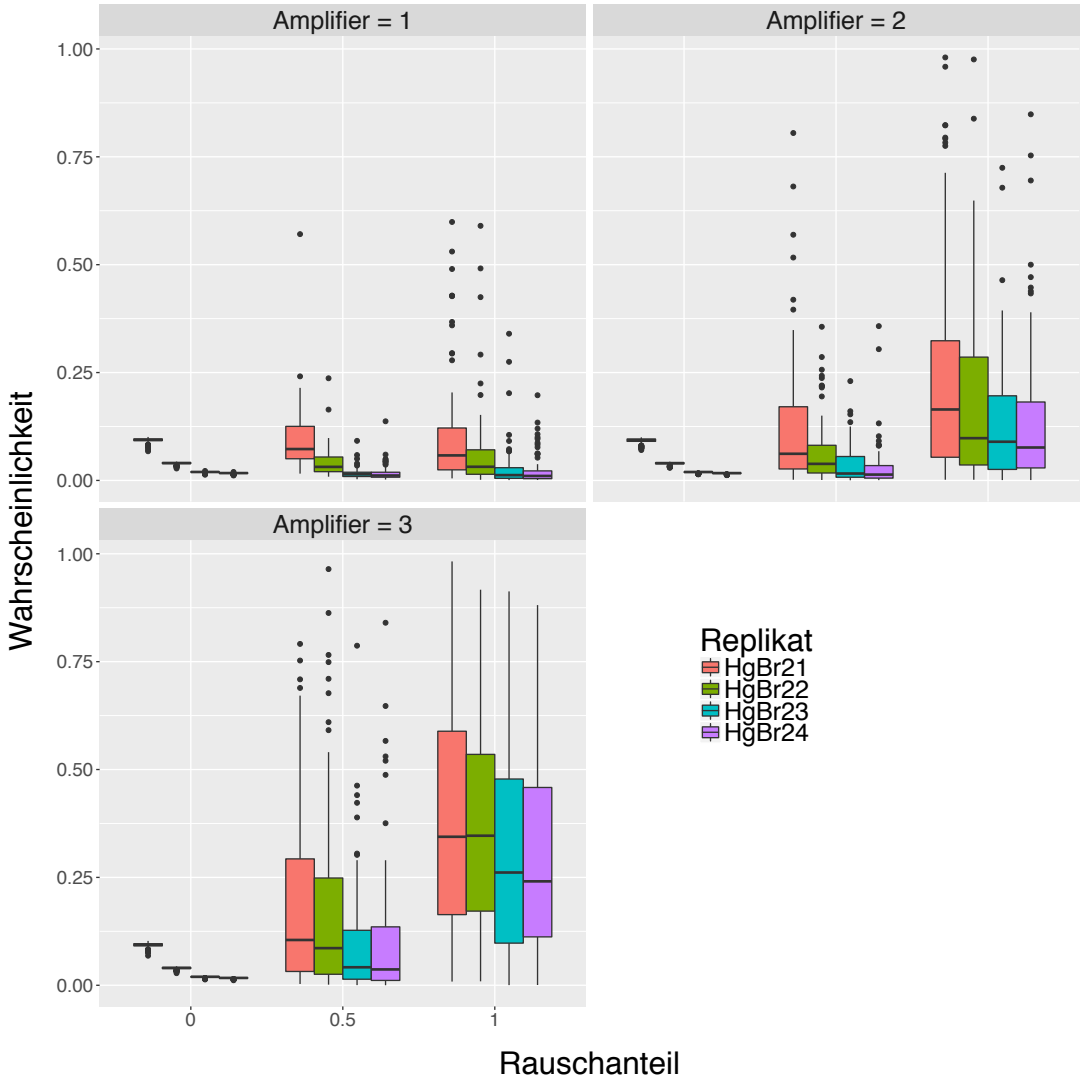

Abbildung 114: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgBr_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

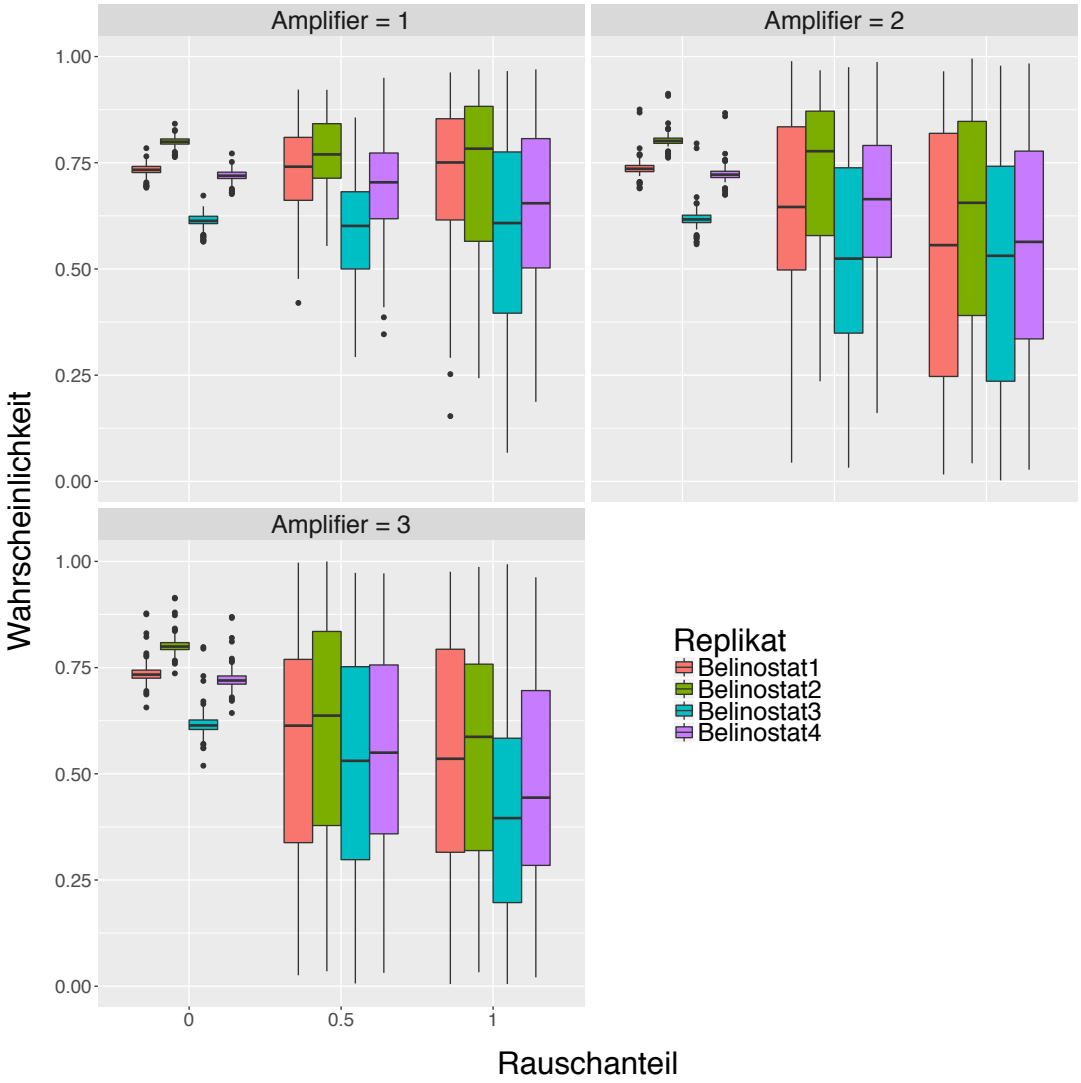

Abbildung 115: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Belinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

