**Antriebe in der Fahrzeugtechnik** 4
Hrsg. von Prof. Dr.-Ing. Michael Günthner

Tobias Gehra

# KI-unterstützte hybride Modellierung von Emissionen im hochtransienten Motorbetrieb

λογος

Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Antriebe in der Fahrzeugtechnik

*Band 4*

# Antriebe in der Fahrzeugtechnik

*Band 4*

Herausgegeben von

**Prof. Dr.-Ing. Michael Günthner**
Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau
Lehrstuhl für Antriebe in der Fahrzeugtechnik

Tobias Gehra

# KI-unterstützte hybride Modellierung von Emissionen im hochtransienten Motorbetrieb

Logos Verlag Berlin

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.




Logos Verlag Berlin GmbH 2025
ISBN 978-3-8325-5940-3
ISSN 2941-4326

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
12681 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.de

# KI-unterstützte hybride Modellierung von Emissionen im hochtransienten Motorbetrieb

Vom Fachbereich Maschinenbau und Verfahrenstechnik
der RPTU Kaiserslautern-Landau
zur Verleihung des akademischen Grades

**Doktor-Ingenieur (Dr.-Ing.)**
genehmigte
**Dissertation**

von
Herrn
**M.Sc. Tobias Gehra**
aus Sindelfingen

D 386
Kaiserslautern 2025

| | |
|---|---|
| Dekan: | Prof. Dr. rer. nat. Roland Ulber |
| Vorsitzender | Prof. Dr.-Ing. Jörg Seewig |
| Berichterstatter: | Prof. Dr.-Ing. Michael Günthner |
| | Prof. Dr.-Ing. Daniel Görges |
| Eingereicht am: | 17.10.2024 |
| Mündliche Prüfung: | 21.02.2025 |

# Vorwort des Herausgebers

Die Diskussion um zukünftige Technologien für Fahrzeugantriebe wird vielfach mit großer Leidenschaft geführt. Umso wichtiger sind belastbare Informationen zu den Möglichkeiten und Grenzen neuer Technologien. Die vorliegende Schriftenreihe des Lehrstuhls für Antriebe in der Fahrzeugtechnik der RPTU Kaiserslautern-Landau möchte objektive, wissenschaftlich fundierte Einblicke in den aktuellen Stand der Forschung und Ausblicke auf zukünftige Entwicklungen im Bereich der Fahrzeugantriebe bieten. Für die schnelle Umsetzung wissenschaftlicher Erkenntnis in praktische Anwendungen sind der Austausch mit anderen Forschern und Entwicklern sowie der erfolgreiche Transfer in die Industrie von besonderer Bedeutung. Neben Vorträgen auf Fachtagungen und Veröffentlichungen in Fachzeitschriften soll hierzu auch diese Schriftenreihe einen Beitrag leisten.

Im Mittelpunkt aktueller Antriebsforschung stehen einerseits besonders schadstoffarme bzw. emissionsfreie Antriebstechnologien und andererseits die Minimierung des $CO_2$-Ausstoßes über den Produktlebenszyklus. Zukunftsfähige Antriebe müssen beiden Anforderungen gerecht werden. Mit dieser Zielsetzung ergibt sich ein äußerst weiter Lösungsraum für zukünftige Antriebe, der von batterieelektrischen Antrieben über Brennstoffzellen-elektrische Konzepte bis hin zu verbrennungsmotorischen Lösungen mit regenerativen Kraftstoffen reicht. Noch weitergehende Gestaltungsmöglichkeiten bieten hybride Antriebskonzepte, beispielsweise durch die Kombination des Hauptantriebs mit weiteren elektrischen, hydraulischen oder pneumatischen Speichern und Antriebskomponenten, ggf. auch für Nebenantriebe.

Im Bereich der Antriebsforschung ist sich die Wissenschaft mittlerweile einig, dass es absehbar nicht „die“ eine einzige und universell anwendbare, ideale Antriebslösung für alle Anwendungen geben wird. Vielmehr ist es wichtig, für jede Antriebsaufgabe die jeweils optimale Lösung zu finden, welche einerseits die technischen Anforderungen des anzutreibenden Fahrzeugs bzw. Arbeitsgeräts zufriedenstellend erfüllt und andererseits den minimal möglichen Einfluss auf die Umgebung nimmt – sowohl im Hinblick auf die Schadstoffemissionen als auch hinsichtlich des Ausstoßes klimaschädlicher Gase. Erklärtes Ziel ist hierbei „zero impact“ – die Konzentrationen der potenziell kritischen Spezies im Abgas sollen auf dem Niveau des Hintergrunds bzw. darunter liegen. Bei Verbrennungsmotoren ist hierfür neben einer hochwirksamen Abgasnachbehandlung insbesondere die Vermeidung der Emissionsbildung bereits an der Quelle, also beim Verbrennungsprozess, von zentraler Bedeutung.

In diesem Zusammenhang beschäftigt sich Herr Gehra im Rahmen seiner hier vorliegenden Dissertation mit neuen Ansätzen zur Modellierung und Vorhersage der Bildung von Schadstoffemissionen in realen Betriebsszenarien von Fahrzeugen. Im Zentrum stehen hochtransiente Betriebszustände, die als besonders kritisch für die Rohemissionsbildung eingeschätzt werden

und in der Praxis einen wesentlichen Beitrag zu den gesamten Emissionen eines Fahrzeugs liefern. In Verbindung mit einer Vorausschau eröffnet die zuverlässige Vorhersage emissionskritischer Situationen die Möglichkeit, das Fahrzeugantriebssystem und ggf. auch seine Peripherie hierauf vorzubereiten bzw. die emissionskritische Situation zu vermeiden oder zumindest abzumildern. Besondere Relevanz gewinnt dies im Kontext hybrider Antriebssysteme, die – innerhalb der durch die jeweilige Systemkonfiguration gegebenen Grenzen – eine Aufteilung der Antriebsleistung zwischen dem verbrennungs- und dem elektromotorischen Anteil erlauben und damit neben der Erhöhung der Effizienz auch eine Reduktion des Schadstoffausstoßes durch optimierte Betriebsstrategien erlauben.

In seiner Arbeit entwickelt und bewertet Herr Gehra besonders effiziente Ansätze zur Modellierung der Schadstoffbildung im transienten Betrieb, einerseits basierend auf physikalischen Zusammenhängen und andererseits unter Zuhilfenahme von Methoden des maschinellen Lernens. In einem weiteren Schritt kombiniert er beide Methoden, was einen erheblichen Innovationsschritt gegenüber bisherigen Ansätzen darstellt. Hierdurch gelingt es ihm, die Genauigkeit der Vorhersage nochmals deutlich zur erhöhen. Durch die Einbindung solcher Modelle in Rahmen der Entwicklung, perspektivisch aber sogar auch in die Antriebssteuerung von (insbesondere Hybrid-) Fahrzeugen zur Anpassung der Betriebsstrategie in Echtzeit sind wesentliche weitere Verbesserungen der Abgasqualität auch in herausfordernden Betriebssituationen zu erwarten, die einen signifikanten Beitrag zum Ziel des „zero impact“ leisten können.

Kaiserslautern, im März 2025 Michael Günthner

# Vorwort

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Antriebe in der Fahrzeugtechnik (LAF) des Fachbereichs Maschinenbau und Verfahrenstechnik der Rheinland-Pfälzischen Technischen Universität Kaiserslautern-Landau. Mein Dank gilt vorab allen, die mich auf meinem Weg begleitet und mich dabei direkt oder indirekt unterstützt haben.

Zunächst möchte ich mich bei Herrn Prof. Michael Günthner für das entgegengebrachte Vertrauen, die Möglichkeit zur Promotion und die stetige Unterstützung während meiner Zeit am Lehrstuhl bedanken. Mein Dank gilt ebenso Herrn Prof. Daniel Görges für die Begutachtung meiner Dissertation und die hervorragende Zusammenarbeit im gemeinsamen Förderprojekt sowie Herrn Prof. Jörg Seewig für die Übernahme des Prüfungsvorsitzes.

Ich habe den LAF als eine eingeschworene Gemeinschaft erleben dürfen, die trotz aller Widrigkeiten stets dafür gesorgt hat, dass eines nie zu kurz kommt: der Spaß an der Arbeit. Hierfür und für die tatkräftige Unterstützung möchte ich mich im höchsten Maße bei meinen Kolleginnen und Kollegen am Lehrstuhl bedanken. Egal, wie groß die Herausforderung war, ich habe mich nie allein gefühlt, und selbst wenn keiner eine Lösung parat hatte, konnte jedes Problem mit Humor entschärft werden. In diesem Zusammenhang möchte ich auch Ganesh Sundaram, Mirjan Heubaum und Jonas Ulmen nennen. Die Zusammenarbeit war mir stets eine Freude und brachte großartige Ergebnisse hervor.

Die Reihe der Kollegen am LAF, zu denen ich jetzt zusätzlich persönliche Worte finden könnte, ist lang, und die dazugehörigen Geschichten würden den Rahmen sprengen. Daher möchte ich in diesem Vorwort nur meinen Bürokollegen Alexander Weigel hervorheben. Alex hat mich bereits vor meiner Zeit am Lehrstuhl durch das Masterstudium begleitet, war stets verlässlich, und seine selbstlose Hilfsbereitschaft ist eine Bereicherung für den gesamten Lehrstuhl – vielen Dank!

Es ist offensichtlich, aber kann nicht deutlich genug gesagt werden: Ich verdanke die mir gegebenen Möglichkeiten zu einem großen Teil meinen Eltern. Sie haben mich immer unterstützt und bedingungslos an mich geglaubt. Ohne sie wäre das alles nicht möglich gewesen.

Abschließend möchte ich mich bei meiner wundervollen Frau für ihre Geduld, die Motivation und ihre aufmunternden Worte in schwierigen Phasen bedanken. Sie unterstützt mich bei allem, was ich tue, und ich könnte mir niemand Besserem an meiner Seite vorstellen. Danke auch an Maja, die jeden Tag zu einem besseren macht.