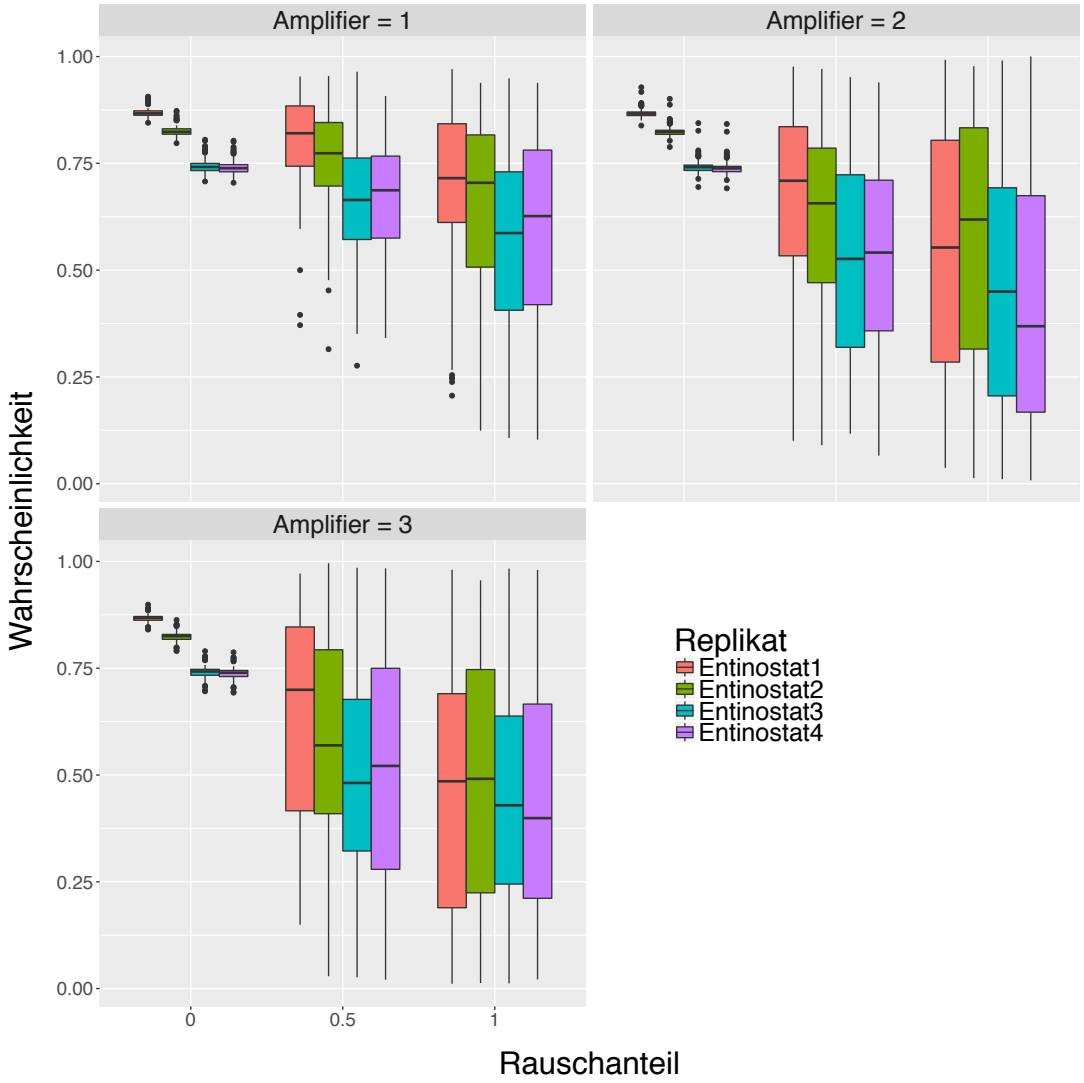

Abbildung 116: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

6.3.6 Rauschplots für UKN1 RF

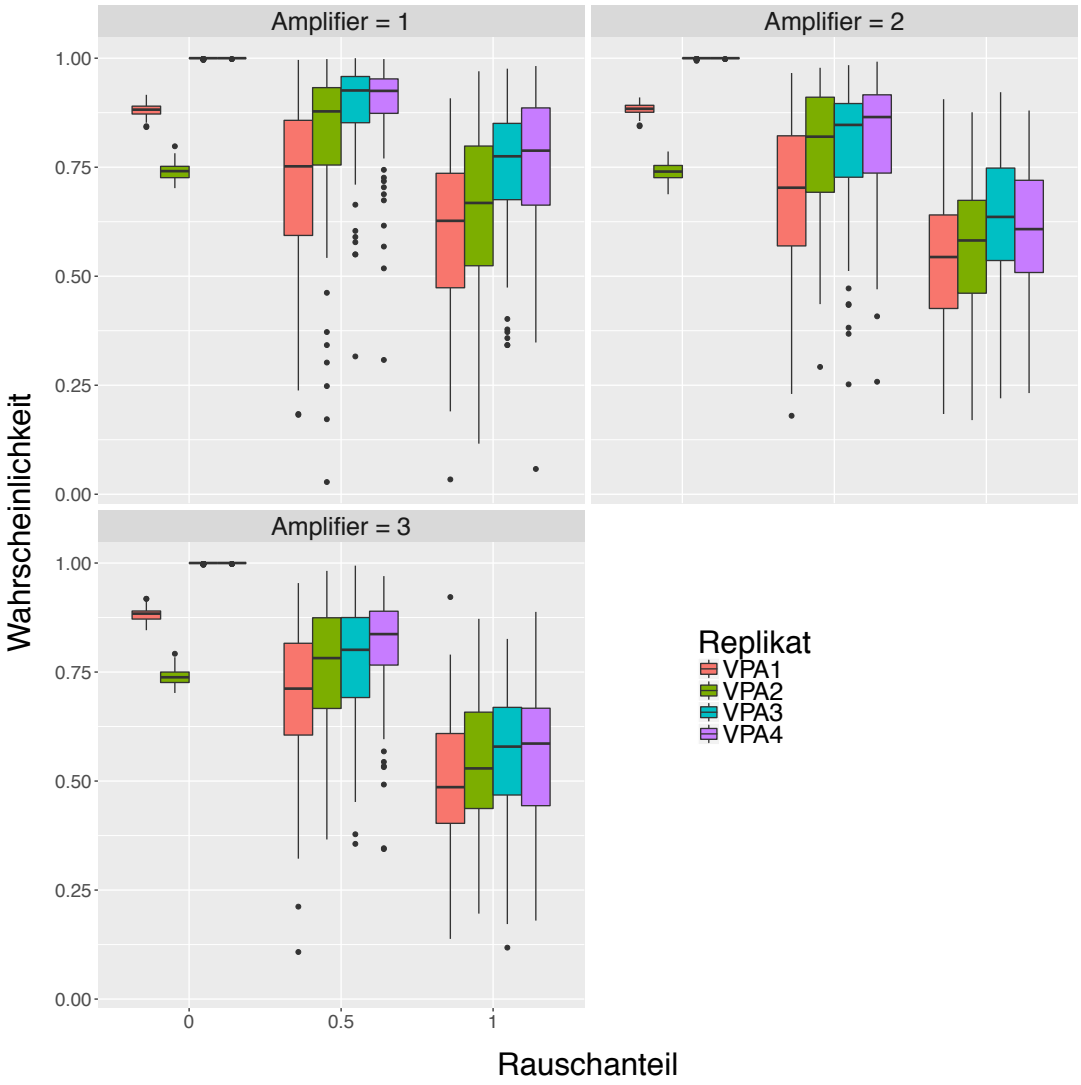

Abbildung 117: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

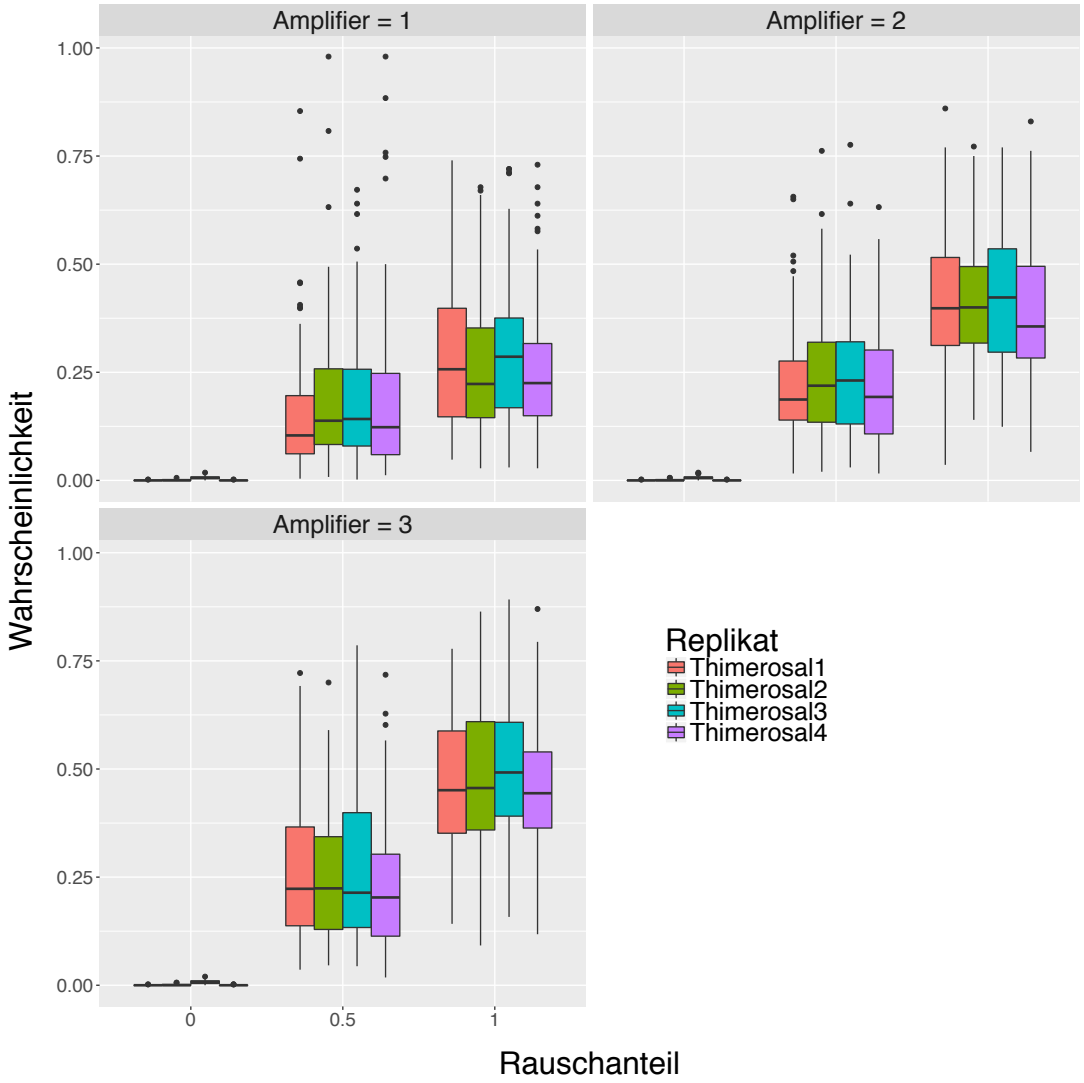

Abbildung 118: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

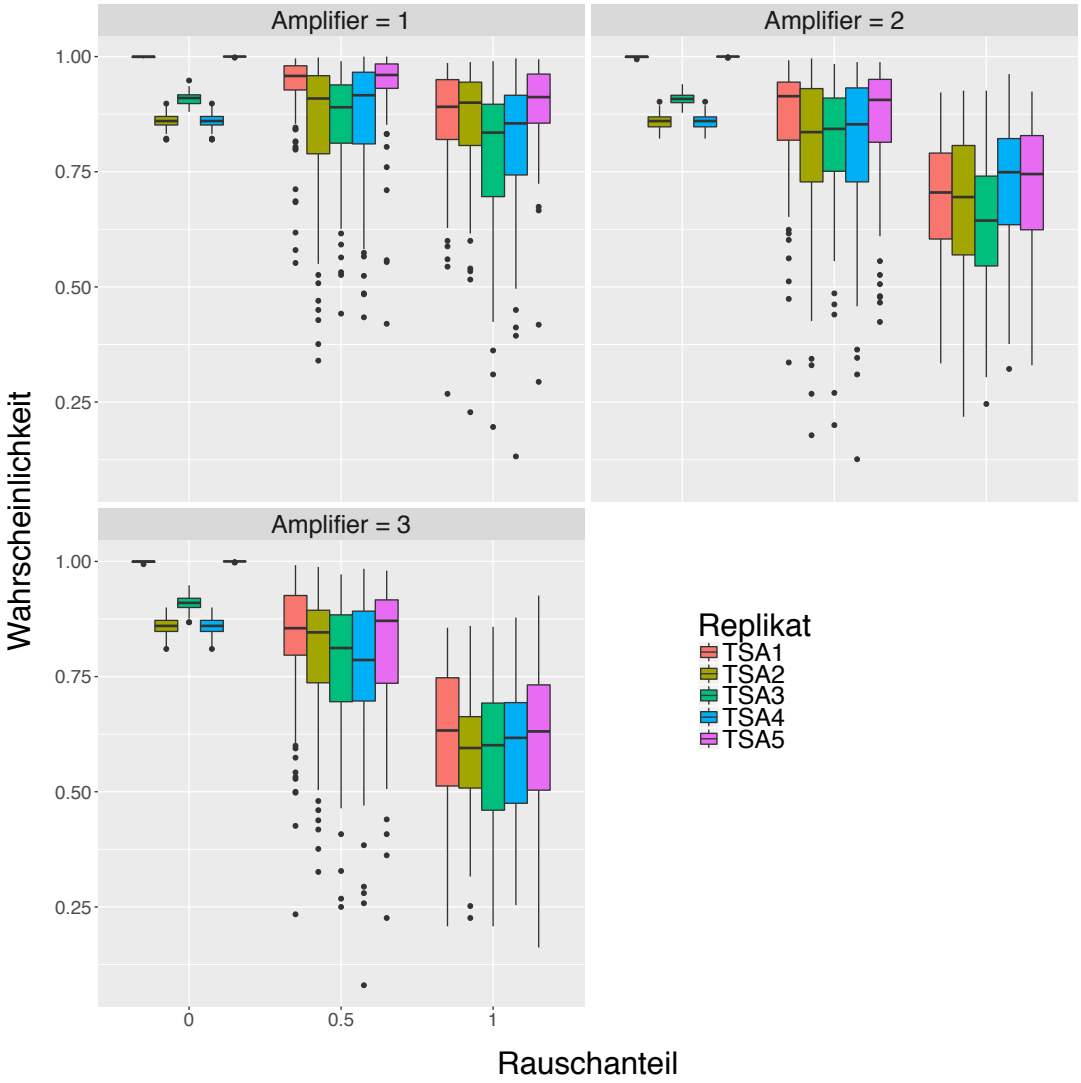

Abbildung 119: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

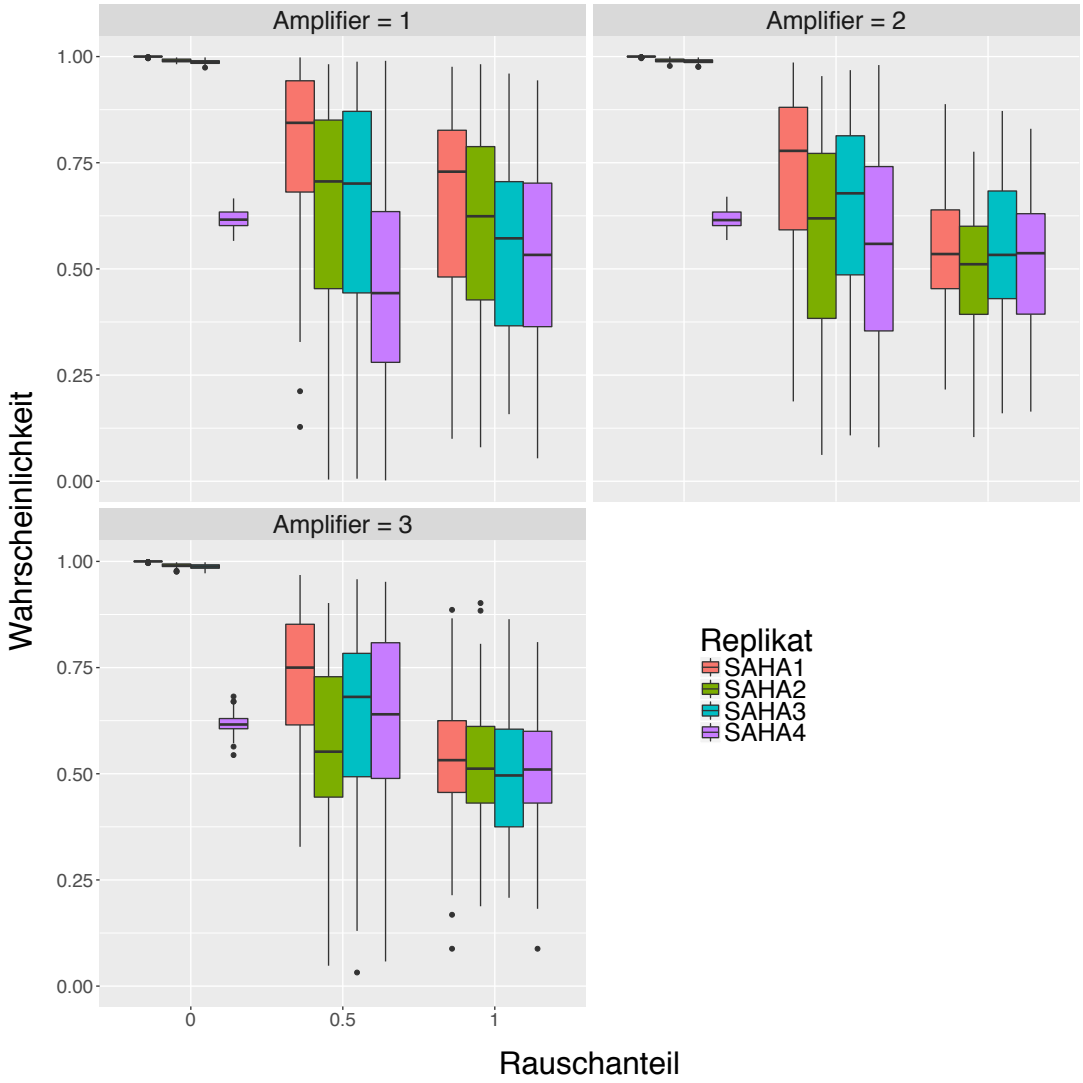

Abbildung 120: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

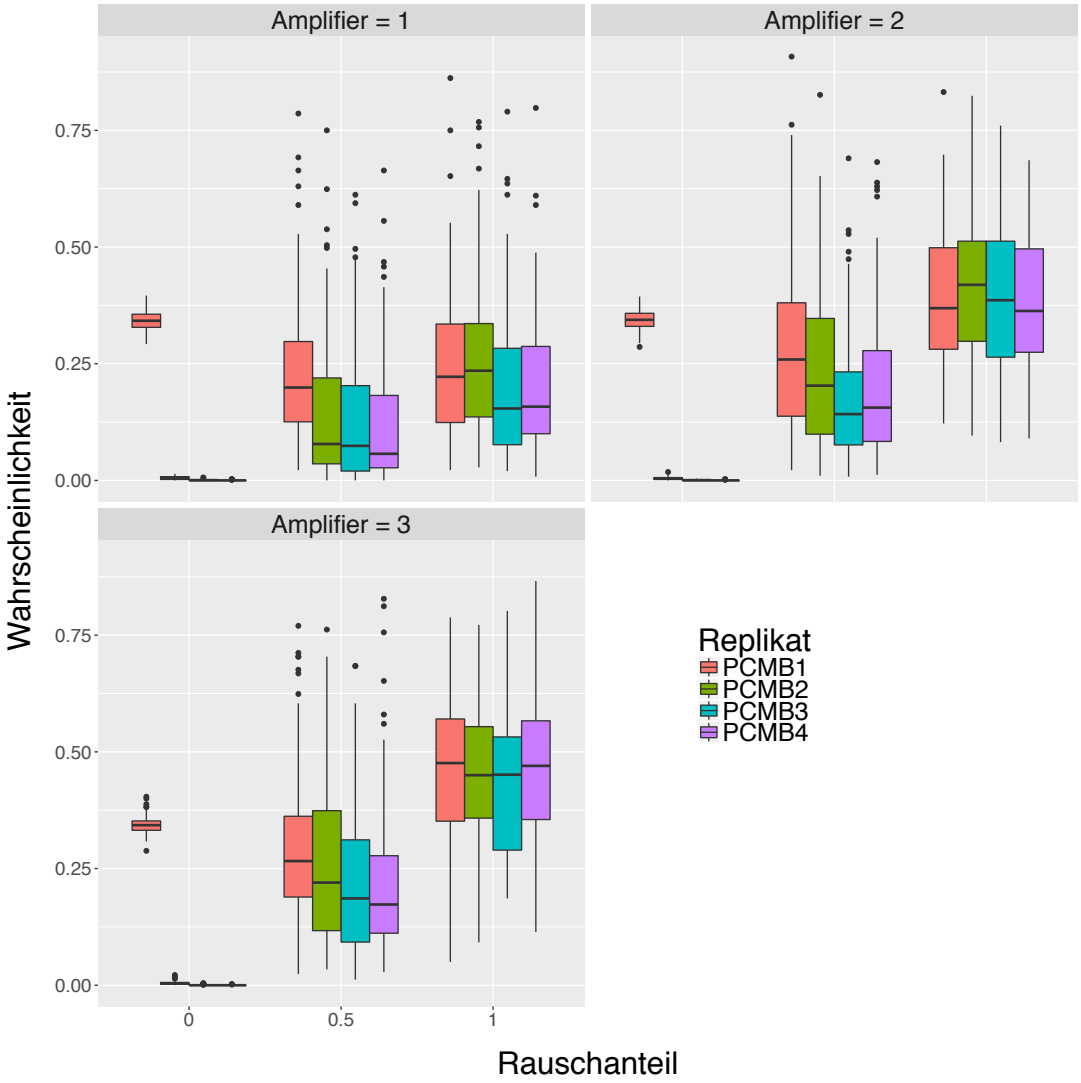

Abbildung 121: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

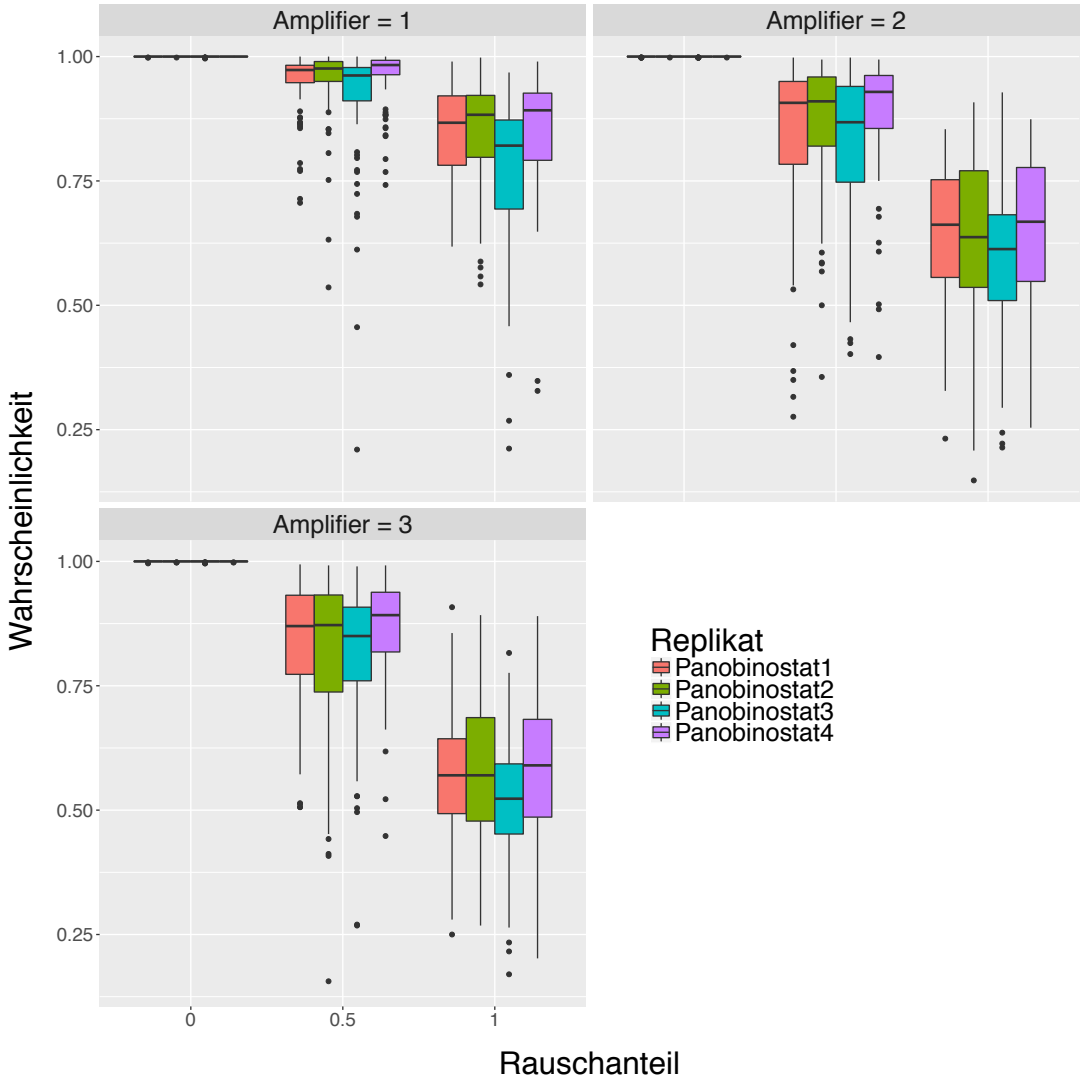

Abbildung 122: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

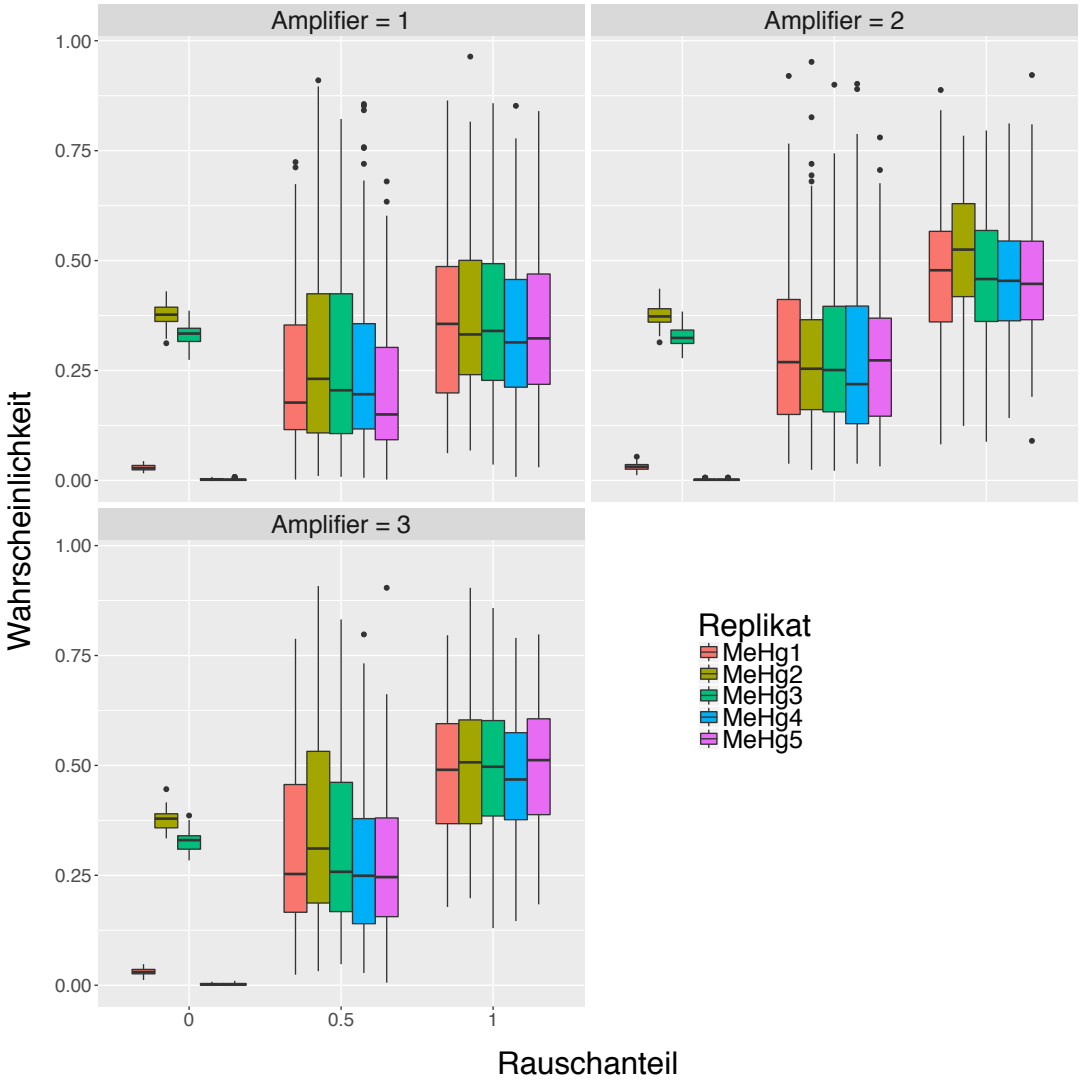

Abbildung 123: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

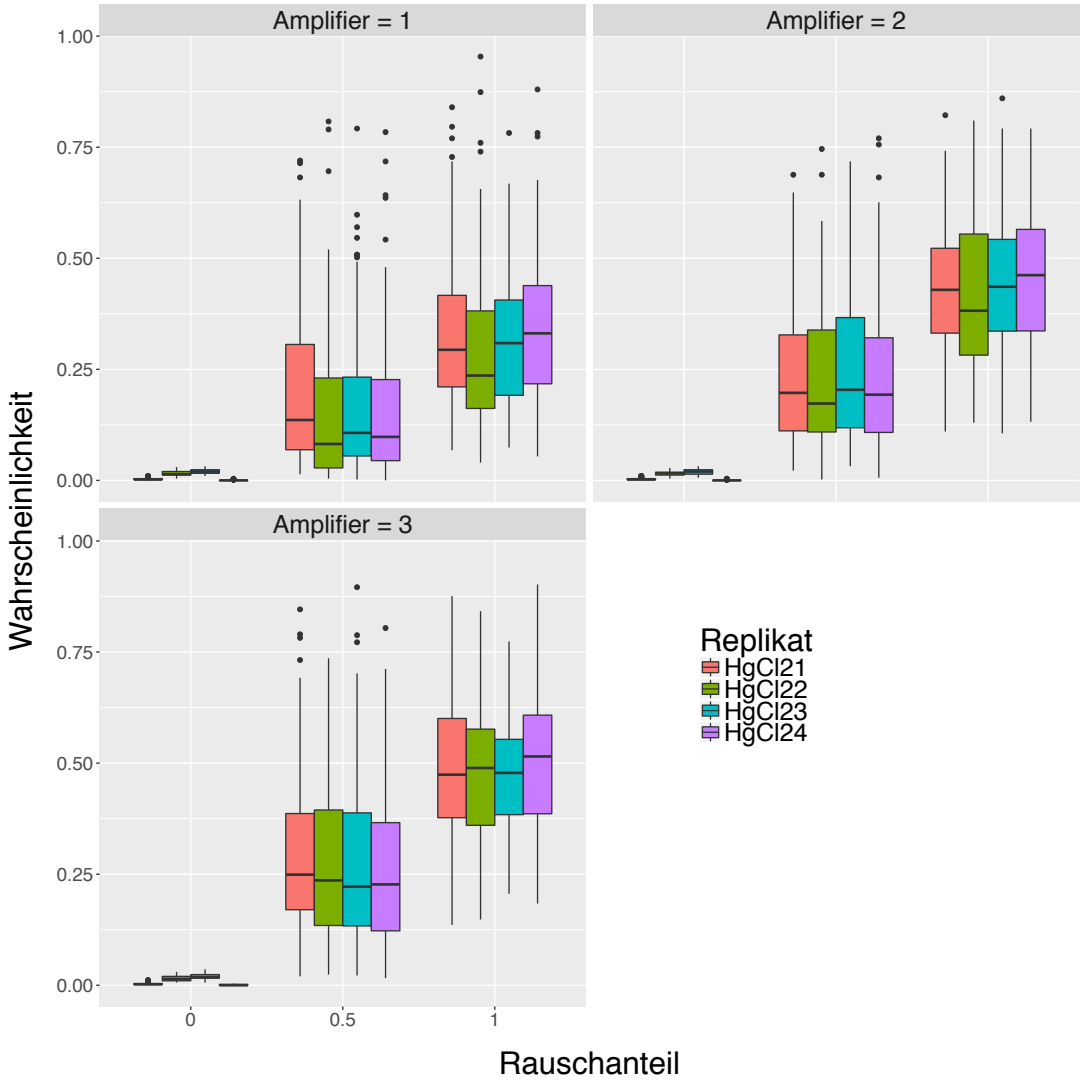

Abbildung 124: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgCl_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

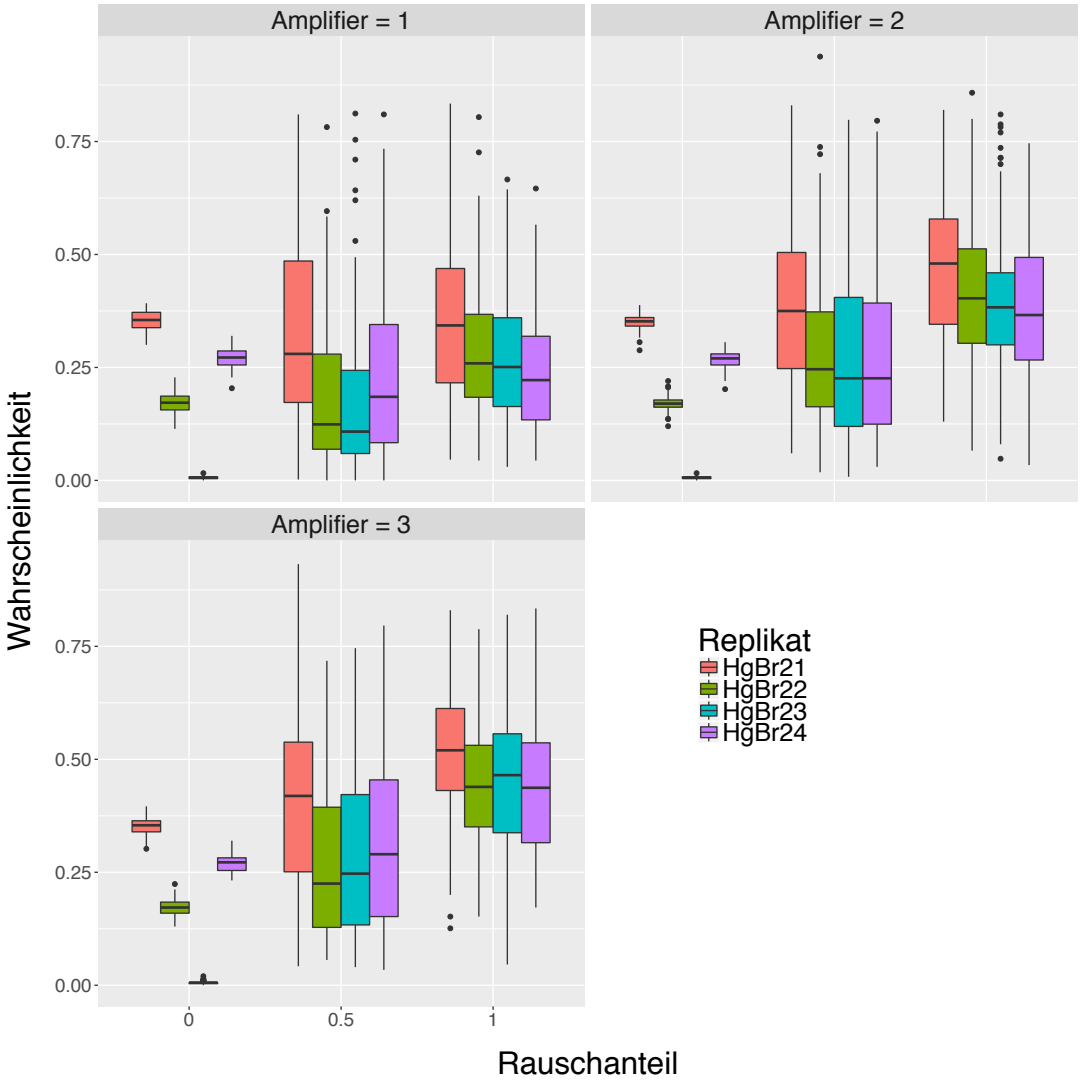

Abbildung 125: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgBr_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

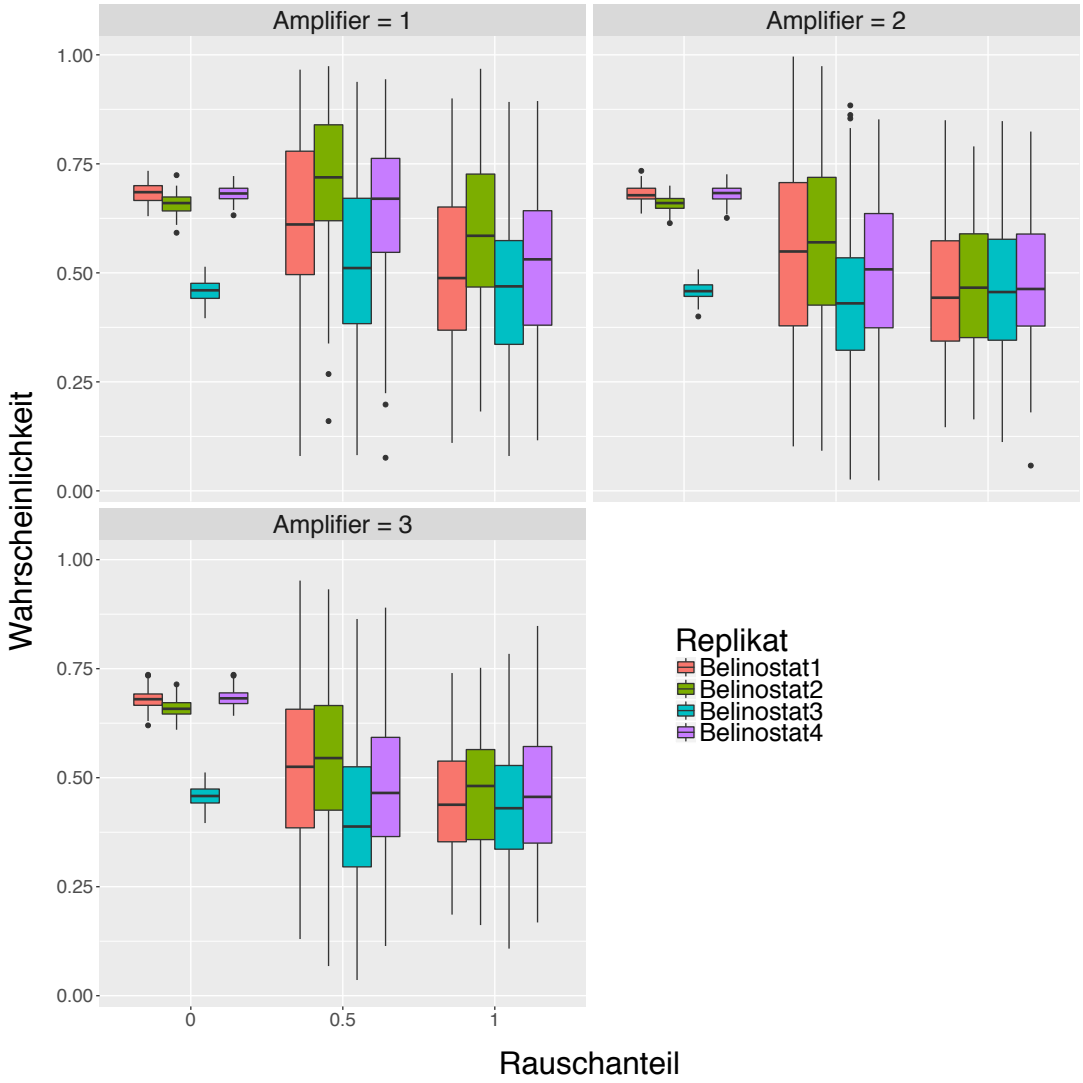

Abbildung 126: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Belinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

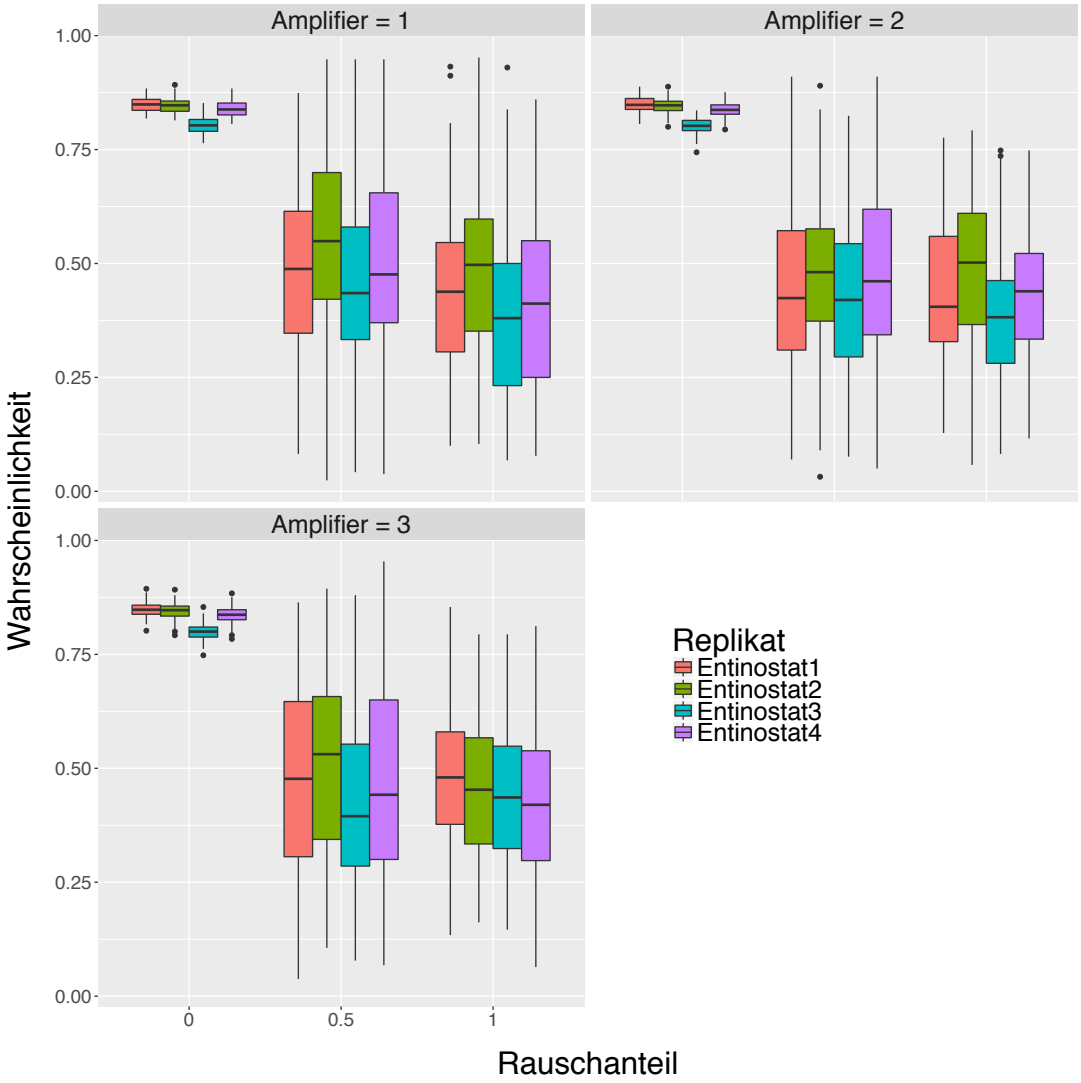

Abbildung 127: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

6.3.7 Rauschplots für UKK RF

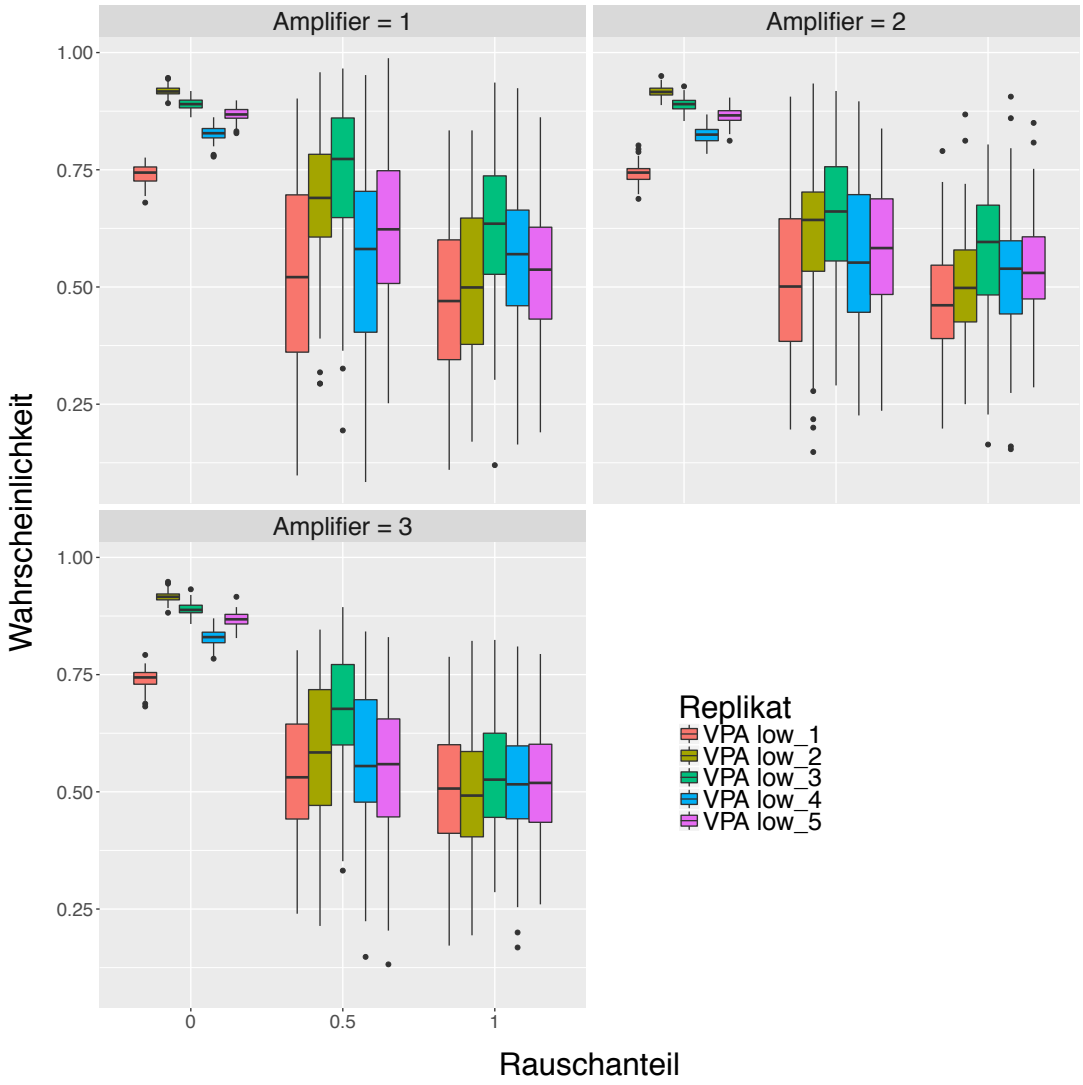

Abbildung 128: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA (kleinere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

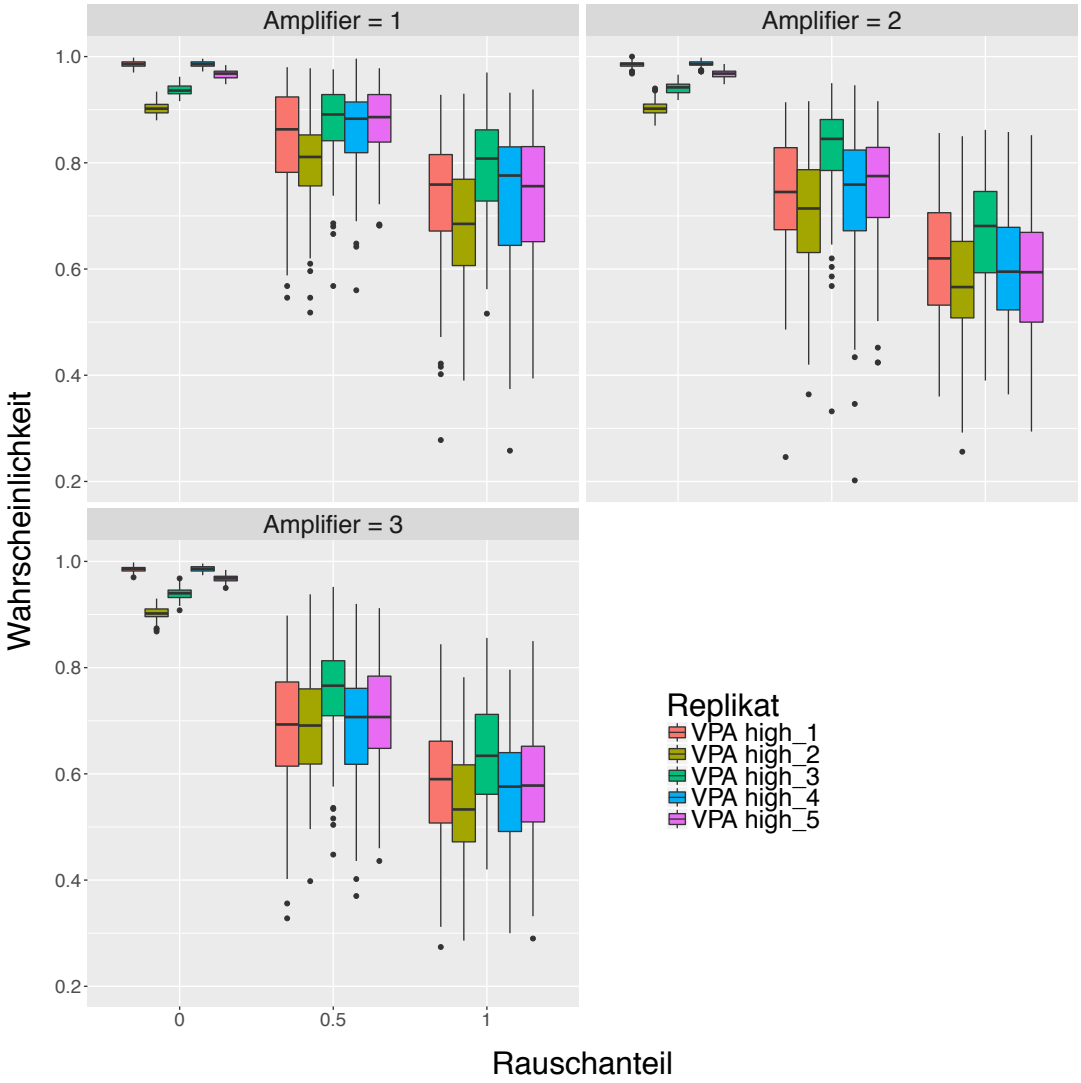

Abbildung 129: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA (höhere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