Kaiserslautern, im März 2025 Tobias Gehra

# Kurzfassung

Bedingt durch den Klimawandel, ist die Notwendigkeit emissionsarmer Antriebstechnologien vermehrt in den Fokus der Forschung gerückt. Batterieelektrische Fahrzeuge (BEV), die lokal emissionsfrei fahren und einen hohen Gesamtwirkungsgrad aufweisen, stellen einen möglichen Lösungsansatz dar. Sie besitzen jedoch auch Nachteile, wie die ausbaufähige Ladeinfrastruktur und vergleichbar geringe Reichweiten im Kompakt- und Mittelklassesegment. Hybridfahrzeuge können diese Nachteile ausgleichen und haben das Potenzial, Vorteile beim Kraftstoffverbrauch und dem Emissionsausstoß gegenüber rein verbrennungsmotorisch betriebenen Fahrzeugen zu erzielen.
Um diese Vorteile zu nutzen und Optimierungsziele (beispielsweise Emissionsreduktion) zu erreichen, sind spezielle Betriebsstrategien notwendig. Diese lassen sich besonders effizient in einer Simulationsumgebung entwickeln, was präzise Modelle der Fahrzeugsysteme und der Umgebung erfordert. In dieser Arbeit werden verschiedene Ansätze zur Modellierung von Emissionen im hochtransienten Motorbetrieb untersucht, mit dem Fokus auf einer präzisen Emissionsvorhersage im realen Straßenbetrieb. Neben den bekannten physikalisch-phänomenologischen Modellen kommen auch Methoden des Maschinellen Lernens (ML) zum Einsatz, und die Kombination beider Ansätze (Hybridmodelle) wird getestet. Es zeigt sich, dass die ML-Modelle im dynamischen Bereich Vorteile gegenüber den physikalisch-phänomenologischen Modellen aufweisen und mit einem sehr geringen Root Mean Square Error (RMSE) den Testzyklus präzise abbilden können. Damit ist Modell für Regelungs- und Optimierungsaufgaben in Echtzeit geeignet. Das parallele Hybridmodell verbessert das Ergebnis um circa 25 % und zeigt, wie datenbasierte Modelle mit physikalischen Informationen bei gleichem Trainingsumfang optimiert werden können.

# Abstract

Due to climate change, the necessity for low-emission propulsion technologies has become an increasingly important topic in development. Battery electric vehicles (BEVs), which can operate locally emission-free and have high overall efficiency, are one possible solution to achieve this goal. However, they also have disadvantages, such as the limited charging infrastructure and comparatively low ranges in the compact and mid-size segments. Hybrid electric vehicles can mitigate these disadvantages and have the potential to offer benefits in terms of fuel consumption and emissions compared to purely internal combustion engine vehicles.
To utilize these advantages, specific operating strategies are necessary to optimize goals like emission reduction. These strategies can be developed and optimized efficiently in simulation environments, which demand precise models of the vehicle's subsystems and environment. This work examines various approaches for modeling emissions in highly transient engine operation, with a focus on accurately forecasting emissions in real driving conditions. Alongside traditional physical-phenomenological emission models, Machine Learning (ML) methods are also used, and hybrid models combining both approaches are tested. The results show that ML methods outperform physical-phenomenological models in dynamic scenarios, achieving a very low RMSE value in test cycles. Thus, the model is suitable for control and optimization tasks and is real-time capable. The parallel hybrid model improves this result by approximately 25 %, demonstrating how a data-based model can be enhanced with physical information without increasing the training scope.

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| 0D, 1D, 3D | Nulldimensional, Eindimensional, Dreidimensional |
| AGR | Abgasrückführung |
| AHRP | Allianz für Hochleistungsrechnen Rheinland-Pfalz |
| ASAM | Association for Standardisation of Automation and Measuring Systems |
| BEV | Battery Electric Vehicle (dt.: Batterieelektrisches Fahrzeug) |
| BMW | Bayerische Motoren Werke |
| CAD | Computer-aided Design (dt.: Rechnerunterstütztes Konstruieren) |
| CAN | Controller Area Network (dt.: etwa Serielles Bussystem) |
| CFD | Computational Fluid Dynamics (dt.: Numerische Strömungsmechanik) |
| CUDA | Compute Unified Device Architecture (dt.: Einheitliche Gerätearchitektur für die Datenverarbeitung) |
| DB | Datenbasis |
| ECU | Engine Control Unit (dt.: Motorsteuergerät) |
| E | Elektrisch |
| FNN | Feed Forward Neural Network (dt.: Vorwärtsgerichtetes Neuronales Netz) |
| GPS | Global Positioning System (dt.: Globales Positionsbestimmungssystem) |
| GPU | Graphics Processor Unit (dt.: Grafikprozessor) |
| HCCI | Homogeneous Charge Compression Ignition (dt.: Homogene Kompressionszündung) |
| KI | Künstliche Intelligenz |
| KW | Kurbelwinkel |
| LAF | Lehrstuhl für Antriebe in der Fahrzeugtechnik |
| LSTM | Long Short-Term Memory (dt.: Langes Kurzzeitgedächtnis) |
| ML | Maschinelles Lernen |
| ML-MoRE | Maschinelles Lernen für die Modellierung und Regelung der Emissionen von Hybridfahrzeugen in Realfahrzeugen |
| MLP | Multy Layer Perceptron (dt.: Mehrlagiges Perzeptron) |
| NEFZ | Neue Europäische Fahrzyklus |
| OB | On-Board-Diagnose-System |
| OT | Oberer Totpunkt |
| PAC | Probably Approximately Correct (dt.: Wahrscheinlich Annähernd Richtig) |
| PEMS | Portables Emissionsmesssystem |

| | |
|---|---|
| PINN | Physikalisch Informierte Neuronale Netze |
| PKW | Personenkraftwagen |
| RDE | Real Driving Emissions (dt.: Abgasemissionsverhalten im realen Fahrbetrieb) |
| ReLU | Rectified Linear Unit – Aktivierungsfunktion („Gleichrichter“) |
| RMSE | Root Mean Square Error (dt.: Wurzel des mittleren quadratischen Fehlers) |
| RNN | Recurrent Neural Network (dt.: Rekurrentes Neuronales Netz) |
| SVR | Support Vector Regression (dt.: Stützvektor Regression) |
| SVM | Support Vector Machine (dt.: Stützvektormaschine) |
| TPA | Three Pressure Analysis (dt.: Dreidruckanalyse) |
| TPE | Tree-structured Parzen Estimator |
| UNFCCC | United Nations Framework Convention on Climate Change (dt.: Klimarahmenkonvektion der Vereinten Nationen) |
| WLTC | Worldwide Harmonized Light Duty Vehicle Test Cycle (dt.: weltweit einheitlicher Leichtfahrzeuge-Testzyklus) |
| WLTP | Worldwide Harmonized Light Duty Vehicle Test Procedure (dt.: weltweit einheitliches Leichtfahrzeuge-Testverfahren) |
| XOR | Ausschließende Disjunktion |

# Formelzeichenverzeichnis

### Chemische Formelzeichen

| Formel | Bezeichnung |
|---|---|
| C | Kohlenstoff |
| CH | Methingruppe |
| CN | Cyanid-Anion |
| CO | Kohlenstoffmonoxid |
| $CO_2$ | Kohlenstoffdioxid |
| $C_xH_yS_qO_z$ | Allgemeiner Ottokraftstoff |
| H | Atomarer Wasserstoff |
| $H_2$ | Molekularer Wasserstoff |
| $H_2O$ | Wasser |
| HC | Allgemeine Kohlenwasserstoffe |
| HCN | Cyanwasserstoff |
| $HO_2$ | Hydroperoxylradikal |
| N | Atomarer Stickstoff |
| $N_2$ | Molekularer Stickstoff |
| $N_2O$ | Distickstoffmonoxid |
| $N_2O_3$ | Distickstofftrioxid |
| $N_2O_4$ | Distickstofftetroxid |
| $N_2O_5$ | Distickstoffpentoxid |
| NCN | Cyanonitren-Radikal |
| NO | Stickstoffmonoxid |
| $NO_2$ | Stickstoffdioxid |
| $NO_3$ | Nitrat-Anion |
| $NO_x$ | Stickstoffoxide |
| O | Atomarer Sauerstoff |
| $O_2$ | Molekularer Sauerstoff |
| OH | Hydroxidion |
| $SO_2$ | Schwefeldioxid |

**Griechische Formelzeichen**

| Zeichen | Einheit | Bezeichnung |
|---|---|---|
| $\Delta$ | - | Differenz |
| $\alpha$ | ° | Winkel |
| $\eta$ | - | Wirkungsgrad |
| $\theta$ | - | Modellattribute/-parameter |
| $\lambda$ | - | Verbrennungsluftverhältnis |
| $\varphi$ | ° | Kurbelwinkel |
| $\Phi$ | - | Äquivalenzverhältnis |

**Lateinische Formelzeichen**

| Zeichen | Einheit | Bezeichnung |
|---|---|---|
| $\hat{f}$ | - | Prädiktion |
| $\dot{m}$ | kg/h | Massenstrom |
| A | - | Aktion |
| C | - | Damköhler Konstante |
| $c_R$ | - | Konstante in der Berechnung der laminaren Flammengeschwindigkeit |
| d | - | Differenz |
| E | J | Energie |
| e | - | Eulersche Zahl |
| f | - | Stoffmengenanteil |
| $f_R$ | - | Stöchiometrischer Restgasgehalt |
| g | - | Modellfunktion |
| H | MJ/kg | Heizwert |
| h | J | Enthalpie |
| H | - | Interner Zustand |
| h | - | Hypothese |
| k | (l/mol)^(n-1)*1/s | Geschwindigkeitskonstante |
| m | kg | Masse |
| m | - | Breite der Schichten |
| max | - | Maximalwert |
| Md | Nm | Drehmoment |
| min | - | Minimalwert |

| Zeichen | Einheit | Bezeichnung |
|---|---|---|
| n | - | Konstante in der Berechnung der turbulenten Flammengeschwindigkeit |
| n | $min^{-1}$ | Drehzahl |
| n | - | Anzahl der Schichten |
| nMax | - | Maximaler Reskalierungsfaktor |
| nMin | - | Minimaler Reskalierungsfaktor |
| p | $N/m^2$ | Druck |
| Pe | - | Péclet-Zahl |
| pGrad | $N/m^2/°KW$ | Druckgradient |
| Q | J | Wärme |
| R | J/(kg*K) | Universelle Gaskonstante |
| R | - | Belohnung |
| s | m/s | Flammengeschwindigkeit |
| S | - | Modellzustand |
| t | s | Zeit |
| T | K | Temperatur |
| U | J | Innere Energie |
| u‘ | - | Turbulente Geschwindigkeitsschwankung |
| V | $m^3$ | Volumen |
| W | J | Arbeit |
| w | - | Gewichtung |
| x/X | - | Modelleingangswerte |
| y/Y | - | Modellausgangswerte |

**Indizes**

| Index | Bezeichnung |
|---|---|
| 0 | Referenzbedingung |
| A | Ausströmend |
| a | Äußere |
| B | Kraftstoff |
| BB | Blowby |
| Br. | Brennstoff |
| E | Einströmend |

| **Index** | **Bezeichnung** |
|---|---|
| f | Final |
| hidden | Verdeckt |
| i | Laufvariable |
| in | Eingang |
| l | Laminar |
| l | Rückreaktion (links) |
| Leck | Leckage |
| max | Maximum |
| min | Minimum |
| out | Ausgang |
| r | Hinreaktion (rechts) |
| t | Zeitpunkt t |
| t | Turbulent |
| u | Unterer |
| verd. | Verdampft |
| W | Wand |
| $\alpha$ | Exponent 1 der laminaren Flammengeschwindigkeit |
| $\beta$ | Exponent 2 der laminaren Flammengeschwindigkeit |

# 1 Einleitung

Das Abkommen von Paris, das 2015 während der Klimarahmenkonvention der Vereinten Nationen (UNFCCC) verabschiedet wurde, zielt darauf ab, die globale Erwärmung zu begrenzen und enthält Bestimmungen, Ziele zur Reduzierung von Treibhausgasemissionen aus anthropogenen Quellen (einschließlich des Verkehrssektors) festzulegen und folglich die Entwicklung und Nutzung emissionsarmer Fahrzeuge zu fördern [1]. Neben diesem globalen Abkommen gibt es zahlreiche weitere Bemühungen, die Treibhausgasemissionen zu reduzieren und definierte Klimaziele zu erreichen. Beispielsweise hat die Europäische Union das spezifische Ziel festgelegt, die Emissionen im Verkehrssektor bis 2050 um 90 % im Vergleich zu den Werten von 1990 zu reduzieren [2]. Daher verlagert sich der Fokus zunehmend auf emissionsfreie Antriebe, wie etwa die batterieelektrischen Fahrzeuge (BEVs), welche aufgrund ihrer hohen Gesamteffizienz als vielversprechende Technologie angesehen werden.