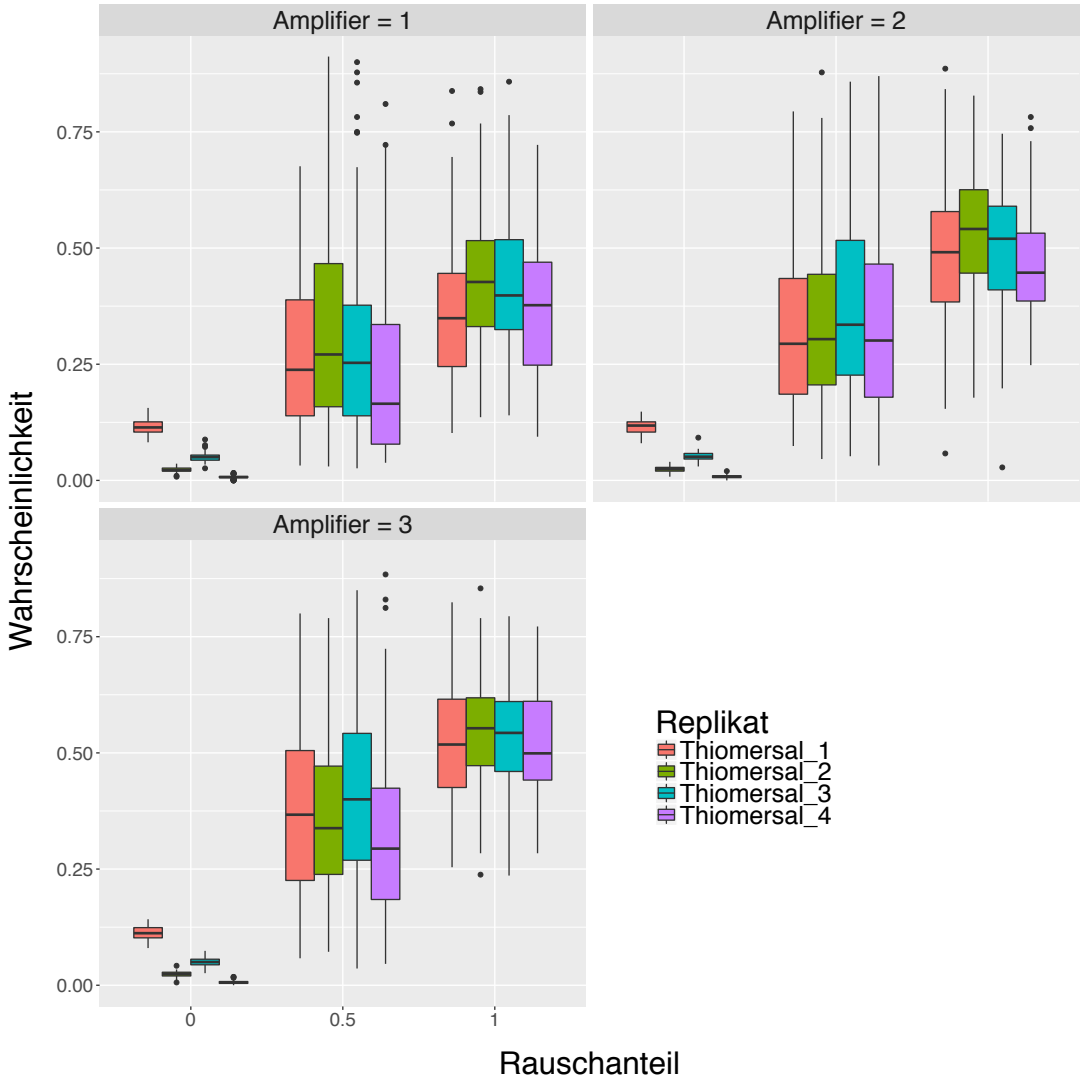

Abbildung 130: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

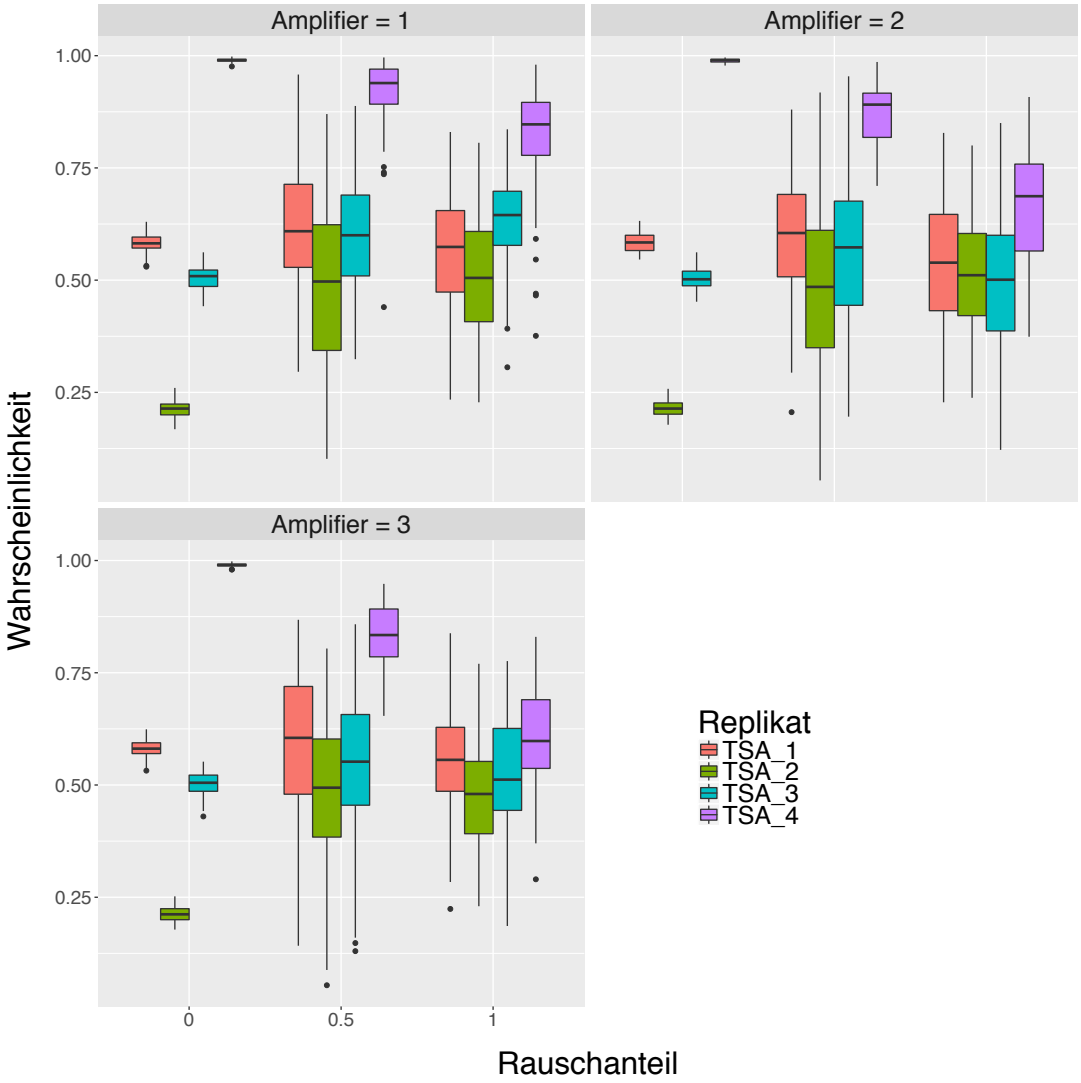

Abbildung 131: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

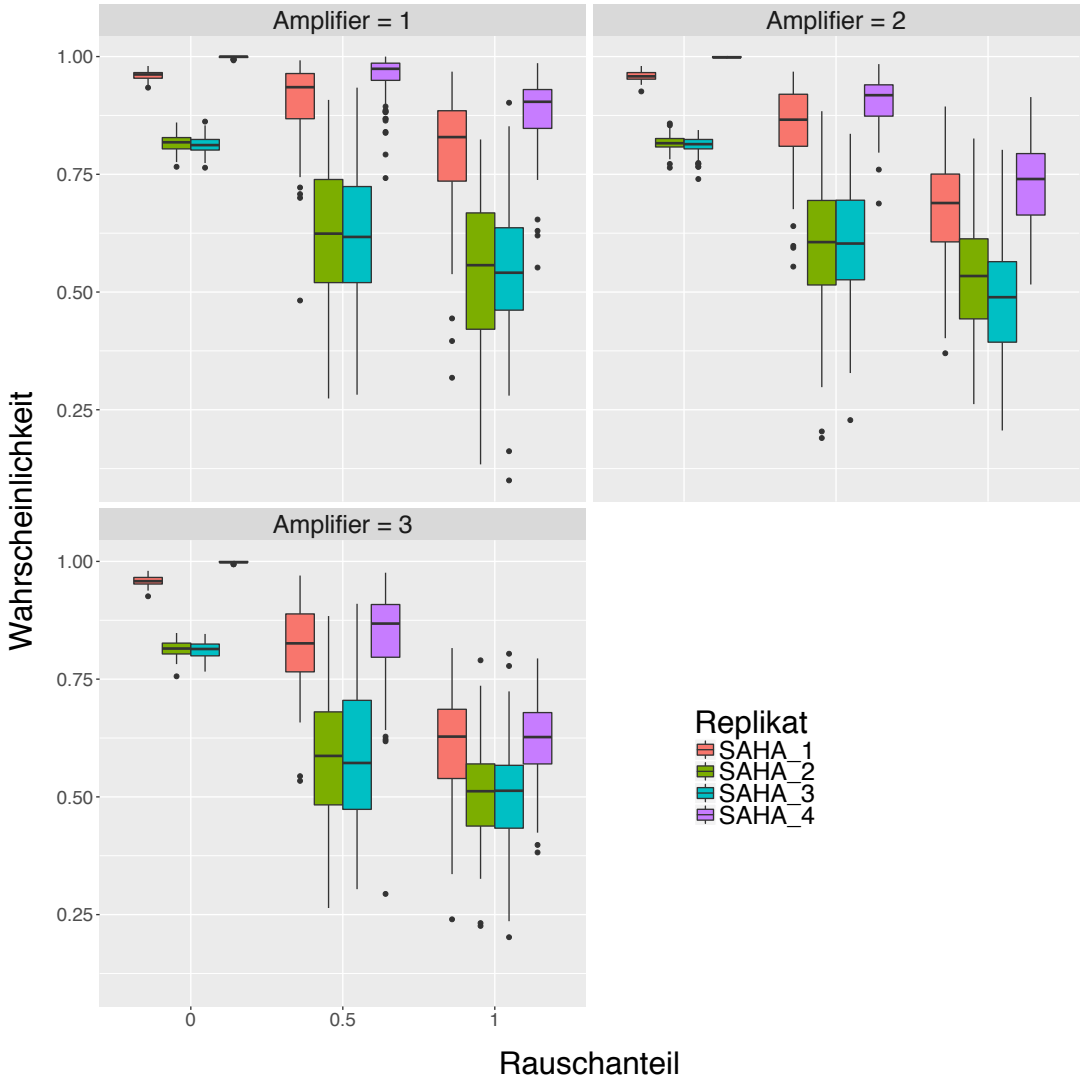

Abbildung 132: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

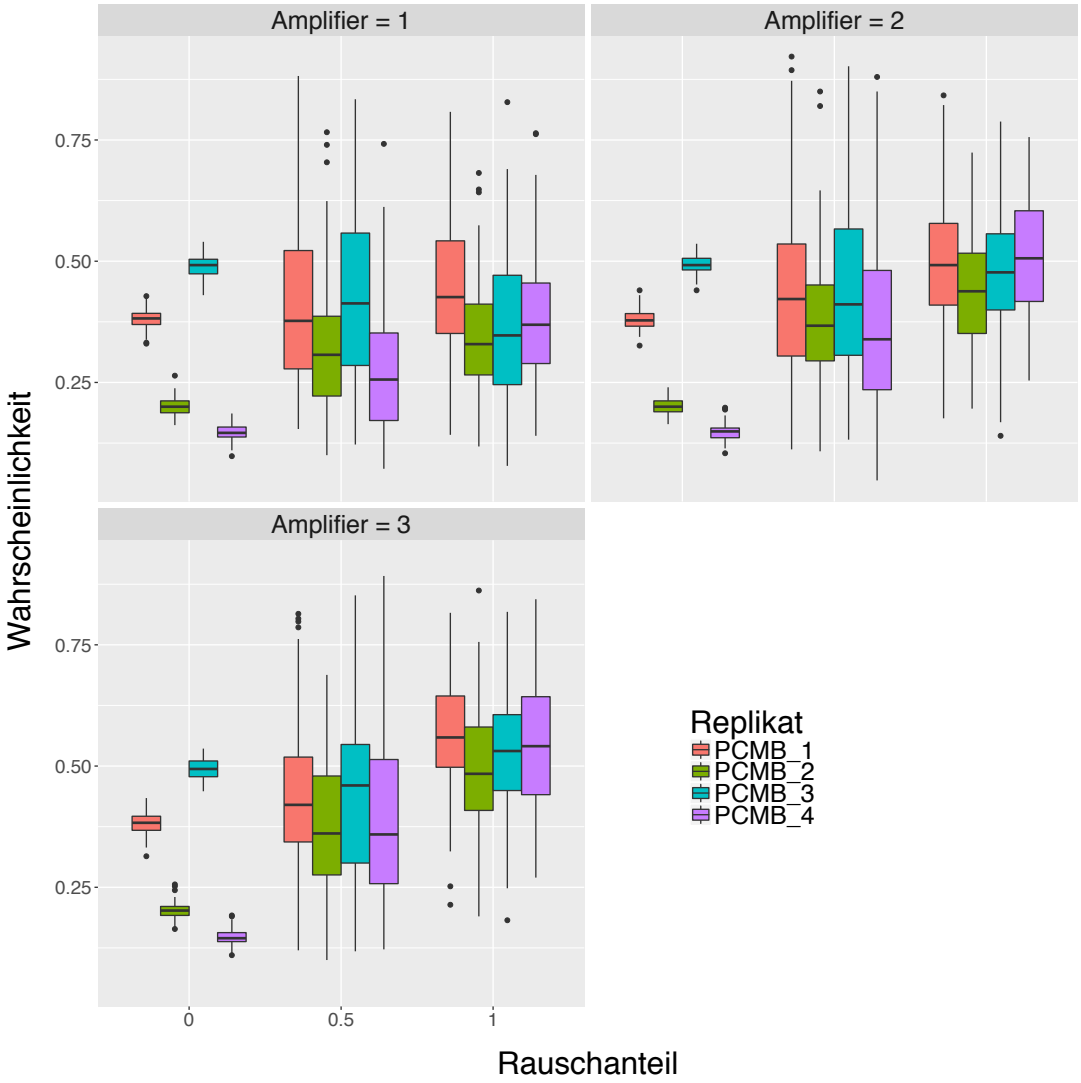

Abbildung 133: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

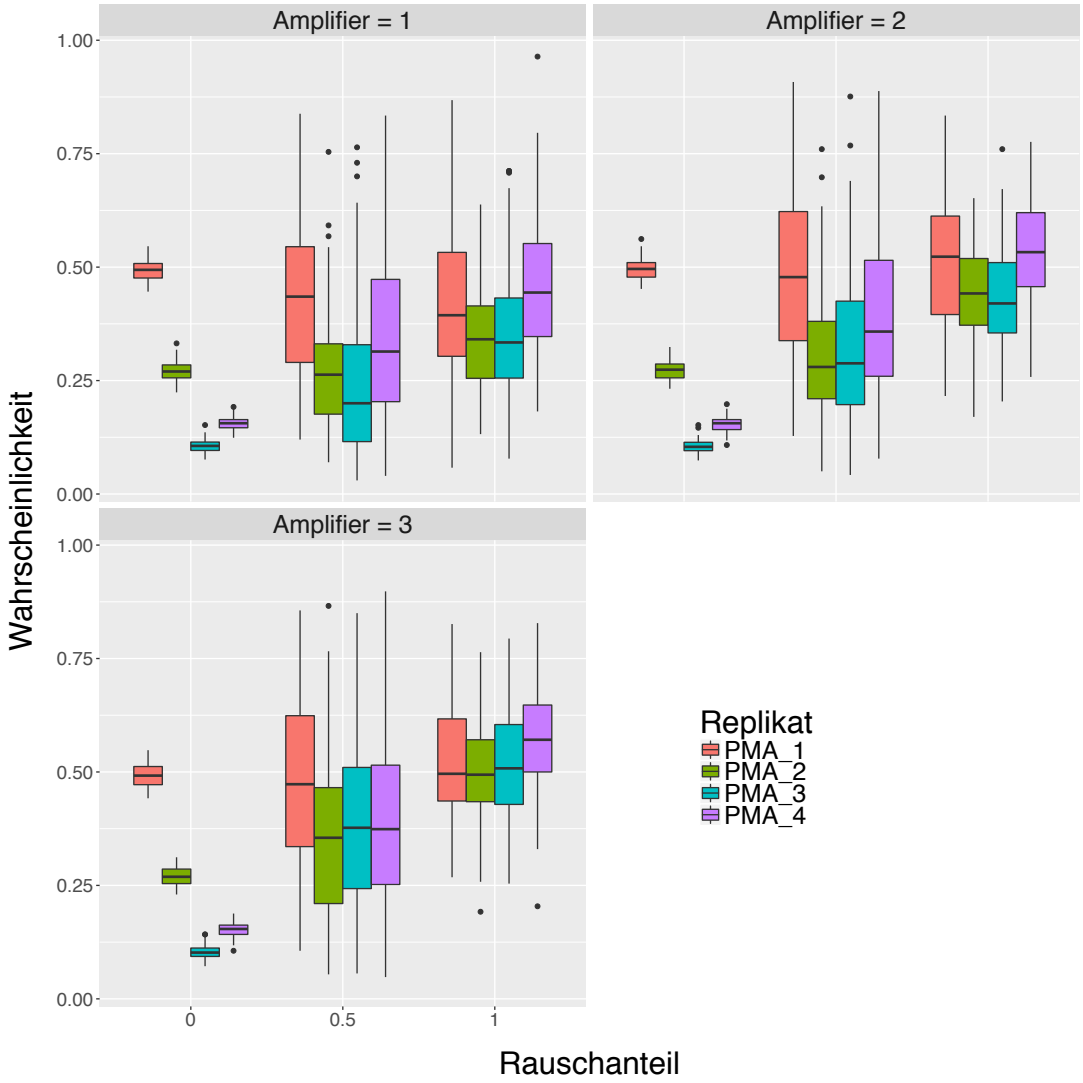

Abbildung 134: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

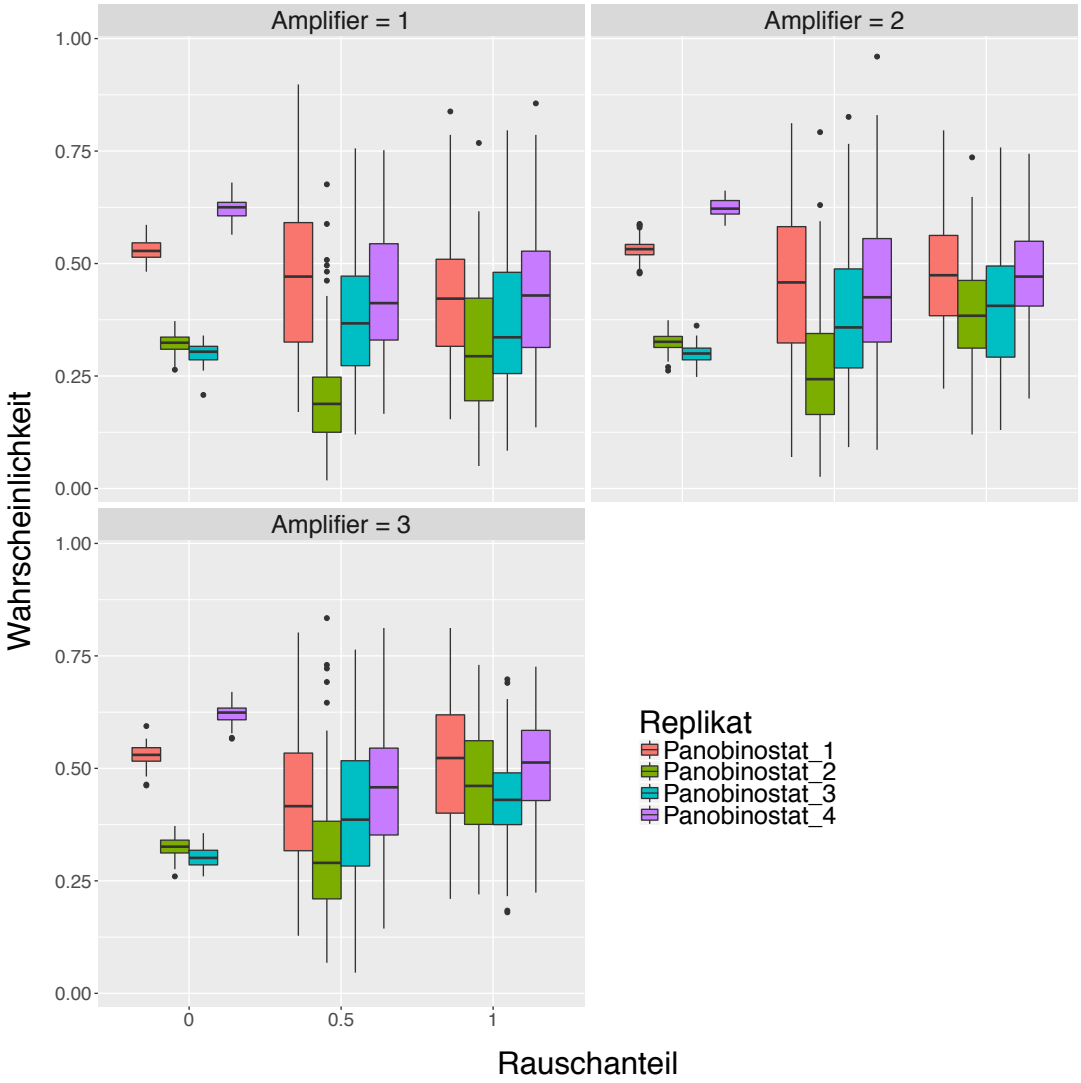

Abbildung 135: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

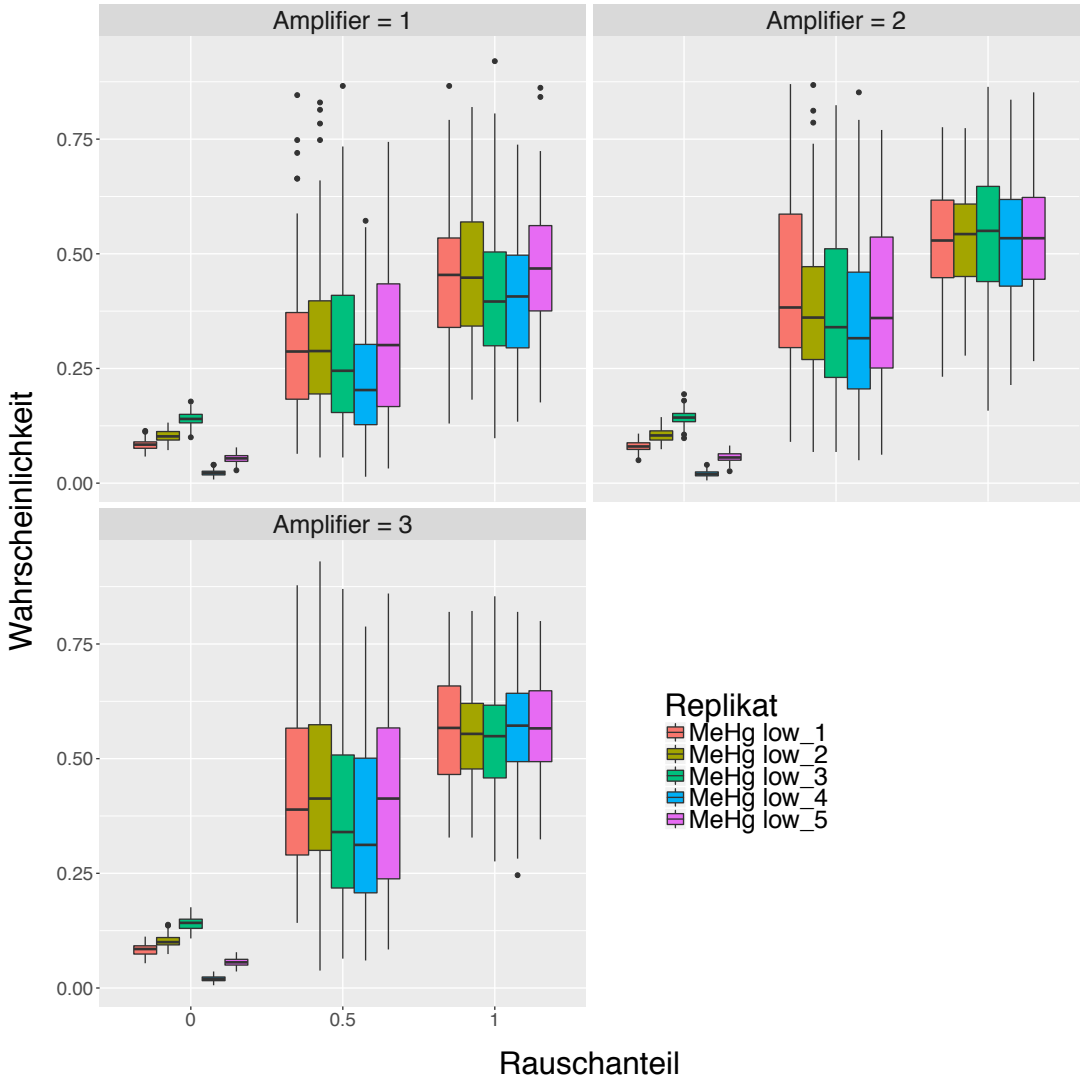

Abbildung 136: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg (kleinere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

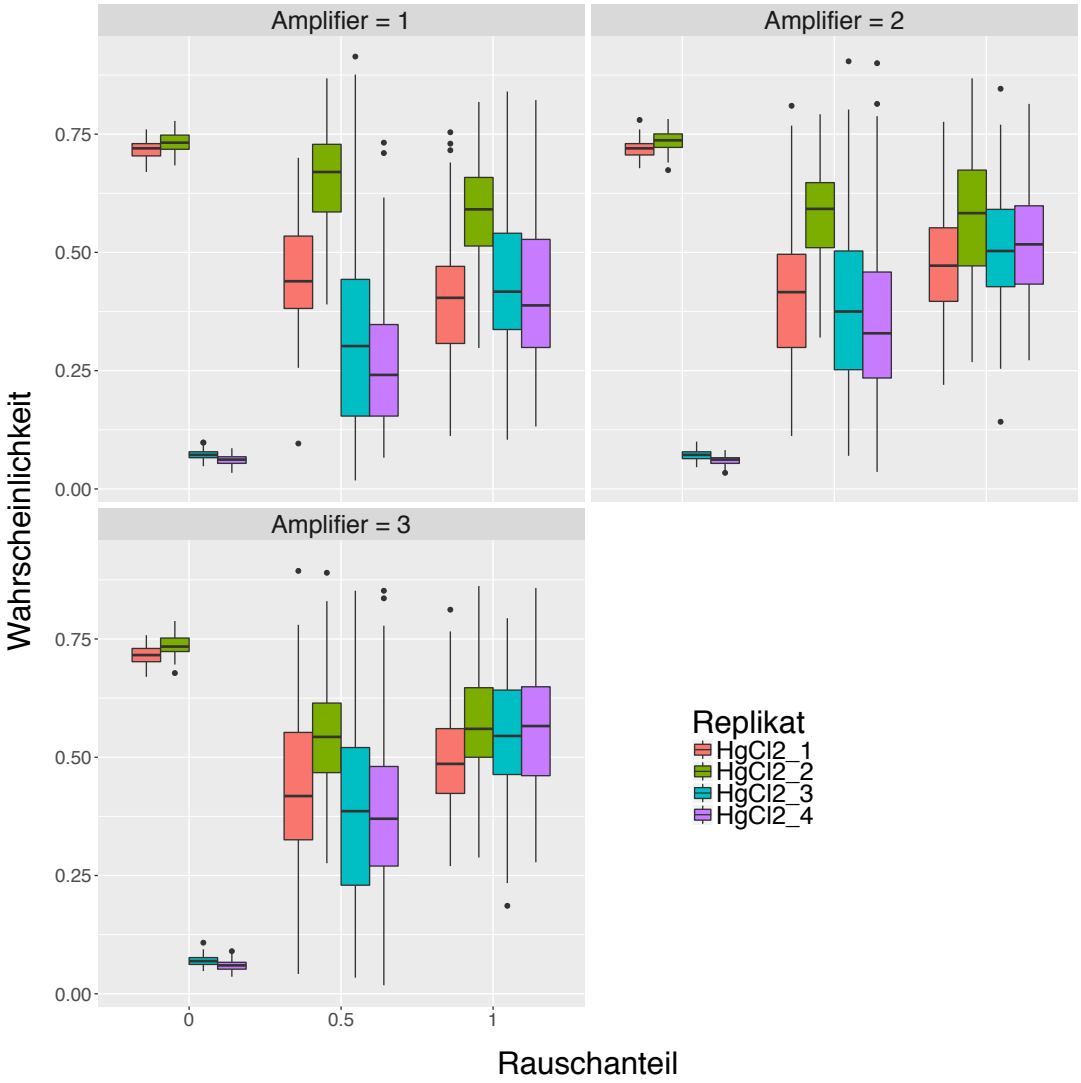

Abbildung 137: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgCl_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

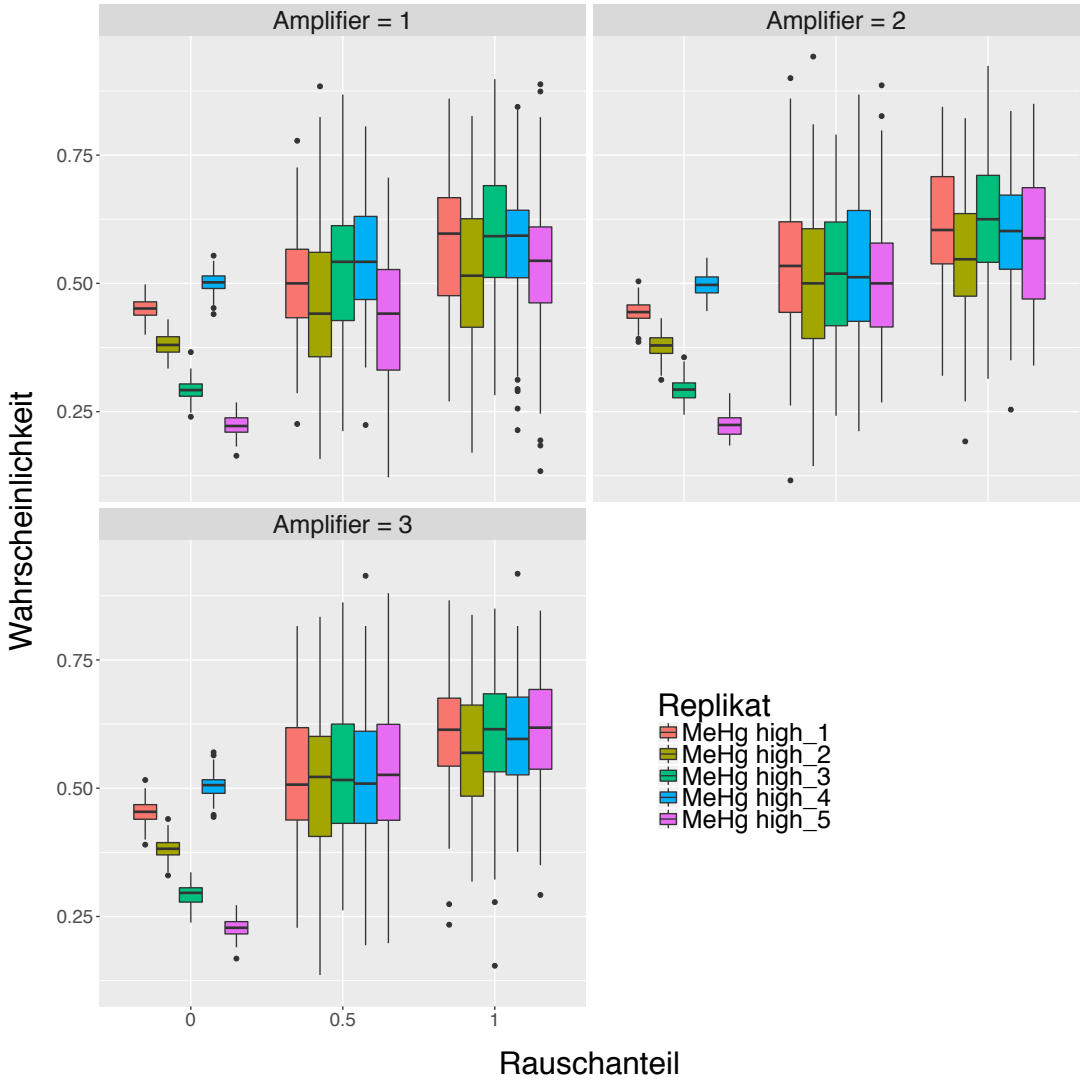

Abbildung 138: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg (höhere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

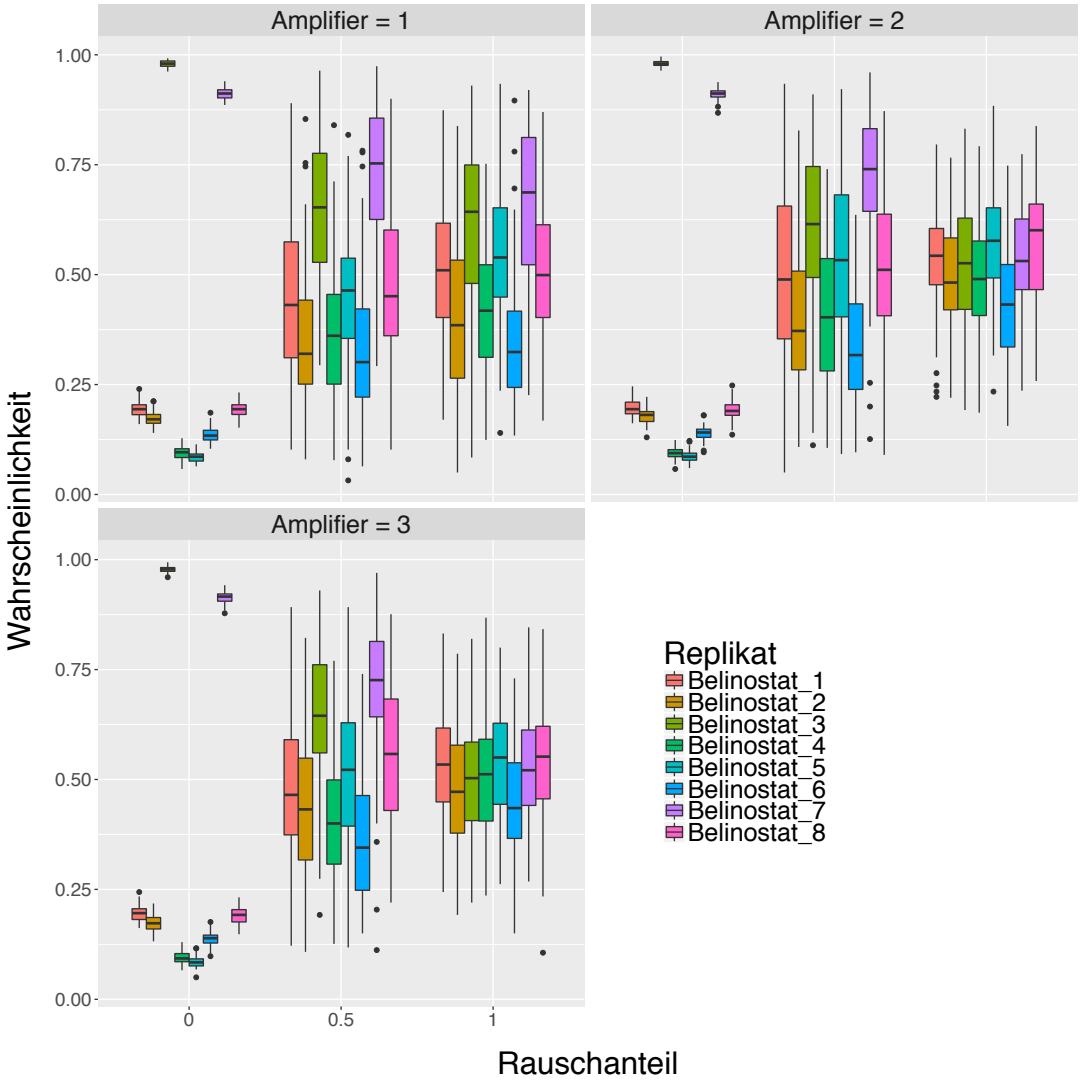

Abbildung 139: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Belinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

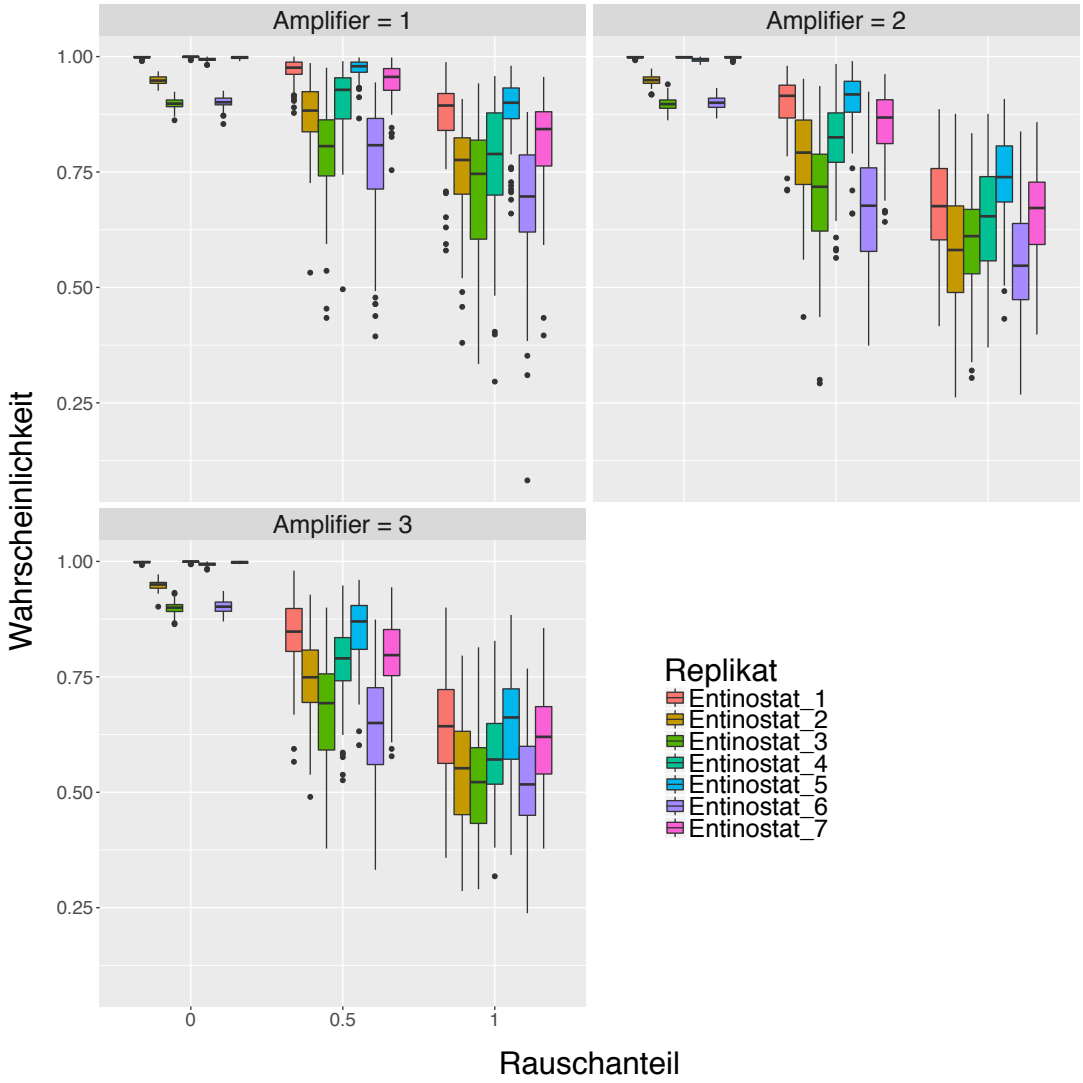

Abbildung 140: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Random Forest verwendet.

6.3.8 Rauschplots für UKK SVM

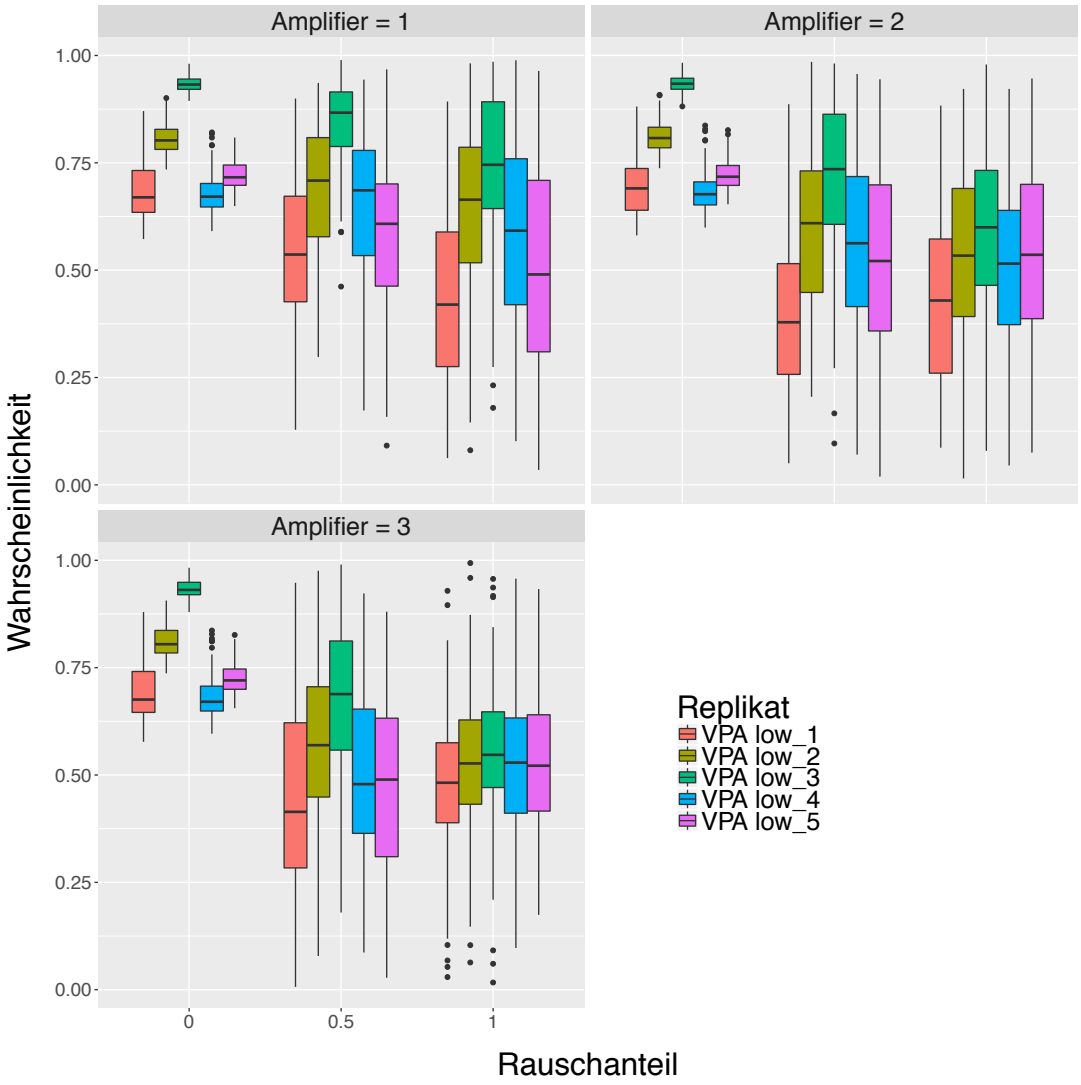

Abbildung 141: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA (kleinere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