Auch für Fahrzeuge mit Verbrennungsmotoren wurden in den letzten zwei Jahrzehnten massive Anstrengungen unternommen, um sowohl die Effizienz als auch die Schadstoffemissionen zu verbessern. Ein Teil dieser Forschung bestand darin, Fahrzeugbetriebszustände zu identifizieren, die zu erhöhten Emissionen führen, was beispielsweise beim Betrieb mit niedrigen Geschwindigkeiten, insbesondere beim Anfahren aus dem Stillstand, der Fall ist. Daraus konnten geeignete Betriebsstrategien entwickelt werden, um sowohl Schadstoff- als auch $CO_2$-Emissionen zu reduzieren, z. B. durch den Betrieb des Motors bei optimaler Last und die Unterstützung des Verbrennungsmotors bei niedrigen Geschwindigkeiten durch den Einsatz eines Elektromotors. Diese Erkenntnisse beschleunigten die Entwicklung von Hybridfahrzeugen unterschiedlichster Antriebskonfiguration, welche auf dem Pfad zur Elektromobilität im Privatkraftwagensektor als wichtige Brückentechnologie betrachtet werden [3].

Hybridfahrzeuge können bei entsprechender Systemauslegung lokal emissionsfrei fahren, Energie beim Bremsen rekuperieren und gleichzeitig weiterhin das bestehende, weitverbreitete Tankstellennetz nutzen. Eine Betriebsstrategie ist erforderlich, um die Interaktion zwischen Verbrennungsmotor und Elektromotor in jeder Fahrsituation zu optimieren, wobei viele Parameter (Route, Verkehr, Fahrer, Wetter, Ladezustand, etc.) berücksichtigt werden müssen. Die Erstellung und Optimierung einer Betriebsstrategie erfordern einen hohen Entwicklungsaufwand, einschließlich zahlreicher praktischer Tests und Fachwissen in verschiedenen Bereichen.

Um eine effiziente Betriebsstrategie zu erstellen, müssen zunächst ein oder mehrere Ziele definiert werden, die als Bewertungsmaßstab für den Fortschritt der Optimierung dienen können. Der Fokus auf der Minimierung des Gesamtenergieverbrauchs war lange Zeit das bevorzugte Entwicklungsziel von Hybridfahrzeugen. Dies kann erreicht werden, indem der Verbrennungsmotor in hocheffizienten Kennfeldbereichen betrieben wird und ungünstige Lastszenarien (z. B. sehr niedrige Lasten) gezielt vermieden werden. Im Ergebnis führt dies zu einer

Verringerung der $CO_2$-Emissionen aufgrund des geringeren Kraftstoffverbrauchs, ignoriert jedoch andere Emissionskomponenten (z. B. Stickstoffoxide). Durch die aktuell herrschenden Abgasgesetzgebungen ist es zusätzlich sinnvoll, die Reduktion unterschiedlicher Abgaskomponenten als Optimierungsziel zu verfolgen. Dies kann neben der Einhaltung gesetzlicher Grenzwerte auch weitere Vorteile bieten, zum Beispiel durch die Möglichkeit das Abgasnachbehandlungssystem zu vereinfachen und somit die Systemkosten zu reduzieren.

Optimierungen können vor allem in realen Versuchsumgebungen sehr zeitaufwändig sein. Begünstigt durch die steigende Rechenleistung und neue Modellierungsmethoden, kann der Einsatz eines digitalen Zwillings daher ein effiziente Alternative darstellen. Hierzu ist eine umfassende Simulationsumgebung erforderlich, die in der Lage sein muss, die realen Betriebsbedingungen mit hoher Genauigkeit zu modellieren. Neben den Fahrzeug- und Umweltmodellen gehören dazu auch spezielle Modelle, um das zu optimierende Verhalten, wie beispielsweise die hochdynamische Emissionsentstehung eines Verbrennungsmotors, zu beschreiben.

Damit der Emissionsausstoß und der Kraftstoffverbrauch von Fahrzeugen unter realistischen Betriebsbedingungen bewertet werden können, wurden verschiedene Fahrzyklen als standardisierte Metrik entwickelt. In Europa wurden 2017 die Worldwide Harmonized Light Duty Vehicle Test Procedure (WLTP) und der damit verbundene Worldwide Harmonized Light Duty Vehicle Test Cycle (WLTC) eingeführt, um realistischere Fahrbedingungen abzubilden als dies der zuvor verwendete Neue Europäische Fahrzyklus (NEFZ) leisten konnte [4]. Als weiteren Schritt verlangt die europäische Emissionsgesetzgebung zum heutigen Zeitpunkt die Bewertung der Schadstoffemissionen im realen Straßenbetrieb gemäß der Real Driving Emissions (RDE)-Gesetzgebung. Diese Rahmenbedingungen erhöhen die Anforderungen an Motor- und Emissionsmodelle, da hochkomplexe Prozesse (Emissionsbildung) in schnell veränderlichen Umgebungen effizient abgebildet werden müssen. Letzteres ist essenziell, um die Modelle auch auf leistungsarmer Hardware, wie beispielsweise Steuergeräten, für unmittelbare Regelungsfunktionen einsetzen zu können.

Zur Modellierung des Verbrennungsmotors und der Emissionsbildung wurden in der Vergangenheit spezifische Methoden, wie physikalisch oder phänomenologisch basierte Ansätze, entwickelt. Obwohl diese Modelle die Emissionen unter stationären Bedingungen präzise vorhersagen können, ist ihre Optimierung für den hochdynamischen Motorbetrieb zeitaufwendig, erfordert ein hohes Maß an Expertise und ist oft mit einer reduzierten Genauigkeit verbunden.

Das zentrale Ziel der vorliegenden Arbeit besteht darin, aktuelle Fortschritte in den Methoden des Maschinellen Lernens zu analysieren und auf die beschriebene Problemstellung anzuwenden. Dadurch sollen insbesondere die bisherigen Limitierungen der dynamischen Emissionsprädiktion überwunden werden. Zusätzlich werden Kombinationen von physikalisch-phänomenologischen und datenbasierten Ansätzen betrachtet (sogenannte Hybridmodelle), um mögliche Vorteile durch die Verknüpfung beider Modellarten zu untersuchen.

# 2 Theoretische Grundlagen

Um die in dieser Arbeit vorgestellten Modelle und Ergebnisse fundiert bewerten zu können, ist es unerlässlich, die Hintergründe und verschiedenen theoretischen Grundlagen zu berücksichtigen. Daher werden nachfolgend zunächst Kenntnisse über die ottomotorische Verbrennung und die physikalisch-phänomenologische Beschreibung dieser Prozesse vermittelt.

Danach wird die Entstehung verschiedener, im Kontext dieser Arbeit relevanter Emissionskomponenten erörtert. Zusätzlich werden physikalisch-phänomenologische Modellierungsansätze dargestellt, welche für die jeweiligen Emissionskomponenten in der Entwicklung von hoher Relevanz sind und ebenfalls in den vorliegenden Untersuchungen angewendet werden.

Abschließend folgt eine Einführung in die datenbasierten Methoden und das Maschinelle Lernen. Durch die Anwendung Maschinellen Lernens auf die komplexen Datenmengen, die aus Experimenten und Simulationen gewonnen werden, können Muster erkannt und Methoden entwickelt werden, die neue und effiziente Emissionsmodelle ermöglichen.

## 2.1 Ottomotor

Der Ottomotor gehört zu den Verbrennungsmotoren, welche sich als Wärmekraftmaschinen dadurch charakterisieren lassen, dass eine Umwandlung von chemisch gebundener in mechanische Energie erfolgt [5], [6], [7]. Die im Kontext dieser Arbeit relevanten Grundlagen werden nachfolgend erörtert.

### 2.1.1 Grundlagen

Das klassische homogene ottomotorische Brennverfahren zeichnet sich nach Merker und Teichmann [8] durch eine zeitliche Abgrenzung von Gemischbildung und Verbrennung aus. Diese vorgemischte Verbrennung ist eine Hauptunterscheidung zu der beim Dieselbrennverfahren vorherrschenden Diffusionsflamme, welche sich nach Heywood [7] dadurch auszeichnet, dass die Durchmischung der Reaktionspartner erst in der Reaktionszone stattfindet.

Beim ottomotorischen Brennverfahren ist der Kraftstoff hingegen bereits vor Zünd- bzw. Verbrennungsbeginn vollständig verdampft. Das Luft-Kraftstoff-Gemisch wird am Ende der Kompressionsphase (in der Regel) durch einen Zündfunken fremdgezündet, da Ottokraftstoff verglichen mit anderen Kraftstoffen (beispielsweise dem Dieselkraftstoff) relativ zündunwillig ist. [8]

Bei der Gemischbildung in Ottomotoren gibt es Unterschiede hinsichtlich der Durchmischung der Komponenten im Brennraum (homogen/inhomogen) und dem Masseverhältnis zwischen Luft und Kraftstoff. Dies wird durch das Vorliegen eines stöchiometrischen Luft-Kraftstoffverhältnisses, bei dem exakt die Luftmasse zur Verfügung steht, welche für die vollständige Verbrennung des Kraftstoffes theoretisch erforderlich ist, oder eines überstöchiometrischen Verhältnisses charakterisiert. Diese Systematik ist in Abbildung 2-1 nach [8] abgebildet.