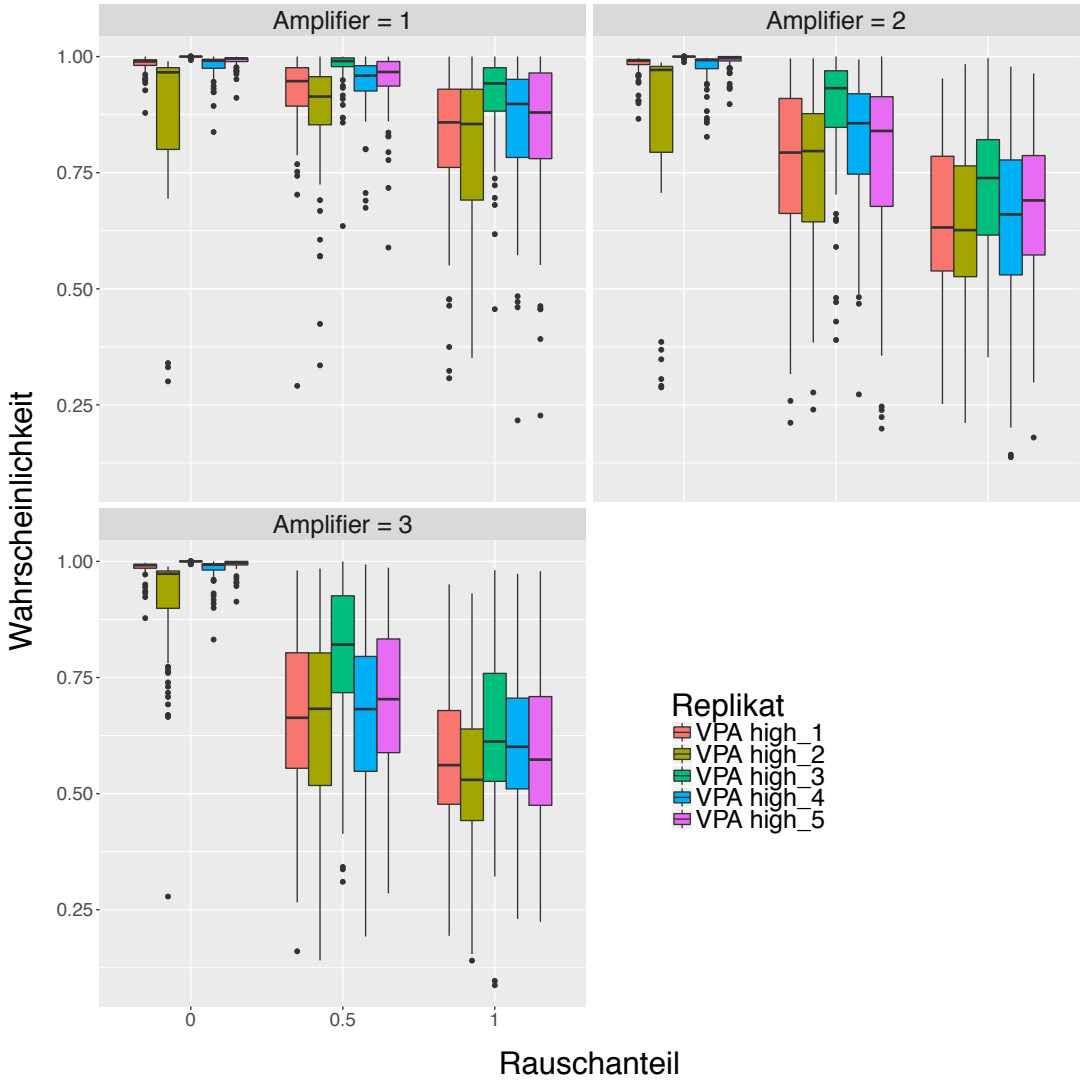

Abbildung 142: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von VPA (höhere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

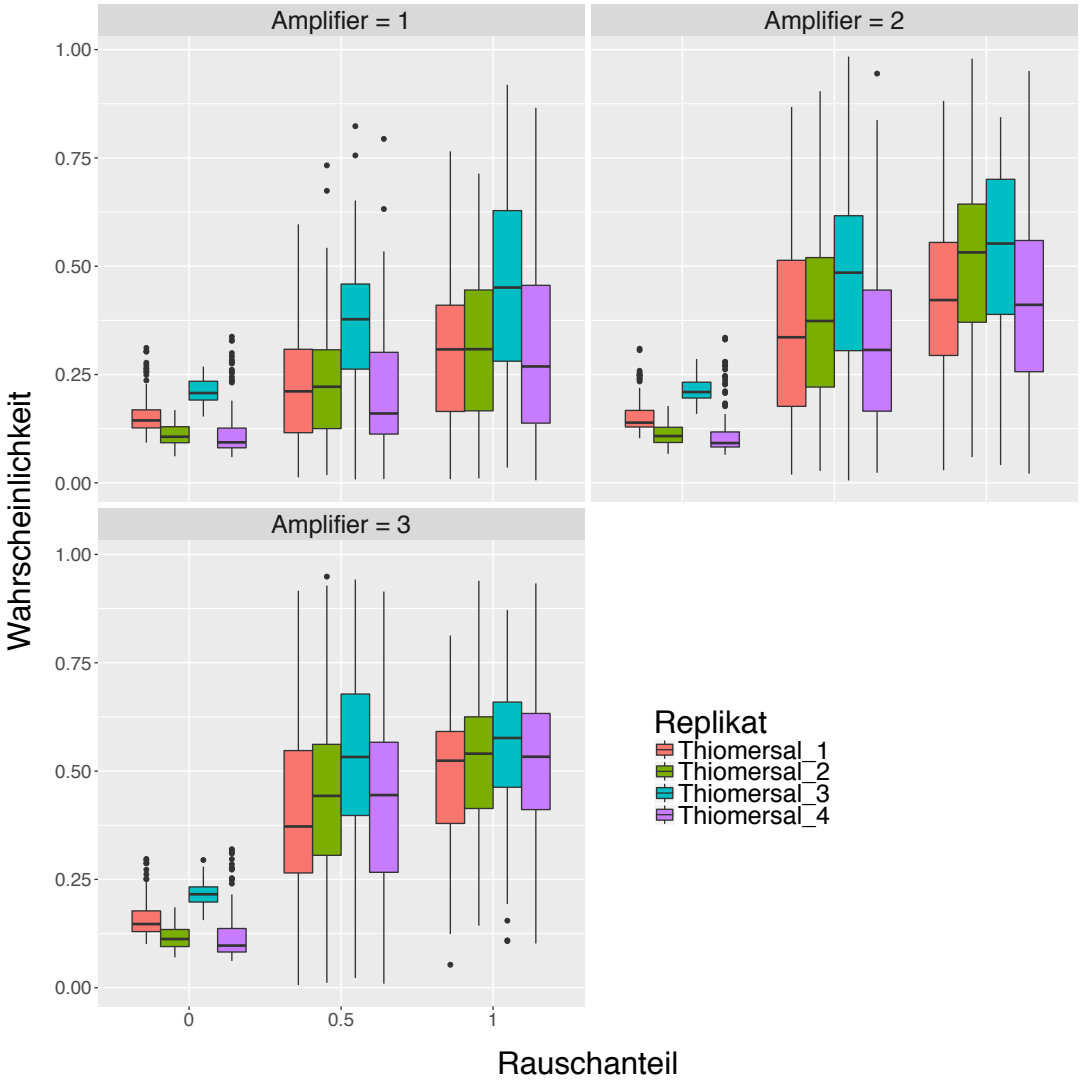

Abbildung 143: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Thiomersal nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

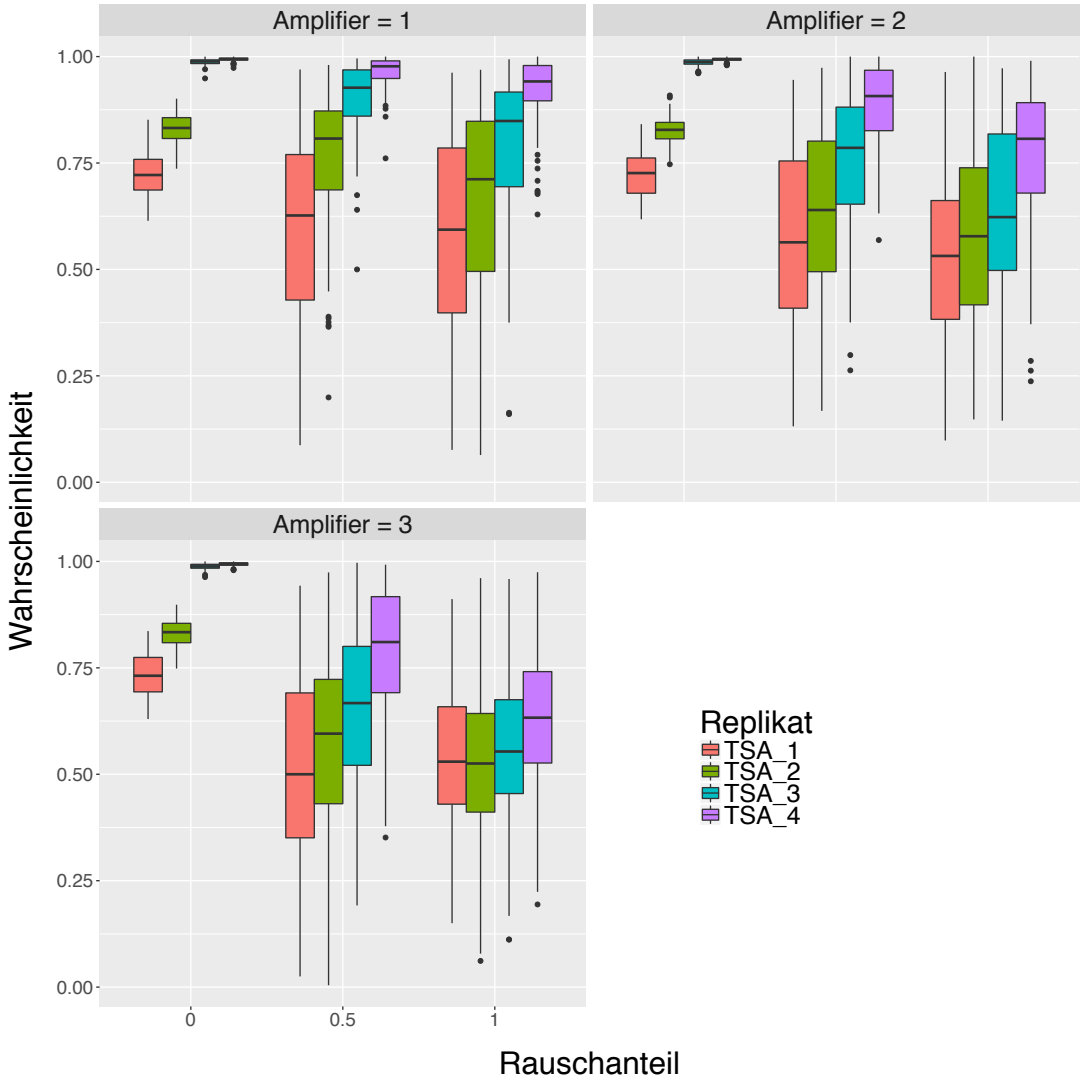

Abbildung 144: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von TSA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

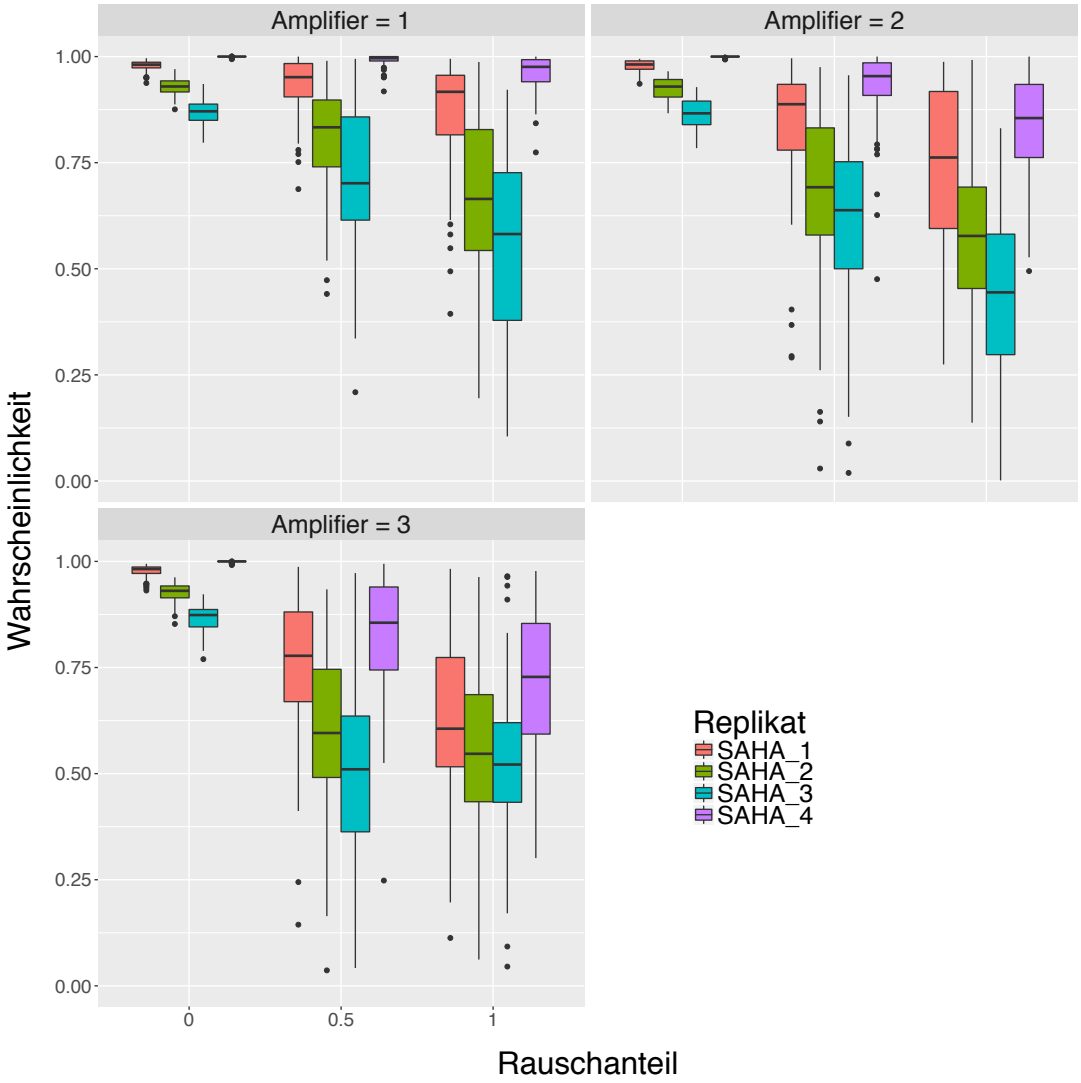

Abbildung 145: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von SAHA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

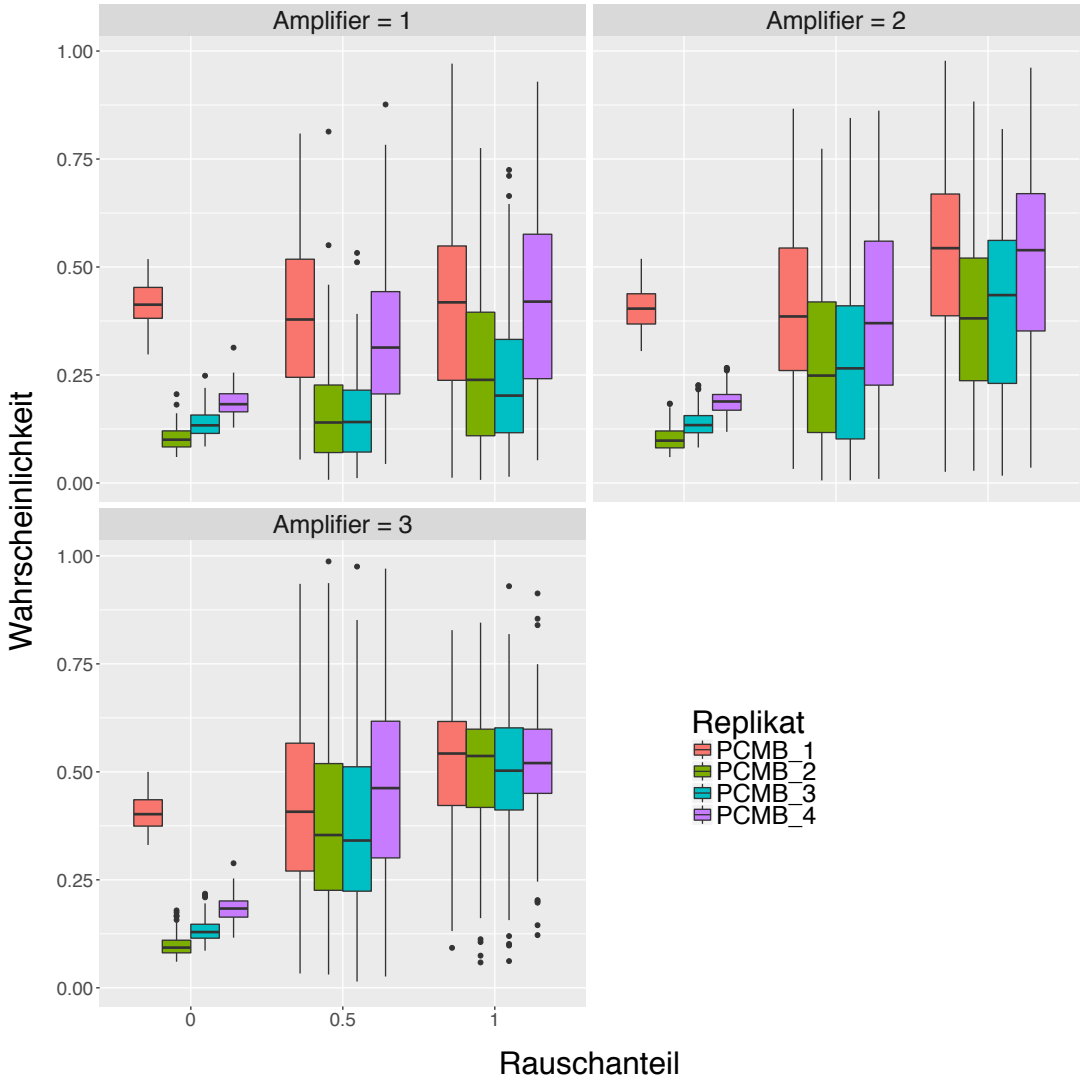

Abbildung 146: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PCMB nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

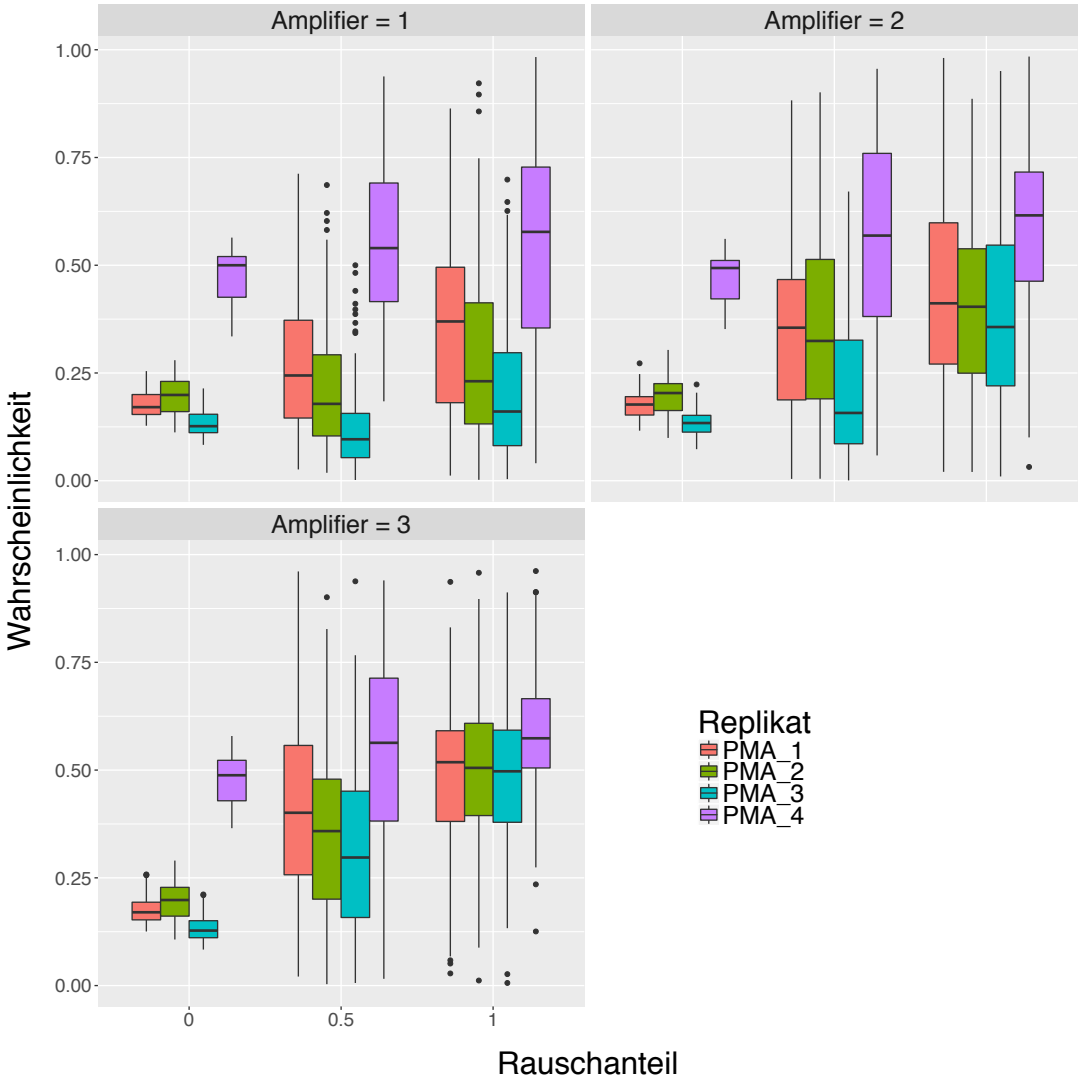

Abbildung 147: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von PMA nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

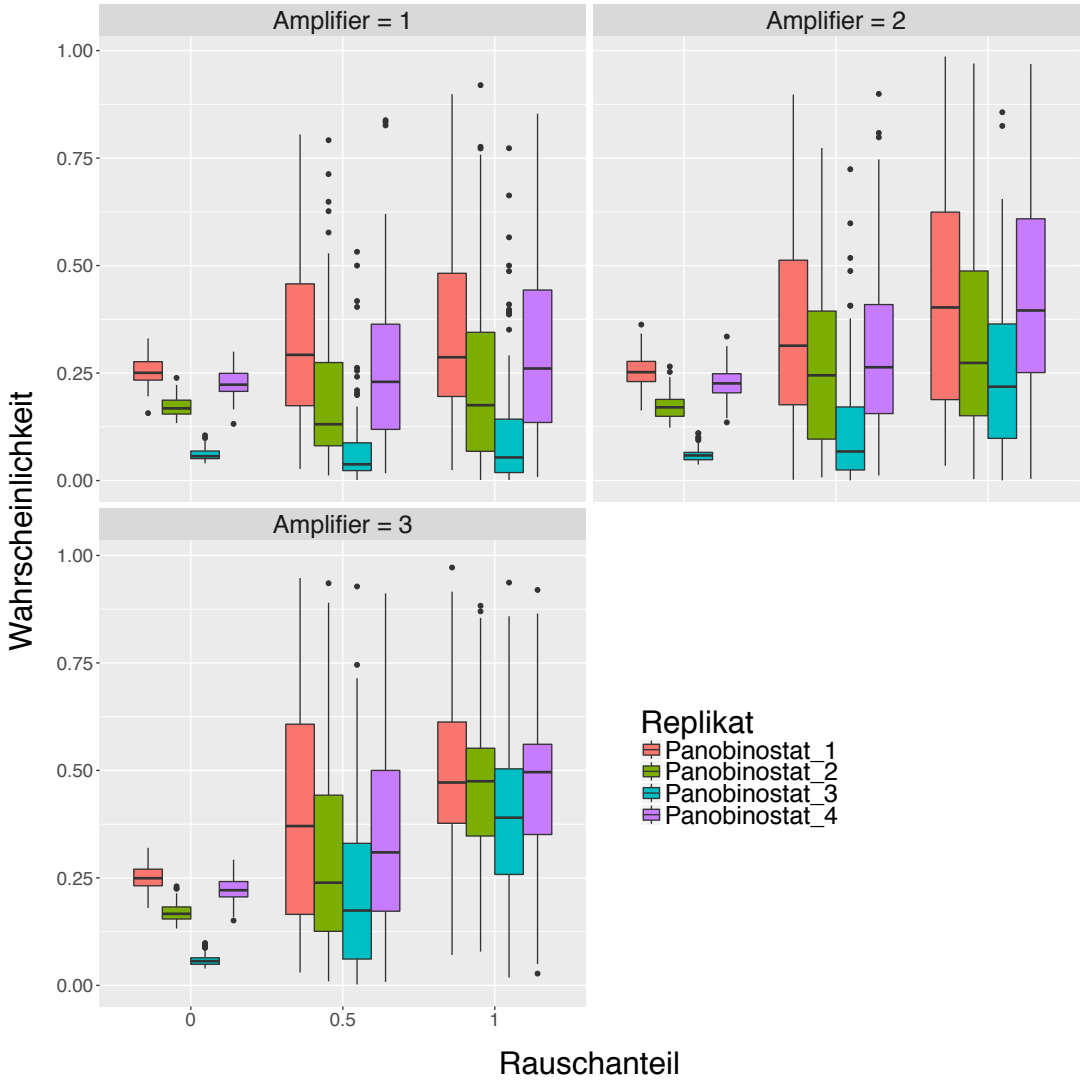

Abbildung 148: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Panobinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

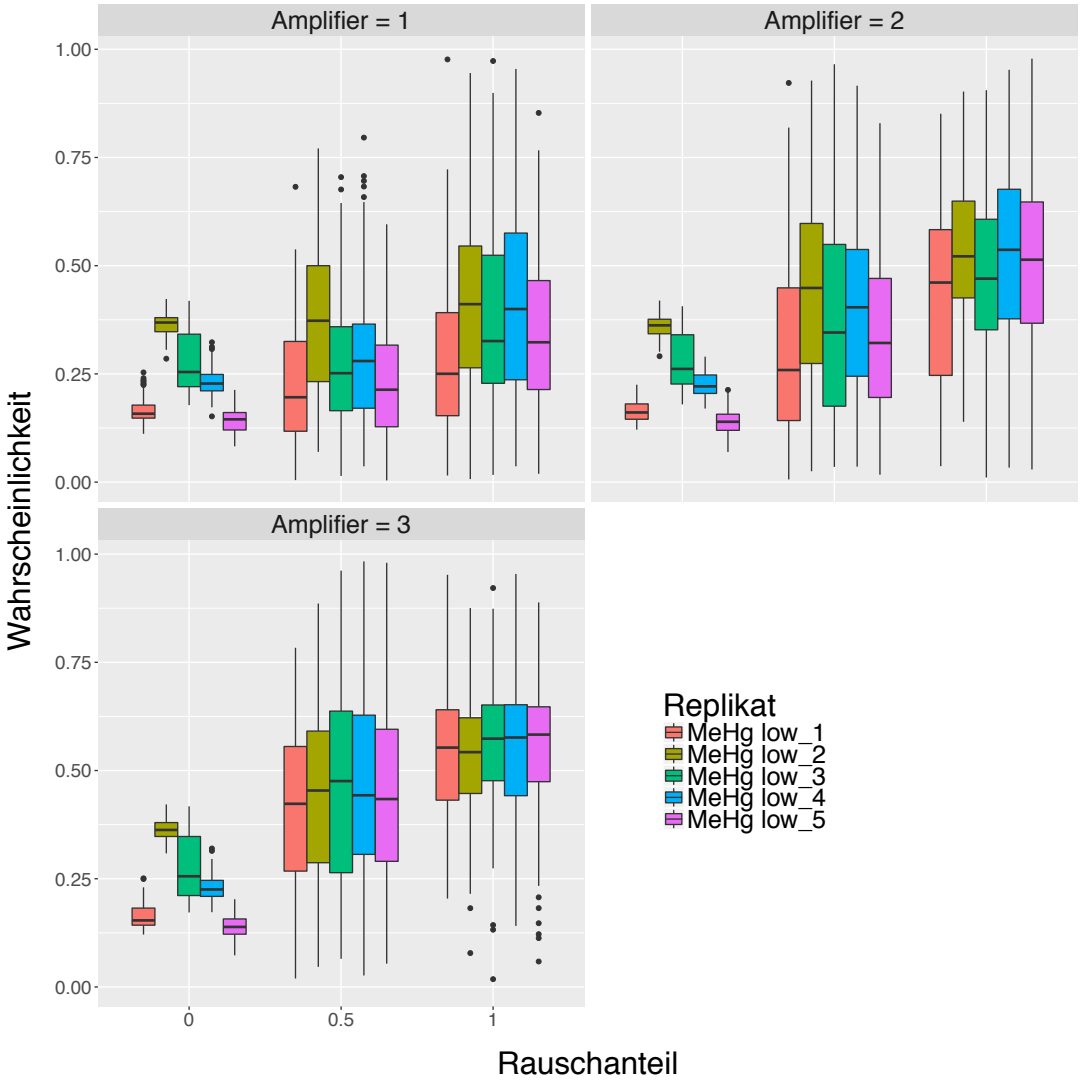

Abbildung 149: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg (kleinere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

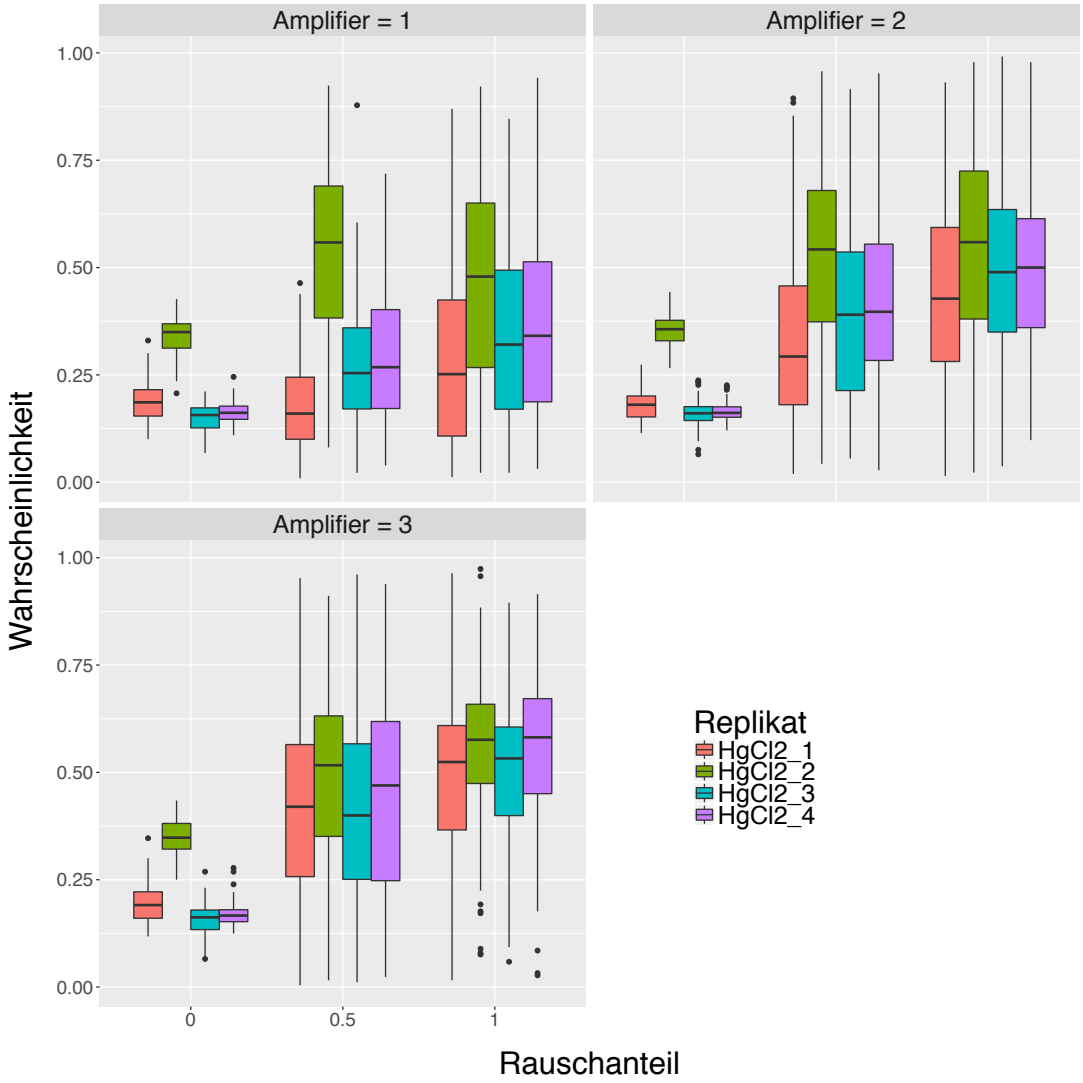

Abbildung 150: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgCl_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

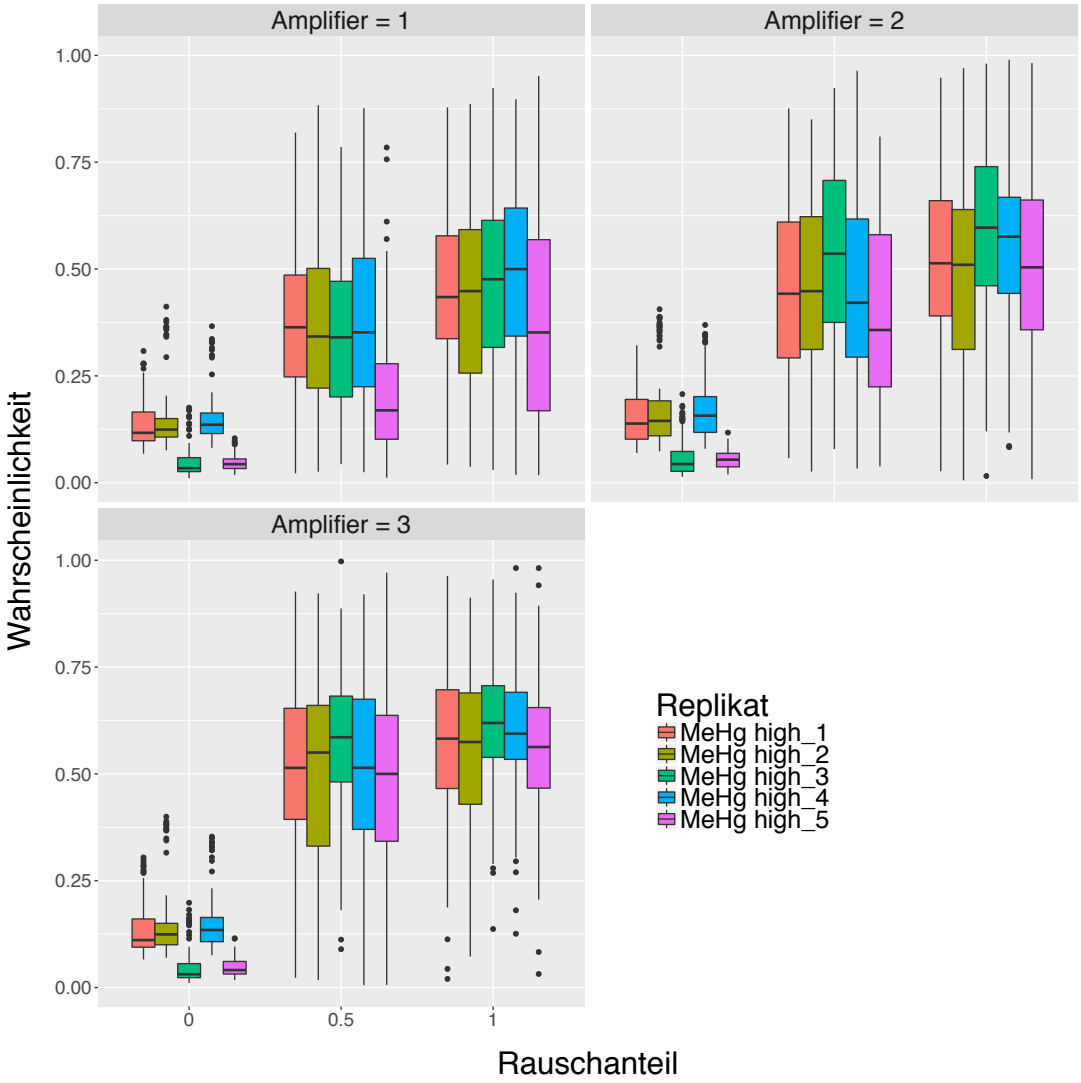

Abbildung 151: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von MeHg (höhere Konzentration) nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

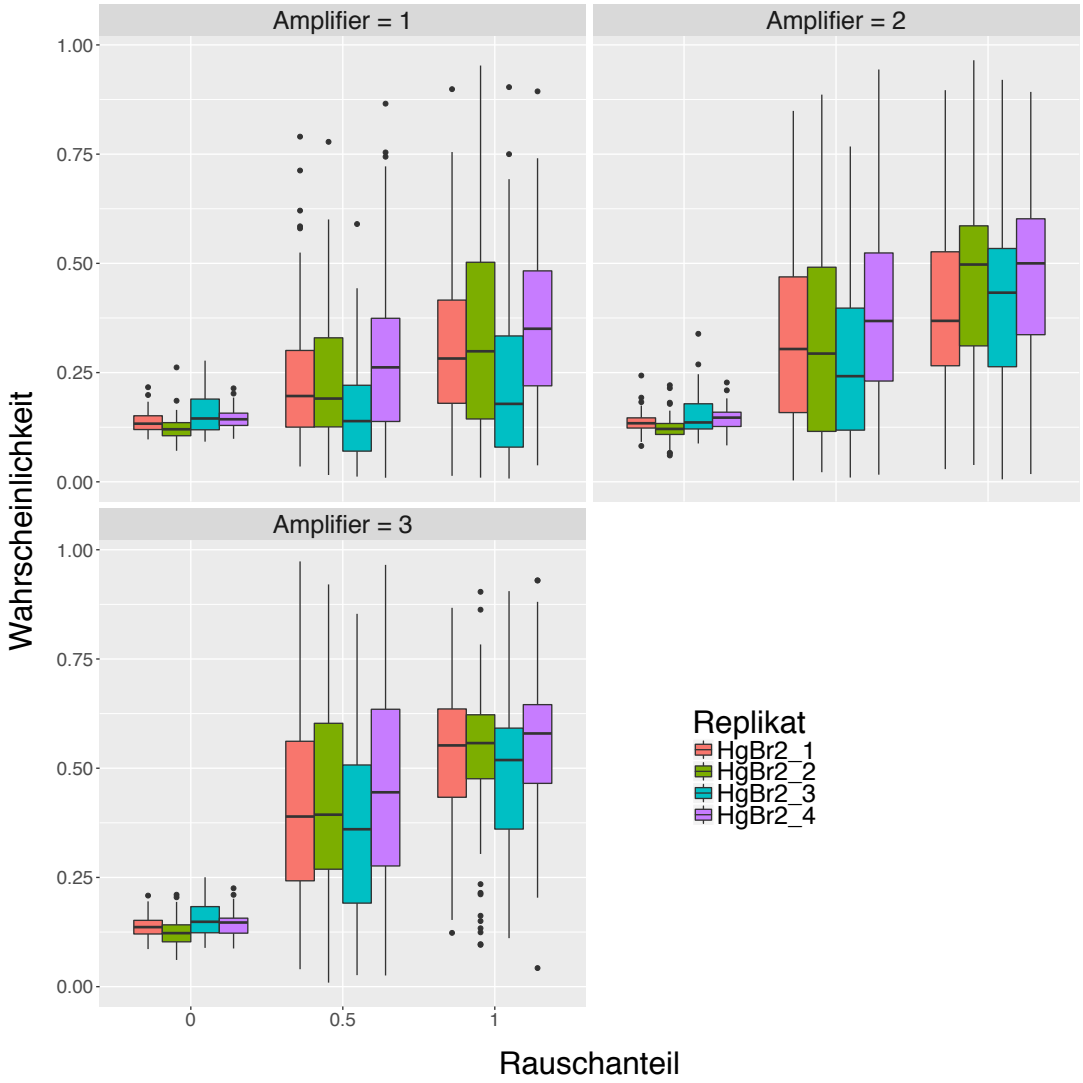

Abbildung 152: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von $HgBr_2$ nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

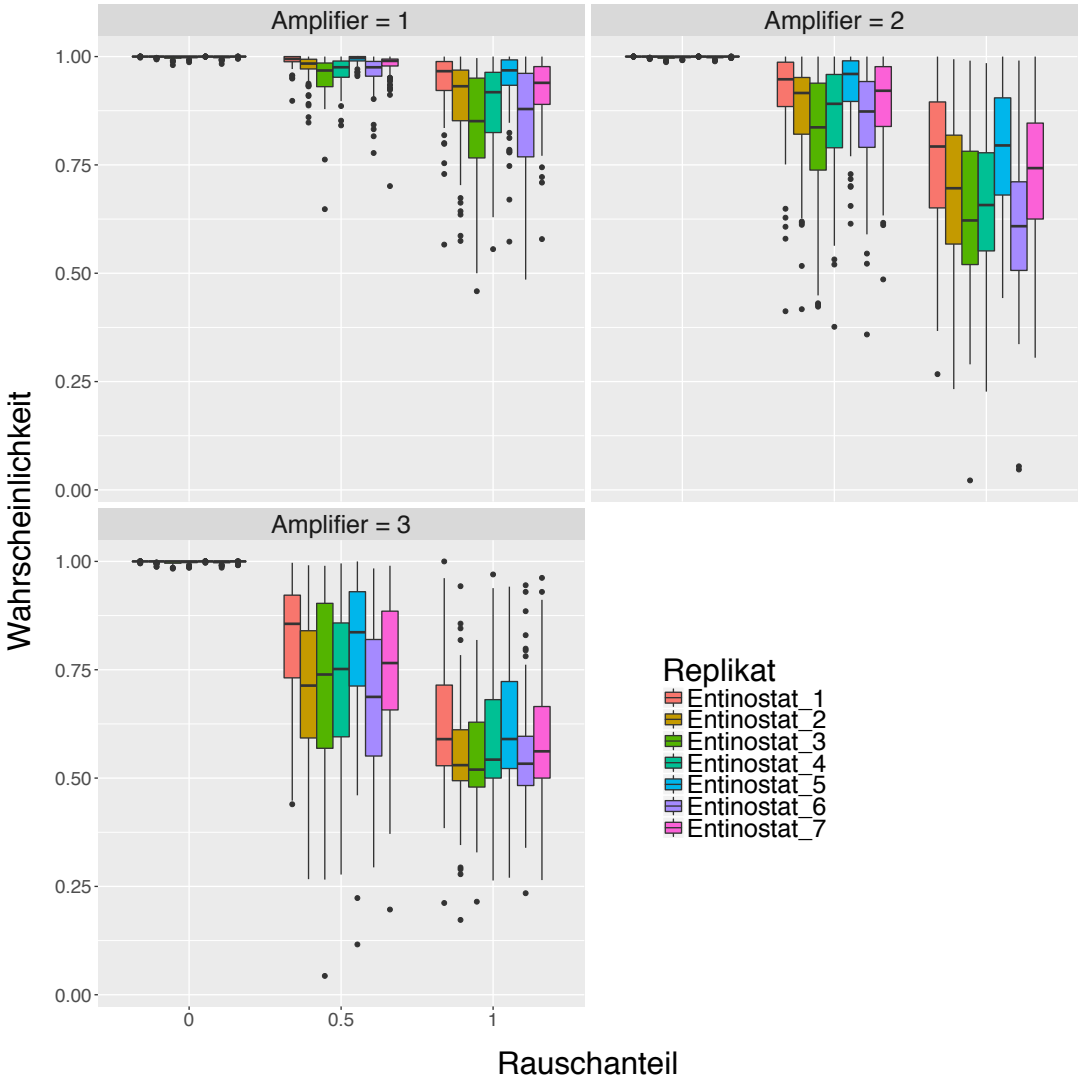

Abbildung 153: Boxplot-Graphiken für die Vorhersagewahrscheinlichkeiten von Entinostat nach dem Verrauschen der Daten für Rauschenstärke 1 (links obere Reihe), Rauschenstärke 2 (rechts obere Reihe) und Rauschenstärke 3 (links untere Reihe). Die Vorhersagewahrscheinlichkeiten für die einzelnen Proben sind durch unterschiedliche Farben gekennzeichnet. Auf der x-Achse ist der Anteil der verrauschten Variablen, auf der y-Achse die Vorhersagewahrscheinlichkeiten abgetragen. Für die Klassifikation wurde das Verfahren Support Vector Machines verwendet.