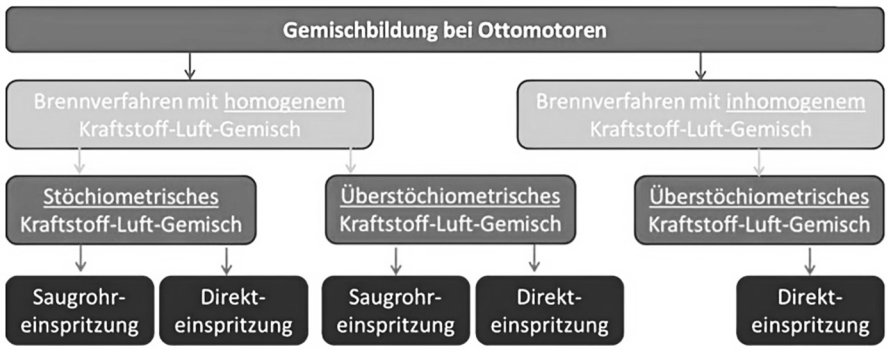


Abbildung 2-1: Systematik zur Unterteilung ottomotorischer Gemischbildungsverfahren nach [8]

Für die Aufrechterhaltung des Umwandlungsprozesses von chemisch gebundener in mechanische Energie muss periodisch frisches Luft-Kraftstoff-Gemisch im Arbeitszylinder bereitgestellt und nach der Verbrennung wieder ausgestoßen werden. Dies geschieht bei Verbrennungsmotoren in Personenkraftwagen meist nach dem Viertakt-Prinzip. Ein Arbeitsspiel besteht dabei aus vier Abschnitten. Im Ansaugtakt gelangt Luft beziehungsweise Luft-Kraftstoff-Gemisch über die Einlassventile in den Brennraum. Danach schließen die Einlassventile und durch eine Bewegung des Kolbens in Richtung oberem Totpunkt (OT) werden die eingeschlossenen Komponenten komprimiert. Durch die Zündkerze wird die Verbrennung vor Erreichen des OT eingeleitet. Die Freisetzung der chemischen Energie des Kraftstoffes führt zu einer Druck- und Temperaturerhöhung im Brennraum, was den Kolben nach unten bewegt. Bei der anschließenden Aufwärtsbewegung des Kolbens werden die Abgase und unter Umständen vorhandenes unverbranntes Gemisch über die Auslassventile ausgestoßen und der Prozess beginnt von neuem. [9]

Infolge der verschiedenen Ausführungen der Ottomotoren und der dazugehörigen Brennverfahren gibt es im dargestellten Viertakt-Prozess Unterschiede in den einzelnen Abschnitten. So wird beispielsweise während der Kompression bei einem Motor mit Saugrohreinspritzung

bereits Luft-Kraftstoff-Gemisch verdichtet, wohingegen bei einem Motor mit Direkteinspritzung und inhomogenem Luft-Kraftstoff-Gemisch (Schichtladung) anfangs nur Luft komprimiert und erst am Ende des Kompressionstaktes Kraftstoff eingespritzt wird. In Abbildung 2-2 wird der Viertakt-Prozess eines Ottomotors mit Saugrohreinspritzung dargestellt.

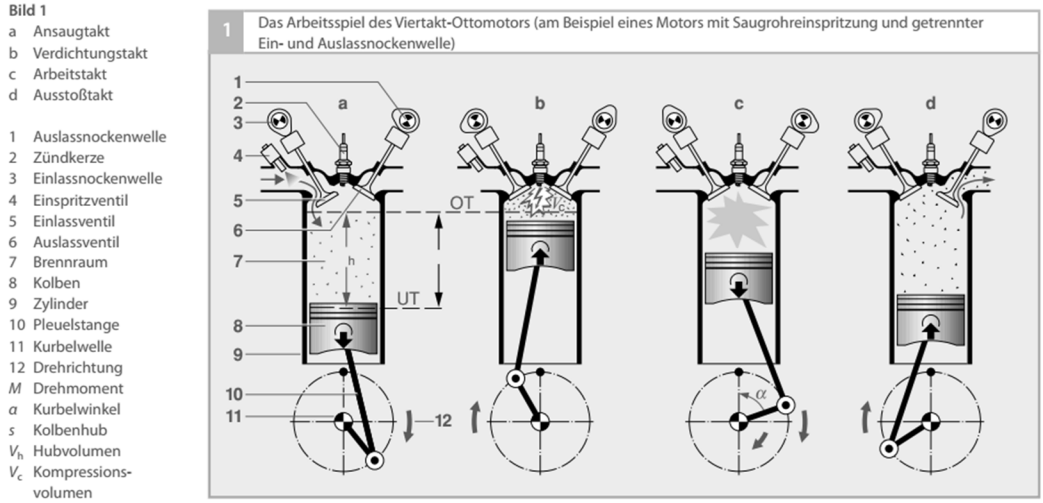


Abbildung 2-2: Arbeitsspiel eines Ottomotors mit Saugrohreinspritzung nach [9]

Auch hinsichtlich der Art der Zündung gibt es verschiedene Varianten. Die herkömmliche Spulenzündung mit einer Hochspannungsquelle (Zündspule) und einer Zündkerze, an welcher der Zündfunke überspringt und dadurch die notwendigen Bedingungen für eine exotherm ablaufende Reaktion schafft, hat sich aufgrund der guten Funktion und vergleichbar geringem Kostenaufwand etabliert [10]. Davon abgesehen gibt es weitere alternative Zündsysteme, welche jedoch – wenn überhaupt – nur in Nischen Anwendung finden (beispielsweise die Plasma- oder Laserzündung) [11].

Neben den zuvor genannten Möglichkeiten, die sich alle unter dem Begriff der Fremdzündung zusammenfassen lassen, gibt es auch beim Ottomotor Konzepte, bei welchen das Gemisch durch die Beeinflussung der Randbedingungen, wie etwa Brennraumtemperatur und -druck, zur Selbstzündung gebracht wird. Diese Form von Niedertemperatur-Brennverfahren werden oft unter dem Begriff „HCCI“ (engl. „Homogeneous Charge Compression Ignition“) zusammengefasst. HCCI ist im deutschsprachigen Raum auch als „homogene Kompressionszündung“ geläufig und bietet die Möglichkeit einer gleichzeitigen Effizienzsteigerung bei reduzierten Schadstoff-Emissionen [12]. Da der Verbrennungsbeginn nicht wie bei den fremdgezündeten Brennverfahren präzise eingeleitet werden kann, liegt darin und hieraus resultierend in der Kontrolle des weiteren Verbrennungsverlaufes eine große Herausforderung.

Wie im Verlauf von Kapitel 5.1.1 näher erläutert wird, handelt es sich beim für diese Arbeit relevanten Versuchsträger um einen Motor mit Direkteinspritzung und einem homogenen,

fremdgezündeten Luft-Kraftstoff-Gemisch, welches in weiten Kennfeldbereichen stöchiometrisch vorliegt. Daher bezieht sich die weitere Aufbereitung der theoretischen Grundlagen (bspw. der Herleitung der Emissionsentstehung) hauptsächlich auf dieses ottomotorische Konzept.

### 2.1.2 Druckverlaufsanalyse

Eine wichtige Grundlage für die spätere Modellierung bzw. Motorprozessrechnung bildet die Druckverlaufsanalyse. Dadurch können solche Größen bestimmt werden, die im Experiment nur schwer oder nicht direkt messbar sind. Hierzu zählen beispielsweise der Brennverlauf oder der Restgasgehalt. In Abbildung 2-3 sind einige charakteristische Größen aufgeführt, welche aus der Analyse des Druckverlaufes folgen.

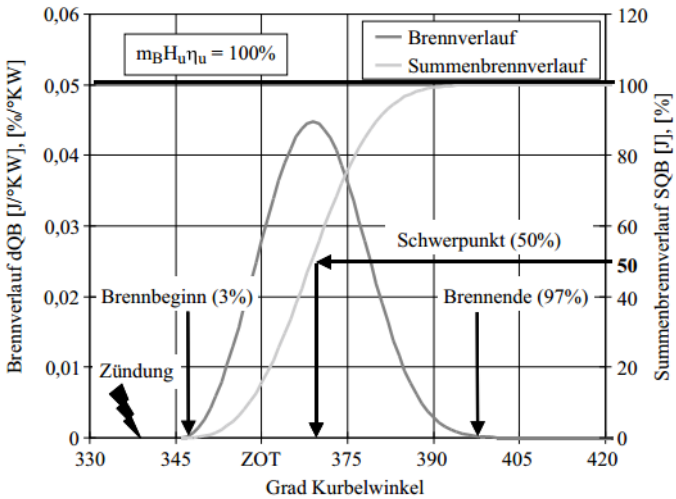


Abbildung 2-3: Beispielhaftes Ergebnis einer Druckverlaufsanalyse [13]

Bei der Bestimmung des Druckverlaufes wird häufig in einen Hochdruck- (Brennraum) und einen Niederdruckbereich (Einlasssystem und Auslasssystem) unterschieden, wobei die zusätzliche Betrachtung des Niederdruckteils den Vorteil bietet, dass genauere Angaben über den Ladungswechselvorgang gemacht werden können. In beiden Fällen wird das erfasste Drucksignal in Korrelation zur Kurbelwellenposition (über den sogenannten Kurbelwinkel) aufgezeichnet, was auch als Druckindizierung bezeichnet wird. Für eine detaillierte Erklärung der Indiziermesskette und der technischen Umsetzung der einzelnen Komponenten kann auf das Werk von Wimmer und Glaser [14] zurückgegriffen werden.

Da der Schwerpunkt der Druckverlaufsanalyse hauptsächlich auf der Verbrennung liegt, wird der Brennraum thermodynamisch in der Regel als geschlossenes System während der Hochdruckschleife betrachtet. Dies bedeutet, dass es keine Enthalpieströme über die Systemgrenze hinweg gibt, wodurch folgender Zusammenhang aufgestellt werden kann: [13]

$$\frac{dQ_B}{dt} = \frac{dU}{dt} + \frac{dQ_W}{dt} + p\frac{dV}{dt}\left[-\frac{dm_{BB}}{dt}h_{BB}\left(-\frac{dm_{Br.,verd.}}{dt}\Delta h_{verd.}\right)\right] \qquad (1)$$

Dabei sind $\frac{dQ_B}{dt}$ der (zeitliche) Brennverlauf, $U$ die innere Energie und $Q_W$ die Wandwärmeverluste. $p\frac{dV}{dt}$ beschreibt die Volumenänderungsarbeit, $m_{BB}$ die Masse des Blowby (Verbrennungsgase, die zwischen Kolben und Zylinderwand in das Kurbelgehäuse entweichen), $h_{BB}$ die dazugehörige Enthalpie, $dm_{Br.,verd.}$ die Masse des verdampften Brennstoffs und $h_{verd.}$ die Verdampfungsenthalpie. In erster Näherung können bei einem Ottomotor mit Direkteinspritzung die Verdampfungsenthalpie und die Blowby-Verluste vernachlässigt werden. [13]

Alle Größen sind in der oben dargestellten Gleichung zeitbezogen, ein Bezug auf die Kurbelwellenposition lässt sich äquivalent umsetzen. Durch Messungen, geometrische Beziehungen und Annahmen (beispielsweise des Polytropenexponenten) kann aus (1) und dem gemessenen Druckverlauf auf den Brennverlauf geschlossen werden. Eine detaillierte Anwendung dieser Methode ist unter anderem in der Arbeit von Witt [15] beschrieben.

### 2.1.3 Nulldimensionale und phänomenologische Simulation der Verbrennung

Für die Simulation von Verbrennungsmotoren gibt es verschiedene Modellarten, welche sich anhand unterschiedlicher Kriterien klassifizieren lassen. In Bezug auf die Dimensionalität ist eine Unterscheidung in 0D-, 1D-, 3D- und quasidimensionale Modelle möglich. Innerhalb dieser Einteilung kann eine weitere Aufschlüsselung in die mathematischen, empirischen, physikalischen und phänomenologischen Modelle vorgenommen werden. [16]

Bei der Auswahl der Modellklasse gibt es keine generell zu bevorzugende Lösung, vielmehr sollte der Modellansatz in Bezug auf den Zweck der Simulation gewählt werden. 0-dimensionale Modelle stellen die einfachste Art der Verbrennungssimulation dar, was mit der Notwendigkeit vieler Vereinfachungen, aber ebenso einer sehr schnellen Berechnungszeit einhergeht. Am anderen Ende des Spektrums lassen sich die 3D- beziehungsweise CFD-Modelle (engl. „Computational Fluid Dynamics“) einordnen, welche den Brennraum räumlich in kleine Volumen-Elemente auflösen und dadurch eine mehrdimensionale Erfassung der Vorgänge erlauben. Dies ermöglicht zwar eine hohe Genauigkeit, impliziert jedoch in der Regel auch eine lange Simulationsdauer. Die erörterten Zusammenhänge sind in Abbildung 2-4 qualitativ dargestellt.

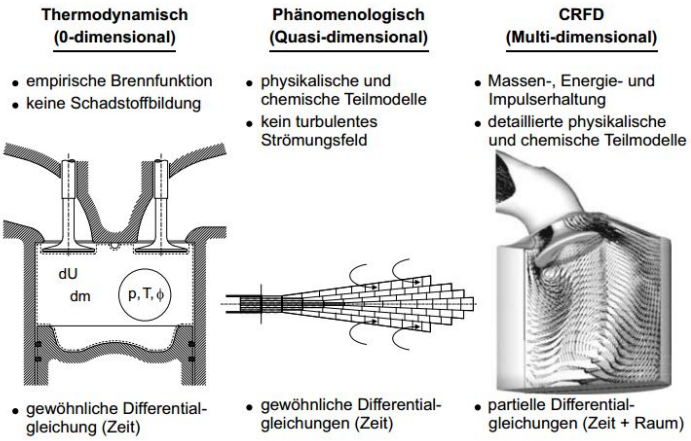


Abbildung 2-4: Einordnung von Modellarten nach Komplexität und Rechendauer [13]

3D-CFD-Simulationen können durch die präzise Auflösung des Strömungsfeldes bereits in sehr frühen Entwicklungsstadien eine Einschätzung bezüglich des späteren Systemverhaltens ermöglichen. Dies kann beispielsweise bei der geometrischen Auslegung des Luftpfades für einen Verbrennungsmotor (Ansaugsystem, Ventile, etc.) hilfreich sein. Dieser lässt sich dann bereits hinsichtlich der Kanalführung optimieren, ohne dass ein physischer Prototyp existiert. Je nach betrachtetem Gesamtvolumen und der Zellgröße erfordert diese Untersuchung jedoch eine hohe Rechenleistung und die Rechenzeit beträgt in der Regel ein Vielfaches der simulierten Dauer. 0D-, 1D-, oder quasidimensionale Modelle können dahingegen die tatsächliche Systemgeometrie nicht oder nur stark eingeschränkt darstellen, da bei der Nutzung dieser Ansätze zahlreiche Vereinfachungen notwendig sind. Der große Vorteil dieser Modellansätze liegt jedoch in der hohen Rechengeschwindigkeit, welche auch Echtzeitfähigkeit ermöglichen kann. Dies ist nicht nur im Hinblick auf stationäre Hardware interessant, sondern besonders auch für die Anwendung in Fahrzeugen mit in der Regel weniger performanter mobiler Hardware.

Da ein Ziel der Arbeit in der Entwicklung eines effektiven Werkzeuges für die Optimierung und Regelung der Emissionen liegt, werden die theoretischen Grundlagen für die 0D- und phänomenologischen Modelle im weiteren Verlauf näher erläutert. Diese Modellarten weisen für den gewählten Einsatzzweck das größte Potenzial auf.

**Nulldimensionale Modellierung**

Bei der 0-dimensionalen Modellierung wird der Brennraum nach bestimmten statistischen Kriterien in eine oder mehrere Zonen unterteilt. Innerhalb dieser Zonen gibt es räumlich gesehen keine Unterschiede und jede Zone wird als homogen betrachtet. Dadurch werden alle Größen innerhalb einer Zone auf eine zeitliche oder kurbelwinkelbasierte Abhängigkeit bezogen, weshalb die 0-dimensionale Modellierung auch öfters als zeitdimensionale Modellierung bezeichnet wird. [17]

Neben dieser zwingenden Voraussetzung für die 0-dimensionale Modellierung gibt es nach [17] noch drei weitere Annahmen, welche oft als Vereinfachung der Berechnung getroffen werden:

- *„Das Arbeitsgas im Brennraum wird als Gemisch idealer Gase behandelt, dessen Komponenten Luft, verbranntes Gas und bei Gemischansaugung Kraftstoffdampf zu jedem Zeitpunkt als vollständig durchmischt angenommen werden.“*
- *„Reibungskräfte im Arbeitsgas werden vernachlässigt, so dass mit der Voraussetzung konstanten Drucks innerhalb jeder Zone der Impulssatz keine Aussage liefert.“*
- *„Die Verbrennung wird im Energieerhaltungssatz durch die Zufuhr der Brennstoffwärme $dQ_B$ dargestellt, die der Energiefreisetzung des chemisch reagierenden Kraftstoffs entspricht, ohne dass Kraftstoffaufbereitung, Verdampfung oder Zündverzug separat modelliert werden.“*

Der Brennraum wird als ein offenes System angenommen, in welchem alle Größen instationär veränderlich sind. Dies lässt sich sowohl räumlich wie auch zeitlich beobachten. Die Vorgänge, welche die Energiebilanz des Systems Brennraum beeinflussen, sind in Abbildung 2-5 dargestellt. [17]

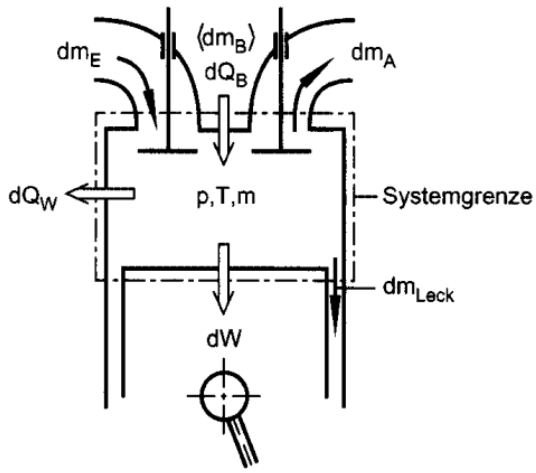


Abbildung 2-5: Brennraum als offenes System nach [17]

Zusammengefasst lassen sich nach [17] die folgenden thermodynamischen Prozesse im System Brennraum beschreiben:

- Stofftransport: Es findet ein Transport der einströmenden Gasmasse ($dm_E$), der ausströmenden Gasmasse ($dm_A$), der Leckagemasse ($dm_{Leck}$) und (bei Direkteinspritzung) der Brennstoffmasse ($dm_B$) über die Systemgrenze statt.
- Energietransport: Durch die Umwandlung der im Kraftstoff gebundenen chemischen Energie wird Wärme ($dQ_B$) freigesetzt. Die Wandwärme ($dQ_W$) und Arbeit ($dW$) werden über die Systemgrenze abgegeben. Zusätzlich sind die Massenströme am Energietransport durch ihre Enthalpie und äußere Energie beteiligt.
- Energieänderung: Die innere, im System gespeicherte Energie ($dU$) und die äußere Energie ($dE_a$) ändern sich während eines Arbeitsspiels.

Das Einzonenmodell erlaubt eine globale Beurteilung der Motorprozesse ohne eine Auflösung lokaler Unterschiede (beispielsweise der Temperaturverteilung) zu berücksichtigen. Wenn dies die Anforderungen an die Simulation erfüllt, so können Einzonenmodelle besonders für luftansaugende Motoren gute Ergebnisse liefern. Zwei- bzw. Mehrzonenmodelle unterteilen den Brennraum in der Regel in je eine Zone mit verbranntem und unverbranntem Gemisch. Dadurch können diese Modelle vorteilhaft bei der Betrachtung von gemischansaugenden Motoren sein, oder wenn die Temperaturverteilung im Brennraum berücksichtigt werden muss, beispielsweise bei der Modellierung von Emissionen. [17]

Nachfolgend werden drei Grundgleichungen für die 0-dimensionale Simulation von Verbrennungsmotoren nach [17] erläutert, wobei dies zum besseren Verständnis anhand des Einzonenmodells für luftansaugende Motoren geschieht:

**Massenerhaltung:** Basierend auf der allgemeinen Kontinuitätsgleichung kann die Änderung der Masse in einer Zone mit der Summe der Massenströme gleichgesetzt werden, welche in diese zu- oder daraus abfließen. Das lässt sich differentiell mit Bezug auf den Kurbelwinkel φ mit folgender Gleichung darstellen.

$$\frac{dm}{d\varphi} = \frac{dm_E}{d\varphi} - \frac{dm_A}{d\varphi} - \frac{dm_{Leck}}{d\varphi} + \frac{dm_B}{d\varphi} \tag{2}$$

Die darin beschriebenen differentiellen Masseänderungen lassen sich in Abbildung 2-5 grafisch nachvollziehen.

**Energieerhaltung:** Für die Herleitung der Energieerhaltung im betrachteten System ist der 1. Hauptsatz der Thermodynamik für instationäre offene Systeme (Gleichung 3) grundlegend von Bedeutung:

$$dW_t + dQ_a + \sum dm_i * (h_i \ + \ e_{ai}) = dU + dE_a \tag{3}$$

Auf der linken Seite der Gleichung werden die Arbeit, Wärme und die durch Stofftransport zugeführte Energie summiert. Dabei ist $W_t$ die technische Arbeit, welche die Systemgrenze passiert; $Q_a$ die äußere Wärme, welche die Systemgrenze überschreitet; $m_i$ beschreibt die Masse, welche über die Systemgrenze fließt und in Bezug auf $m_i$ ist $h_i$ die dazugehörige spezifische Enthalpie, $e_{ai}$ stellt die spezifische äußere Energie dar. Der rechte Term von Gleichung 3 beschreibt die im System gespeicherte innere ($U$) und äußere Energie ($E_a$).

Diese Gleichung lässt sich auf den Brennraum als offenes System unter Beachtung der am Kapitelbeginn getroffenen Annahmen, des gewählten Bilanzraumes (siehe Abbildung 2-5) und der erneuten Bezugnahme auf den Kurbelwinkel $\varphi$ wie folgt formulieren:

$$-p\frac{dV}{d\varphi} + \frac{dQ_B}{d\varphi} - \frac{dQ_W}{d\varphi} + h_E\frac{dm_E}{d\varphi} - h_A\frac{dm_A}{d\varphi} - h_A\frac{dm_{Leck}}{d\varphi} = \frac{dU}{d\varphi} \tag{4}$$

$p\frac{dV}{d\varphi}$ beschreibt dabei die Volumenänderungsarbeit und wird aus der Multiplikation des Zylinderdrucks und der Zylindervolumenänderung berechnet. Der Brennverlauf $\frac{dQ_B}{d\varphi}$ und die Wandwärmeverluste $\frac{dQ_W}{d\varphi}$ bilden zusammen den Heizverlauf ab. $h_E\frac{dm_E}{d\varphi}$, $h_A\frac{dm_A}{d\varphi}$ und $h_A\frac{dm_{Leck}}{d\varphi}$ sind die Enthalpieströme der einströmenden-, der ausströmenden und der Leckagemasse. Diese Terme werden der Änderung der inneren Energie im betrachteten System (Brennraum) $\frac{dU}{d\varphi}$ gleichgesetzt.

**Zustandsgleichung:** Wie bereits erwähnt, wird das Arbeitsgas im Brennraum als Gemisch idealer Gase betrachtet. Diese können allgemein anhand der thermischen Zustandsgleichung idealer Gase beschrieben werden, welche in verschiedenen, äquivalenten Formen aufgestellt werden kann. Für die weitere Herleitung der Zustandsgleichung für das Einzonenmodell luftansaugender Motoren wird die folgende extensive Form verwendet:

$$p * V = m * R * T \tag{5}$$

Der Zustand des Gases wird in (5) durch den Druck $p$, das Volumen $V$, die Masse $m$, der Gaskonstante $R$ (wobei in diesem Fall die spezifische Gaskonstante genutzt wird) und die Temperatur $T$ bestimmt.

Erneut wird diese Gleichung nach dem Kurbelwinkel $\varphi$ abgeleitet und es muss berücksichtigt werden, dass es sich über ein Arbeitsspiel (zumindest abschnittsweise) um ein Gasgemisch handelt, dessen Gaskonstante $R$ von der momentanen Gemischzusammensetzung abhängt.

Dadurch folgt:

$$p\frac{dV}{d\varphi} + V\frac{dp}{d\varphi} = mR\frac{dT}{d\varphi} + mT\frac{dR}{d\varphi} + RT\frac{dm}{d\varphi} \tag{6}$$

Mit den hergeleiteten Gleichungen kann das System Brennraum berechnet werden, wobei Annahmen für eventuell fehlende Größen getroffen werden müssen. Außerdem sind für eine eindeutige Lösung, da es sich um eine Differentialgleichung handelt, Anfangsbedingungen erforderlich. Dies kann beispielsweise dadurch erfüllt sein, dass zu einem Kurbelwinkel $\varphi$ die Werte aller Variablen vorliegen. Durch die Anzahl der Unbekannten in den oben genannten Gleichungen – wie etwa Druck- und Temperaturverläufe, die Zusammensetzung des Arbeitsgases oder der umgesetzten Brennstoffwärme – ist häufig vor der Simulation eine Analyse des realen Prozesses anhand des gemessenen Zylinderdruckverlaufes erforderlich, um fehlende Parameter (besonders den Brennverlauf) zu bestimmen. Dies wird in Kapitel 2.1.2 näher erläutert. [3]

Für die Bestimmung der in den Gleichungen (1)-(6) enthaltenen Größen stehen weitere Gleichungen zur Verfügung, in welchen erneute Annahmen bzw. Näherungen getroffen werden. Als weiterführende Literatur kann diesbezüglich auf die Lehrbücher von Pischinger et al. [17], Maurer [18] oder Merker und Teichmann [13] zurückgegriffen werden.

**Phänomenologische Modellierung**

Phänomenologische Motormodelle berechnen die Vorgänge im System Brennraum unter Berücksichtigung verschiedener chemischer und physikalischer Phänomene. Hierzu gehören unter anderem die Gemischbildung, die Zündung, die Reaktionskinetik oder die Strahlausbreitung beim Kraftstoffeintrag in den Brennraum. Für diese Ansätze ist in der Regel die statistische Aufteilung des Systems in mehrere Zonen erforderlich, welche sich hinsichtlich der Parameter Druck, Temperatur und stofflicher Zusammensetzung unterscheiden können. Aus diesem Grund werden phänomenologische Modelle als quasidimensional bezeichnet. Im Gegensatz zu den 3D-CFD-Modellen wird jedoch keine explizite räumlich-geometrische Auflösung des Rechengebiets angestrebt, um die Simulationszeit deutlich zu verringern. [13]

Für den Verbrennungsvorgang in Ottomotoren, welche homogen betrieben werden, werden bei der phänomenologischen Modellierung nach [13] in der Regel folgende Annahmen zur Vereinfachung getroffen:

- Kraftstoff, Luft und Restgas sind homogen vermischt.
- Das Volumen der Reaktionszone ist im Vergleich zum gesamten Brennraumvolumen sehr klein.
- Die Flamme selbst ist nach dem Flamelet-Ansatz (siehe Peters [19]) infinitesimal dünn.
- Der Brennraum ist in eine verbrannte und eine unverbrannte Zone unterteilt.

Für den Verlauf der Verbrennung ist die Umsatzrate des unverbrannten Gemisches entscheidend. Diese hängt mit der Ausbreitungsgeschwindigkeit der turbulenten Flammenfront zusammen, die ihren Ursprung an der Zündkerze hat und in Wandnähe erlischt. Zu den größten Herausforderungen bei der phänomenologisch basierten Modellierung von Ottomotoren gehört die Beschreibung des turbulenten Strömungsfeldes. Hierzu fehlen in der Regel relevante Parameter, wie etwa die turbulente Schwankungsgeschwindigkeit oder turbulente Längenskalen. Die Abschätzung dieser Parameter ist daher eine zentrale Aufgabe in der Modellierung, welcher sich zahlreiche Untersuchungen angenommen haben. [13]

So haben beispielsweise Bossung et al. [20] einen Ansatz vorgestellt, welcher basierend auf einem $\kappa - \varepsilon$ Turbulenzmodell nulldimensionale Turbulenzparameter berechnet, um den Brennverlauf für einen Ottomotor mit homogener Gemischverteilung prädiktiv bestimmen zu können.

Aufgrund der zentralen Bedeutung für die phänomenologische Modellierung werden nachfolgend Ansätze für die Beschreibung der laminaren und turbulenten Flammengeschwindigkeit dargestellt. Die laminare Flammengeschwindigkeit wird nach [13] dadurch gekennzeichnet, dass es sich um die Ausbreitungsgeschwindigkeit einer *„(…) vorgemischten Flammenfront in einem ruhenden Brennstoff-Luft-Gemisch“* handelt. Für die Bestimmung dieser Geschwindigkeit stehen verschiedene Möglichkeiten zur Verfügung. Eine empirische Beziehung für die laminare Flammengeschwindigkeit $s_l$ wurde bereits 1982 von Metghalchi und Keck [21] aufgestellt:

$$s_l = s_{l,0} * \left(\frac{T_u}{T_0}\right)^{\alpha} * \left(\frac{p}{p_0}\right)^{\beta} * (1 - c_R * f_R) \tag{7}$$

Dabei wird die Abhängigkeit der laminaren Flammengeschwindigkeit von der Temperatur und dem Druck unter Berücksichtigung von Referenzbedingungen ($T_0$= 298 K; $p_0$= 100 kPa) dargestellt. $f_R$ ist der stöchiometrische Restgasgehalt, $c_R$ eine Konstante, für welche in der Literatur unterschiedliche Angaben gemacht werden. Ursprünglich wurde diese von Metghalchi und Keck [16] mit 2.1 angenommen, Wallesten [22] nimmt circa 20 Jahre später hingegen $c_R = 3$ an.

Da es sich um einen empirischen Zusammenhang handelt, müssen die Randbedingungen beziehungsweise der Gültigkeitsbereich definiert werden. Exemplarisch sind in Tabelle 2-1 die Berechnungsformeln von $\alpha$, $\beta$ für Isooktan als Brennstoff und die dazugehörigen Randbedingungen nach [21] angegeben.

Tabelle 2-1: Parameter für die Bestimmung der laminaren Flammengeschwindigkeit nach [21]

| Exponenten | | Randbedingungen | | |
|---|---|---|---|---|
| α | β | $\Phi = 1/\lambda$ | T | p |
| 2.18–0.8(1/ λ-1) | -0.16+0.22(1/ λ-1) | 0.8-1.2 | 298-700 K | 0.4-50 bar |

Da im realen Brennraum die Strömung nicht als laminar angenommen werden kann, muss die Turbulenz bei der Modellierung der Flammenfront berücksichtigt werden. Zur Vereinfachung kann hierbei davon ausgegangen werden, dass die Reaktionsrate lokal gleich bleibt, jedoch die Flammenfront eine größere Oberfläche durch den Einfluss turbulenter Wirbel besitzt. Dies erhöht die Brenngeschwindigkeit und muss daher durch die Abbildung der turbulenten Flammengeschwindigkeit $s_t$ erfasst werden. [13]

Damköhler [23] hat bereits 1940 einen Ansatz zur Bestimmung der turbulenten Flammengeschwindigkeit $s_t$ auf der Grundlage von experimentellen Untersuchungen aufgestellt. Darauf aufbauend hat Peters [24] folgenden Zusammenhang abgeleitet:

$$s_t = s_l \left(1 + C \frac{u´}{s_l}\right)^n \tag{8}$$

$u´$ ist die turbulente Geschwindigkeitsschwankung, welche durch Messungen bestimmt wird. Die Damköhler-Konstante $C$ ist in erster Linie von der Flammendicke und der turbulenten Längenskala abhängig. $n$ ist eine Konstante, welche in der Literatur zwischen 0.5 und 1 liegt. [13]

Neben der Ausbreitung der Flammenfront gibt es noch zahlreiche weitere Mechanismen, die experimentell untersucht und im Rahmen der phänomenologischen Modellierung des Verbrennungsmotors genutzt werden. Dazu zählen beispielsweise die Wärmefreisetzung [25],[26], das Klopfverhalten [27], [28], [29] oder der Zündvorgang [30], [31]. Aufgrund der Breite der Thematik wird im Rahmen dieser Arbeit auf die weiterführende Literatur verwiesen.

## 2.2 Emissionsentstehung und physikalisch-phänomenologische Abbildung beim Ottomotor mit Direkteinspritzung

Das folgende Kapitel behandelt die theoretischen Grundlagen der Emissionsentstehung bei Ottomotoren mit Direkteinspritzung. Ein besonderes Augenmerk wird dabei auf Stickstoffoxide ($NO_x$), Kohlenstoffmonoxid ($CO$), Kohlenstoffdioxid ($CO_2$) und unverbrannte Kohlenwasserstoffe ($THC$) gelegt. Dies resultiert daraus, dass diese gasförmigen Abgaskomponenten folgende Voraussetzungen erfüllen:

- Sie sind für die betrachtete Art von Verbrennungsmotor relevant.
- Die vorhandene Messtechnik erlaubt eine präzise Erfassung der Komponenten im dynamischen Betrieb.
- Die verwendeten Simulationswerkzeuge ermöglichen eine effiziente physikalisch-phänomenologische Modellierung.

In der datenbasierten Modellierung bestehen bezüglich der letztgenannten Voraussetzung keinerlei Einschränkungen, da sich unter Einhaltung bestimmter Randbedingungen (beispielsweise Erfassung relevanter Eingangsparameter, welche in Korrelation zur betrachteten Ausgangsgröße stehen) beliebige Größen zeiteffizient darstellen lassen.

Die Grundlagen der Emissionsentstehung und die Charakteristika der jeweiligen Emission sind wichtig, um die physikalisch-phänomenologischen Ansätze nachzuvollziehen, und sie bilden die Basis für einen möglichen Vorteil der hybriden Modellansätze, worauf im späteren Verlauf dieser Arbeit näher eingegangen wird.

Theoretisch entsteht bei einer vollständigen Verbrennung eines Ottokraftstoffes mit Luft, dessen allgemeiner chemischer Aufbau sich durch die Summenformel $C_xH_yS_qO_z$ beschreiben lässt, von den oben genannten Emissionskomponenten nur Kohlenstoffdioxid. Da der hauptsächlich in der Luft enthaltene Stickstoff idealisiert nicht an der Verbrennung teilnimmt, lässt sich folgende Reaktionsgleichung aufstellen: [8]

$$C_xH_yS_qO_z + \left(x + \frac{y}{4} + q - \frac{z}{2}\right) * O_2 \Rightarrow \; x * CO_2 + \frac{y}{2} * H_2O + q * SO_2 \tag{9}$$

Dabei sind $x$, $y$, $q$ und $z$ Indizes, welche die Anzahl der Atome im jeweiligen Molekül beschreiben, was wiederum vom verwendeten Kraftstoff abhängt. In der Regel handelt es sich hierbei um ein Stoffgemisch, sodass die Indizes der statistisch gemittelten Kraftstoffzusammensetzung entsprechen. Aus der Reaktionsgleichung lässt sich direkt ableiten, dass zu einer vollständigen Kraftstoffverbrennung die exakte Menge (stöchiometrisch) an (Di-) Sauerstoff aus der Luft vorhanden sein muss, was in der Praxis nicht immer gegeben ist (beispielsweise durch lokale Unterschiede in der Gemischzusammensetzung). Dieses und weitere Phänomene führen dazu, dass bei der Verbrennung zusätzliche Komponenten entstehen, was im weiteren Verlauf des Kapitels näher erläutert wird.

Die theoretischen Grundlagen der Emissionsentstehung beinhalten Reaktionsgleichungen und Formulierungen, welche die Basis für eine physikalisch-phänomenologische Modellierung der Emissionskomponenten bilden können. Dadurch besteht eine direkte Verknüpfung zwischen den empirischen Beobachtungen und den Voraussagen, weshalb neben der Emissionsentstehung auch die dazugehörige Modellierung im jeweiligen Unterkapitel beschrieben wird. Dies beschränkt sich jedoch auf die primär im Rahmen der Arbeit genutzten Modellansätze, welche in der Simulationssoftware GT-Suite v2022 von Gamma Technologies LLC implementiert sind.

### 2.2.1 Stickstoffoxide

Stickstoffoxide (Stickoxide) beziehungsweise $NO_x$ ist ein Oberbegriff, welcher die Moleküle $NO$, $NO_2$, $NO_3$, $N_2O$, $N_2O_3$, $N_2O_4$ und $N_2O_5$ umfasst. Diese entstehen im motorischen Prozess hauptsächlich aus dem Sauerstoff und dem Stickstoff, welche aus der Luft stammen. Die wichtigsten Stickstoffoxide bei der ottomotorischen Verbrennung sind $NO$ und $NO_2$, wobei das Verhältnis $NO$ zu $NO_2$ in der Regel über 99:1 liegt. Stickstoffmonoxid wird jedoch nach einer Zeit unter atmosphärischen Bedingungen nahezu komplett zu Stickstoffdioxid oxidiert. [11]

**Zeldovich-Mechanismus:** Aufgrund des Anteils von Stickstoffmonoxid von über 99 % am gesamten Rohemissionsausstoß der Stickstoffoxide werden die maßgeblichen Mechanismen zur Bildung von $NO$ hergeleitet. Die Bildung von Stickstoffmonoxid lässt sich primär in zwei Prozesse unterteilen. Hierzu gehört die thermische $NO$-Bildung, welche 1946 von Zeldovich [32] erstmalig formuliert und 1970 von Lavoie et al. [33] zum sogenannten erweiterten Zeldovich-Mechanismus ergänzt wurde, welcher aus den folgenden drei Reaktionen besteht: [13]

$$O + N_2 \overset{k1}{\leftrightarrow} NO + N \tag{10}$$

$$N + O_2 \overset{k2}{\leftrightarrow} NO + O \tag{11}$$

$$N + OH \overset{k3}{\leftrightarrow} NO + H \tag{12}$$

$k1$, $k2$ und $k3$ sind Geschwindigkeitskonstanten, die empirisch ermittelt werden müssen. Dabei werden in der Literatur [7], [34], [35] unterschiedliche Angaben gemacht, was einen maßgeblichen Unsicherheitsfaktor bei der Modellierung der thermischen $NO$-Bildung mit dem erweiterten Zeldovich-Mechanismus darstellt. Die Geschwindigkeitskonstanten sind außerdem stark von der Temperatur abhängig. Die $NO$-Bildungsrate lässt sich aus den obigen drei Reaktionsgleichungen folgendermaßen formulieren: [13]

$$\frac{d[NO]}{dt} = k_{1,r}[O][N_2] + k_{2,r}[N][O_2] + k_{3,r}[N][OH] - k_{1,l}[NO][N] - k_{2,l}[NO][O] - k_{3,l}[NO][H] \quad (13)$$

Die Indizes $k_i$ sind mit dem Zusatz $r$ oder $l$ gekennzeichnet, da beim Zeldovich-Mechanismus eine Hin- ($r$) und eine Rückreaktion ($l$) berücksichtigt wird. Aus der *NO*-Bildungsrate lässt sich ableiten, dass es motorische Zustände geben kann, welche eine Bildung von *NO* begünstigen können. Es kann zusätzlich vorkommen, dass die Rückbildung von *NO* verlangsamt oder verhindert wird. Dies resultiert aus der beschriebenen Temperaturabhängigkeit der Geschwindigkeitskonstanten $k_i$ und der momentan vorherrschenden Konzentration der an der jeweiligen Reaktion beteiligten Stoffe. Für einen maßgeblichen Ablauf der Rückreaktion, also eines *NO*-Abbaus, muss die *NO*-Konzentration in Bezug auf die vorherrschende Brennraumtemperatur über der Gleichgewichtskonzentration liegen. Ein solches Verhältnis kann jedoch meist erst am Verbrennungsende vorliegen, wo jedoch die allgemeine Reaktionsgeschwindigkeit durch die niedrigen Temperaturen bereits abgesunken ist. Dadurch ist die Hinreaktion besonders bei hohen Temperaturen („thermische *NO*-Bildung“) in Bezug auf die *NO*-Bildungsrate bestimmend. So wird beispielsweise die *NO*-Bildung der ersten Teilreaktion verfünfzigfacht, wenn die Temperatur von 2000 K auf 2500 K ansteigt. Durch die sehr viel schneller veränderlichen physikalischen Zustände im Brennraum im Vergleich mit der chemischen Reaktionskinetik weicht die *NO*-Konzentration praktisch immer von der Gleichgewichtskonzentration ab. Die Hinreaktionen (11) und (12) laufen bedingt durch die empirisch ermittelten Geschwindigkeitskonstanten um ein Vielfaches schneller ab als die in Gleichung (10), was dazu führt, dass der in Gl. (10) produzierte Stickstoff umgehend umgesetzt wird und dadurch die Bildungsrate von $N$ gleich 0 gesetzt werden kann. Dies führt zu folgender Vereinfachung der *NO*-Bildungsrate: [13]

$$\frac{d[NO]}{dt} = 2 * k_{1,r}[O][N_2] - 2 * \ k_{1,l}[NO][N] \quad (14)$$

$[N]$ lässt sich nach [13] vereinfacht folgendermaßen bestimmen:

$$[N] = \frac{k_{1,r}[O][N_2] + k_{2,l}[NO][O] + k_{3,l}[NO][H]}{k_{1,l}[NO] + k_{2,r}[O_2] + k_{3,r}[OH]} \quad (15)$$

Werden Gl. (14) und Gl. (15) betrachtet, sind, abgesehen von der gesuchten Größe $[NO]$, noch die Konzentrationen $[O]$, $[N_2]$, $[H]$ und $[OH]$ unbekannt, können jedoch nach [13] unter der Annahme eines partiellen Gleichgewichts bestimmt werden. Dies ist unter anderem in [7] detailliert dargestellt.

**Prompt-*NO*:** Der zweite Mechanismus, welcher im Kontext der ottomotorischen Verbrennung für die Entstehung von Stickstoffmonoxid betrachtet wird, ist für die Entstehung des sogenannten „Prompt-*NO*“ verantwortlich. Die Prompt-*NO*-Bildung wurde von Fenimore [36] im Jahre 1971 beobachtet und beschrieben. Fenimore suchte dabei nach einer Erklärung für

die schnelle Bildung von *NO* in der primären Reaktionszone (Flammenfront), was sich mit den damals bekannten Mechanismen - wie dem vorgestellten Zeldovich-Mechanismus – nicht herleiten ließ. Die Grundlagen seiner Beobachtungen waren Experimente an Brennern mit verschiedenen Gemischen, wobei sich herausstellte, dass unter dem Vorhandensein von Kohlenwasserstoffen bereits bei einem Reaktionszeitpunkt nahe null Sekunden eine *NO*-Konzentration vorhanden war, ohne dass *NO* in den Reaktionspartnern enthalten war. Als relevante Gleichungen für die Vorgänge in der primären Reaktionszone formulierte Fenimore:

$$CH + N_2 \Leftrightarrow HCN + N \tag{16}$$

$$C_2 + N_2 \Leftrightarrow 2CN \tag{17}$$

Einerseits kann aus dem Stickstoff aus Gleichung (16) *NO* durch die im Zeldovich-Mechanismus beschriebenen Vorgänge gebildet werden, andererseits kann *HCN* im weiteren zeitlichen Verlauf auf verschiedene Weisen reagieren, sodass erneut Stickstoff freiwerden kann und als weitere *NO*-Quelle dient. Dies wurde durch Miller und Bowman [37] festgestellt.

Neben dem von Fenimore vorgestellten Prompt-*NO*-Mechanismus mit *HCN* als Zwischenprodukt, ist heute primär der im Jahre 2000 von Moskaleva et al. [38] formulierte Reaktionsablauf mit der Zwischenspezies *NCN* von Relevanz:

$$CH + N_2 \Rightarrow H + NCN \tag{18}$$

Dieser Ansatz korrelierte besser mit experimentellen Untersuchungen, was 2008 von Sutton et al. [39] validiert wurde. Auch hier ist *NCN* nicht das Endprodukt und es können verschiedene Reaktionen folgen, aus welchen Stickstoff oder direkt Stickstoffmonoxid gebildet werden. Im Gegensatz zum erweiterten Zeldovich-Mechanismus ist bei der Bildung von Prompt-*NO* auch nach heutigem Stand von einem erheblichen Forschungsbedarf auszugehen, um die Bildung von Stickstoffmonoxid in der primären Reaktionszone exakt zu modellieren [13].

Der Prompt-*NO*-Mechanismus wird im Gesamtkontext der *NO*-Bildung im Vergleich zum Zeldovich-Mechanismus als geringfügig eingestuft. Dies liegt daran, dass die Flammenfront als primäre Reaktionszone aufgrund der hohen Drücke, welche in einer Verbrennungskraftmaschine auftreten, sehr dünn ($\sim 0.1$ mm) ist. Zusätzlich ist in der Regel davon auszugehen, dass der Verbrennungsdruck in einem Großteil der Verbrennung weiter ansteigt, sodass das in einer frühen Verbrennungsphase umgesetzte Gemisch weiter verdichtet und erhitzt wird, was einer *NO*-Bildung förderlich ist. [7]

### 2.2.2 Kohlenstoffmonoxid

Die Bildung von Kohlenstoffmonoxid in Verbrennungsmotoren hängt primär vom Verbrennungsluftverhältnis λ ab. Im Gegensatz zu Dieselmotoren, welche in der Regel mager (λ deutlich größer als 1) betrieben werden (und somit kaum Kohlenstoffmonoxid-Emissionen erzeugen), ist bei Ottomotoren ein stöchiometrischer Betrieb die Regel. Dieser Bereich wird je nach Applikation in höheren Lasten zu Gunsten des Bauteilschutzes auch unterschritten (λ kleiner als 1). In beiden Fällen sind Kohlenstoffmonoxid-Emissionen beim Ottomotor als relevant einzustufen, was durch Abbildung 2-6 verdeutlicht wird. [7]

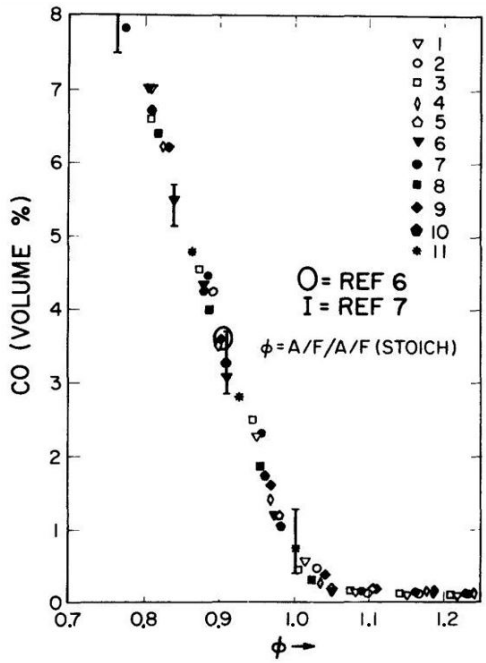


Abbildung 2-6: *CO*-Emission in % in Abhängigkeit des Verbrennungsluftverhältnisses (hier $\Phi = \lambda$) [40]

Abbildung 2-6 zeigt die Auswirkungen unterschiedlicher Verbrennungsluftverhältnisse auf die Kohlenstoffmonoxid-Emissionen. Dies wurde in [40] anhand elf verschiedener Kraftstoffe getestet, wobei sich unabhängig von der Zusammensetzung eine klar definierte Kurve ausbildete.

Die Bildung von Kohlenstoffmonoxid wird einem primären Reaktionspfad in der Verbrennung von Kohlenwasserstoffen zugeschrieben und lässt sich nach Bowman [41] folgendermaßen schematisch darstellen:

$$RH \Rightarrow R \Rightarrow RO_2 \Rightarrow RCHO \Rightarrow RCO \Rightarrow CO \tag{19}$$

$R$ ist hierbei im oberen Ablaufschema ein freies Kohlenwasserstoff-Radikal. Das während des Verbrennungsprozesses entstehende Kohlenstoffmonoxid wird anschließend zu Kohlenstoffdioxid oxidiert, wobei diese Reaktion vergleichsweiße langsam abläuft: [7]

$$CO + OH \Leftrightarrow CO_2 + H \tag{20}$$

Es wird angenommen, dass bei Bedingungen, welche unmittelbar hinter der Flammenfront herrschen können (Temperaturen > 2500 K, Drücke 15 – 40 bar), das System aus Kohlenstoff, Sauerstoff und Wasserstoff einen Gleichgewichtszustand einnimmt und damit Gl. (20) folgt. Dieser Zustand ist gegen Ende der Verbrennung durch eine zunehmende Expansion und Abkühlung des Arbeitsgases nicht mehr gegeben und es kann lokal zu Abweichungen von dieser Gesetzmäßigkeit kommen. [7]

Newhall [42] stellte in seiner Arbeit fest, dass die Kohlenstoffmonoxid-Oxidation während des Expansionsvorganges kinetisch limitiert ist und dadurch bei fortwährender Brenndauer die tatsächliche Kohlenstoffmonoxid-Konzentration die aus dem Gleichgewichtszustand berechnete Konzentration aus Gl. (20) zunehmend übersteigt, siehe Abbildung 2-7.

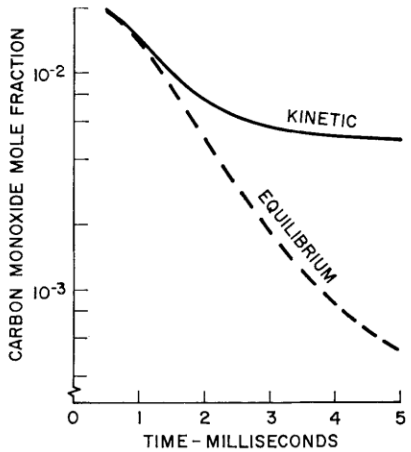


Abbildung 2-7: Kohlenstoffmonoxid-Konzentration während des Expansionsvorgangs; Kraftstoff = $C_8H_{18}$; $\lambda = 1$; Drehzahl = 4000 $min^{-1}$ [42].

Newhall [42] zeigte, dass es drei Reaktionsabläufe gibt, welche Einfluss auf die sehr schnelle und dominante Reaktion aus Gl. (20) nehmen können (z. B. durch eine Reduktion des verfügbaren $OH$) und deren Berücksichtigung die Modellvorhersagen besonders gegen Ende des Expansionsvorganges deutlich verbessern können.

$$H + H + M \Leftrightarrow H_2 + M \tag{21}$$

$$H + OH + M \Leftrightarrow H_2O + M \tag{22}$$

$$H + O_2 + M \Leftrightarrow HO_2 + M \tag{23}$$

### 2.2.3 Kohlenstoffdioxid

Die Emission von Kohlenstoffdioxid ($CO_2$) ist ein zentraler Prozess bei der Verbrennung von Kohlenwasserstoffverbindungen, zu welchen auch Ottokraftstoffe zählen. Dies folgt aus den Bruttoreaktionsgleichungen einer vollständigen Verbrennung (nach dem Zerfall der Kohlenwasserstoffverbindung) [13]:

$$H_2 + \frac{1}{2} O_2 = \ H_2O \tag{24}$$

$$C + O_2 = CO_2 \tag{25}$$

Wie aus Gleichungen (24) und (25) folgt, sind Wasser und Kohlenstoffdioxid bei dieser theoretischen Betrachtungsweise die einzigen anzunehmenden Reaktionsprodukte. Eine vollständige Umsetzung der Kohlenwasserstoffverbindung kann in einem motorischen Prozess jedoch nur gelingen, wenn genügend Sauerstoff verfügbar ist, was sich dadurch formulieren lässt, dass das Verbrennungsluftverhältnis $\lambda \ \geq 1$ gegeben sein muss. In diesem idealisierten Szenario ist die Bestimmung der Kohlenstoffdioxid-Emission bei der Kenntnis der chemischen Zusammensetzung der Kohlenwasserstoffverbindung unmittelbar möglich.

Die Forderung eines Verbrennungsluftverhältnisses größer 1 ist jedoch nicht nur global auf den gesamten Brennraum beschränkt, sondern muss auch örtlich sehr begrenzt (lokal) der Fall sein, was in der Praxis im gesamten Brennraum jedoch kaum sicherzustellen ist. Zusätzlich kann es vorkommen, dass wichtige Reaktionen nicht schnell genug ablaufen, was dazu führt, dass der chemische Gleichgewichtszustand nicht erreicht wird. [13]

Die Herausforderungen bei der Modellierung der Kohlenstoffdioxid-Emissionen sind eng mit denen der Kohlenstoffmonoxid-Emissionen verknüpft, weshalb auf das vorherige Kapitel und besonders Gl. (20) verwiesen wird.

### 2.2.4 Unverbrannte Kohlenwasserstoffe

Bei der ottomotorischen Verbrennung von Kohlenwasserstoffen und einem Verbrennungsluftverhältnis $> 1$ sind unmittelbar hinter der Flammenfront keine unverbrannten Kohlenwasserstoffe messbar. Das lässt darauf schließen, dass die nicht umgesetzten Kohlenwasserstoffe nicht oder nur teilweise an der Verbrennung teilnehmen und dadurch unverändert oder nur teiloxidiert vorliegen. [13]

Cheng et al. [43] haben die wichtigsten Mechanismen für die Emission von unverbrannten Kohlenwasserstoffen in Ottomotoren untersucht und folgendermaßen zusammengefasst:

- Gemisch, das in Spalten (wie beispielsweise dem Feuersteg) vorliegt, wird nicht oder nur teilweise von der Flammenfront erreicht
- Kraftstoff reichert sich vor der Verbrennung im Ölfilm an und desorbiert über die Zeit
- Kraftstoff, der sich in Ablagerungen/Verbrennungsrückständen anreichert, desorbiert und nimmt nicht an der Verbrennung teil
- Unvollständige Verbrennung durch Flame-Quenching (an der Zylinderwand oder an Spalten)
- Flüssiger Kraftstoff in Spalten, der nicht an der Gemischbildung