6.4 Verwendeter R-Code

Listing 1: R-Code für Hauptkomponentenplot

```
# function for pca plot
# INPUT:  data = expression data
#         nr = number of selected probesets
#         Overview = data frame of clinical variables
#         name.slot.1 = variable in Overview
pca.paper <- function(data, nr = 100, Overview, name.slot.1 = "compound",
                      name1 = "Compound", ppi = 600, height = 10, width = 10, name_file,
                                          xl, yl){
require(ggplot2)

myPalette <- c("#FF6347", "#6E8B3D", "#7FFF00","#FF34B3","#228B22","#3A5FCD",
               "#4876FF", "#7EC0EE", "#FF3E96", "#4F94CD","#27408B","#FF0000", "#6959CD",
               "#CD3700", "#8B2252")
names(myPalette) <- levels(Overview[,name.slot.fun])
levels.order <- c(2,3,5,8,7,13,10,6,11,1,4,9,12,14,15)

gene.var <- apply(data,1,var)
gen.sel <- order(gene.var, decreasing=TRUE)

esnatsII.pca <- prcomp(t(data[gen.sel[1:nr],]), retx = TRUE, scale.=FALSE, center=TRUE)

var.pca <- esnatsII.pca$sdev^2/sum(esnatsII.pca$sdev^2)
PC1.lab <- paste("PC1 (",round(100*var.pca[1],1), " %)",sep = "")
PC2.lab <- paste("PC2 (",round(100*var.pca[2],1), " %)",sep = "")

name.slot.fun <- which(names(Overview)==name.slot.1)



data2 <- data.frame(x = esnatsII.pca$x[,1], y= esnatsII.pca$x[,2]
                    ,type = Overview[,name.slot.fun])

mySequence <- levels(data2$type)[levels.order]


colScale  <- scale_color_manual(name = name1, breaks = mySequence, values = myPalette, drop = TRUE)

plot2 <- ggplot(data2, aes(x=x, y=y, col=type)) + geom_point(size = 4) +
  coord_fixed() +
  xlab(PC1.lab) + ylab(PC2.lab) + theme_bw() + xlim(xl) + ylim(yl) +
  # eliminates background and gridlines
  theme(
    axis.title.x = element_text(size=20)
    ,axis.text.x = element_text(size = 16)
    ,axis.title.y = element_text(size = 20)
    ,axis.text.y = element_text(size = 16)
    ,plot.background = element_blank()
    ,panel.grid.major = element_blank()
    ,panel.grid.minor = element_blank()
    ,legend.key = element_blank()
    ,legend.text = element_text(size=18)
    ,legend.title = element_text(size = 20)
     )

tiff(filename=name_file, width= width*ppi, height=height*ppi, compression="lzw", res= ppi)
print(plot2 +   colScale + guides(col=guide_legend(ncol=1,title.hjust=0.3,title.vjust=0.5)))
dev.off()
}
```

Listing 2: R-Code für GO-Anreicherungsanalyse

```
# This functions calculate for vector of differential expressed probesets the p-values of enrichment within all GO-
    groups.

# Input: GeneSig     = vector of significant probesets
#        GeneUniv    = vector of all probesets on chip


# Output NameErg.RData with data frame with variables
#           GO.ID                = GO identifier,
#            Term                = Term of GO group (sometimes abridged, full Term to get with GOTERM function from GO
    .db,
#       Annotated                = how many probesets are annotated to the specific GO group,
```

```
#     Significant                 = how many probesets from GO are significant,
#        Expected                 = how many probesets are expected to be significant given null hypothesis,
#            elim                 = p-value of enrichment using elimCount

# packages for the chip
# to be changed in case of another chip

library(hgu133plus2.db)
library(topGO)

affyLib <- "hgu133plus2.db"

GO_ana_list <- function(GeneSig, GeneUniv){
        geneList <- factor(as.integer(GeneUniv %in% GeneSig))
        names(geneList) <- GeneUniv

        GOdata <- new("topGOdata", ontology = "BP", allGenes = geneList, annot = annFUN.db, affyLib = affyLib)

        test.stat.elim <- new("elimCount", testStatistic = GOFisherTest, name = "Fisher test")
        resultElim <- getSigGroups(GOdata, test.stat.elim)

        Result.Elim <- GenTable(GOdata, elim = resultElim, topNodes = length(resultElim@score))
        return(Result.Elim)

}
```

Listing 3: R-Code für Training von SVM und Klassifikation

```
# training and prediction with svm
# INPUT: expr.norm = normalized expression data which will be split in training und test set
#        compounds = compounds in the test set, levels in the description.classifier$compound
#        description.classifier = data.frame with columns "compound", "type" with classes of compounds
#        ngenes = number of probesets to be used
#        kern = kernel of the svm
# OUTPUT: matrix with predicted probabilities of compounds in the test set, column names are levels of description.
    classifier$type


classification.svm <- function(expr.norm, compounds, description.classifier, ngenes = 100, kern = "linear"){

  require(e1071)

  id.train <- which(!(as.character(description.classifier$compound) %in% compounds))
  id.test <- which(as.character(description.classifier$compound) %in% compounds)

  test.compounds <- droplevels(description.classifier$compound)[id.test]

  train <- expr.norm[,id.train]
  test <- expr.norm[, id.test]

  gene.var <- apply(train,1,var)
  gene.order <- order(gene.var, decreasing=TRUE)

  train.data <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                           t(train[gene.order[1:ngenes],]),
                           row.names = 1:ncol(train))

  if(ngenes == 1){
        test.data <- data.frame(test[gene.order[1:ngenes],], row.names = 1:ncol(test))
        names(test.red) <- "X1"
        train.data <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                           train[gene.order[1:ngenes],],
                           row.names = 1:ncol(train))
        names(train.data) = c("Klasse","X1")
        }
  else {
        test.data <- data.frame(t(test[gene.order[1:ngenes],]), row.names = 1:ncol(test))
        train.data <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                           t(train[gene.order[1:ngenes],]),
                           row.names = 1:ncol(train))
        }


  tuning <- tune(svm, Klasse~., data = train.data,
                 ranges = list(gamma = 2^(-5:2), cost = 2^(-5:2)),
                 probability = TRUE, kernel = kern)

  prediction <- predict(tuning$best.model, test.data, probability=TRUE)
  prediction.matrix <- attr(prediction, "probabilities")

  rownames(prediction.matrix) <- test.compounds
  return(prediction.matrix)

}
```

Listing 4: R-Code für Training von Random Forest und Klassifikation

```
# training and prediction with Random Forest
# INPUT: expr.norm = normalized expression data which will be split in training und test set
#        compounds = compounds in the test set, levels in the description.classifier$compound
#        description.classifier = data.frame with columns "compound", "type" with classes of compounds
#        ngenes = number of probesets to be used
#        mtry.n = parameter for Random Forest optimization

# OUTPUT: matrix with predicted probabilities of compounds in the test set, column names are levels of description.
    classifier$type

classification.rf <- function(expr.norm, compounds, description.classifier,ngenes,mtry.n){

id.train <- which(!(as.character(description.classifier$compound) %in% compounds))
id.test <- which(as.character(description.classifier$compound) %in% compounds)

test.compounds <- droplevels(description.classifier$compound)[id.test]

train <- expr.norm[,id.train]
test <- expr.norm[, id.test]

gene.var <- apply(train,1,var)
gene.order <- order(gene.var, decreasing=TRUE)

train.data <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                         t(train[gene.order[1:ngenes],]),
                         row.names = 1:ncol(train))

test.data <- data.frame(t(test[gene.order[1:ngenes],]), row.names = 1:ncol(test))

model <- train(Klasse ~., train.data, method = 'rf'
               , tuneGrid=expand.grid(mtry = mtry.n)
               , trControl=trainControl(method='cv', number = 3, classProbs=TRUE))

pred <- predict(model,test.data,"prob")
rownames(pred) <- paste(as.character(test.compounds)
                         ,1:length(test.compounds),sep = "_")

return(pred)
}
```

Listing 5: R-Code für SVM basierte Vorhersage unter Auslassen von Replikaten

```
# training and prediction with SVM with subset of available replicas

# INPUT:
# expr.norm = normalized expression data which will be split in training und test set
#       compounds = compounds in the test set, levels in the description.classifier$compound
#       description.classifier = data.frame with columns "compound", "type" with classes of compounds
#       ngenes = number of probesets to be used
# kern = kernel used for svm
# number_keep = number of used replicas
# n_rep = number of drawn samplings

# OUTPUT: list of altogether n_rep matrices with predicted probabilities of compounds in the test set, column names are
     levels of description.classifier$type


class.esnats.red <- function(expr.norm, compounds, description.classifier
                             , ngenes = 100, kern = "linear"
                             , number_keep, n_rep = 100){

  id.train <- which(!(as.character(description.classifier$compound) %in% compounds))
  id.test <- which(as.character(description.classifier$compound) %in% compounds)

  test.compounds <- droplevels(description.classifier$compound)[id.test]

  clinic.train = droplevels(description.classifier[id.train,])


  train <- expr.norm[,id.train]
  test <- expr.norm[, id.test]

  results <- vector('list', length = n_rep)

  for(i in 1:n_rep){

  id.train.keep = sort(unlist(tapply(X = 1:nrow(clinic.train), INDEX = clinic.train$compound, FUN = function(x) sample(
       x, number_keep))))

  train_loop = train[, id.train.keep]

  gene.var <- apply(train_loop,1,var)
  gene.order <- order(gene.var, decreasing=TRUE)

  train.red <- data.frame(Klasse = droplevels(clinic.train$type[id.train.keep]),
                          t(train_loop[gene.order[1:ngenes],]),
                          row.names = 1:ncol(train_loop))

  if(ngenes == 1){test.red <- data.frame(test[gene.order[1:ngenes],], row.names = 1:ncol(test)); names(test.red) <- "X1
       ";   train.red <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                          train_loop[gene.order[1:ngenes],],
                          row.names = 1:ncol(train_loop)); names(train.red) = c("Klasse","X1")} else {  test.red <-
                               data.frame(t(test[gene.order[1:ngenes],]), row.names = 1:ncol(test));   train.red <-
                               data.frame(Klasse = droplevels(clinic.train$type[id.train.keep]),
                          t(train_loop[gene.order[1:ngenes],]),
                          row.names = 1:ncol(train_loop))}

  #print(train.red)

  tuning <- tune(svm, Klasse~., data = train.red,
                 ranges = list(gamma = 2^(-5:2), cost = 2^(-5:2)),
                 probability = TRUE, kernel = kern)

  prediction <- predict(tuning$best.model, test.red, probability=TRUE)
  pred.matrix <- attr(prediction, "probabilities")

  rownames(pred.matrix) <- test.compounds
  results[[i]] = pred.matrix
  }
  return(results)

}
```

Listing 6: R-Code für RF basierte Vorhersage unter Auslassen von Replikaten

```
# training and prediction with Random Forest with subset of available replicas

# INPUT:
# expr.norm = normalized expression data which will be split in training und test set
#       compounds = compounds in the test set, levels in the description.classifier$compound
#       description.classifier = data.frame with columns "compound", "type" with classes of compounds
#       ngenes = number of probesets to be used
#       mtry.n = parameter for Random Forest optimization
# number_keep = number of used replicas
# n_rep = number of drawn samplings

# OUTPUT: list of altogether n_rep matrices with predicted probabilities of compounds in the test set, column names are
      levels of description.classifier$type

class.esnats.red <- function(expr.norm, compounds
                             , description.classifier, ngenes
                             , mtry.n, n_rep, number_keep){

  id.train <- which(!(as.character(description.classifier$compound) %in% compounds))
  id.test <- which(as.character(description.classifier$compound) %in% compounds)

  test.compounds <- droplevels(description.classifier$compound)[id.test]

  clinic.train = droplevels(description.classifier[id.train,])


  train <- expr.norm[,id.train]
  test <- expr.norm[, id.test]

  results <- vector('list', length = n_rep)

  for(i in 1:n_rep){

    id.train.keep = sort(unlist(tapply(X = 1:nrow(clinic.train), INDEX = clinic.train$compound, FUN = function(x)
        sample(x, number_keep))))

    train_loop = train[, id.train.keep]

    gene.var <- apply(train_loop,1,var)
    gene.order <- order(gene.var, decreasing=TRUE)

    train.red <- data.frame(Klasse = droplevels(clinic.train$type[id.train.keep]),
                            t(train_loop[gene.order[1:ngenes],]),
                            row.names = 1:ncol(train_loop))

   test.red <- data.frame(t(test[gene.order[1:ngenes],]), row.names = 1:ncol(test))

   model <- train(Klasse ~., train.red, method = 'rf'
                  , tuneGrid=expand.grid(mtry = mtry.n)
                  , trControl=trainControl(method='cv', number = 3, classProbs=TRUE))

   pred <- predict(model,test.red,"prob")
   rownames(pred) <- paste(as.character(test.compounds)
                           , 1:length(test.compounds),sep = "_")

    results[[i]] = pred
  }
  return(results)

}
```

Listing 7: R-Code für SVM basierte Vorhersage von verrauschten Daten

```
# training and prediction with RF with noisy data
# INPUT:
# noise = proportion of perturbed variables, numeric within [0, 1]
# amplifier = strengt of noise, integer
# expr.norm = normalized expression data which will be split in training und test set
#       compounds = compounds in the test set, levels in the description.classifier$Compound
#       description.classifier = data.frame with columns "Compound", "Klasse" with classes of compounds
#       ngenes = number of probesets to be used
#       mtry.n = parameter for Random Forest optimization
# test.noise = perturb test set, logical
# train.noise = perturb train set, logical

# OUTPUT: matrix with predicted probabilities of compounds in the test set, column names are levels of description.classifier$Klasse


noise2fun_rf <- function(noise, amplifier, expr.norm, compounds
                         , description.classifier, ngenes
                         , mtry.n, test.noise = TRUE, train.noise = TRUE){

  id.train <- which(!(as.character(description.classifier$Compound) %in% compounds))
  id.test <- which(as.character(description.classifier$Compound) %in% compounds)

  test.compounds <- droplevels(description.classifier$Compound)[id.test]

  train <- expr.norm[,id.train]
  test <- expr.norm[, id.test]

  gene.sd <- apply(train,1,sd)
  gene.order <- order(gene.sd, decreasing=TRUE)
  gene.sel = gene.order[1:ngenes]

  train = train[gene.sel, ]
  test = test[gene.sel, ]

  ###
  sigma = gene.sd[gene.sel] # standard deviation of selected features

  prob_val = matrix(data = runif(n = ngenes*ncol(expr.norm), min = 0, max = 1) # uniformly distributed random variable
                    , ncol = ngenes
  )

  prob_matr = qnorm(prob_val) # transformation in normally distributed variable

  cut_matr = matrix(as.numeric(prob_val > 1-noise),ncol = ngenes) # add noise if  prob_val > 1-noise
  add_matrix = t((cut_matr*prob_matr) %*% diag(amplifier*sigma))

  train.new = train + add_matrix[,id.train] # perturb train set
  test.new = test + add_matrix[,id.test] # perturb test set

  if(!train.noise){train.new = train}
  if(!test.noise){test.new = test}

  train.data <- data.frame(Klasse = droplevels(description.classifier$Klasse[id.train]),
                           t(train.new),
                           row.names = 1:ncol(train.new))

  test.data <- data.frame(t(test.new), row.names = 1:ncol(test.new))

  # build the probabilistic classifier
  model <- train(Klasse ~., train.data, method = 'rf'
                 , tuneGrid=expand.grid(mtry = mtry.n)
                 , trControl=trainControl(method='cv', number = 3, classProbs=TRUE))

  pred <- predict(model,test.data,"prob")
  rownames(pred) <- paste(as.character(test.compounds)
                          , 1:length(test.compounds),sep = "_")

  return(pred)}
```

Listing 8: R-Code für RF basierte Vorhersage von verrauschten Daten

```
# training and prediction with SVM with noisy data
# INPUT:
# noise = proportion of perturbed variables, numeric within [0, 1]
# amplifier = strengt of noise, integer
# expr.norm = normalized expression data which will be split in training und test set
#       compounds = compounds in the test set, levels in the description.classifier$Compound
#       description.classifier = data.frame with columns "Compound", "Klasse" with classes of compounds
#       ngenes = number of probesets to be used
#       mtry.n = parameter for Random Forest optimization
# test.noise = perturb test set, logical
# train.noise = perturb train set, logical

# OUTPUT: matrix with predicted probabilities of compounds in the test set, column names are levels of description.
    classifier$Klasse

noise2fc_svm = function(noise, amplifier, n.genes, compounds
                    , expr.norm, description.classifier, kern
                    , test.noise = TRUE, train.noise = TRUE){


id.train <- which(!(as.character(description.classifier$compound) %in% compounds))
id.test <- which(as.character(description.classifier$compound) %in% compounds)

test.compounds <- droplevels(description.classifier$compound)[id.test]

train <- expr.norm[,id.train]
test <- expr.norm[, id.test]

gene.sd <- apply(train,1,sd)
gene.order <- order(gene.sd, decreasing=TRUE)
gene.sel = gene.order[1:n.genes]

train = train[gene.sel, ]
test = test[gene.sel, ]

###
sigma = gene.sd[gene.sel] # standard deviation of selected features

prob_val = matrix(data = runif(n = n.genes*ncol(expr.norm), min = 0, max = 1) # uniformly distributed random variable
                  , ncol = n.genes
                  )

prob_matr = qnorm(prob_val) # transformation in normally distributed

cut_matr = matrix(as.numeric(prob_val > 1-noise),ncol = n.genes) # add noise if  prob_val > 1-noise
add_matrix = t((cut_matr*prob_matr) %*% diag(amplifier*sigma))

train.new = train + add_matrix[,id.train]
test.new = test + add_matrix[,id.test]

if(!train.noise){train.new = train}
if(!test.noise){test.new = test}

train.red <- data.frame(Klasse = droplevels(description.classifier$type[id.train]),
                        t(train.new),
                        row.names = 1:ncol(train.new))

test.red <- data.frame(t(test.new), row.names = 1:ncol(test.new))

tuning <- tune(svm, Klasse~., data = train.red,
               ranges = list(gamma = 2^(-5:2), cost = 2^(-5:2)),
               probability = TRUE, kernel = kern)

prediction <- predict(tuning$best.model, test.red, probability=TRUE)
pred.matrix <- attr(prediction, "probabilities")
rownames(pred.matrix) <- test.compounds
pred.matrix
}
```

Erklärung

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

$Unterschrift$: $Ort, Datum$: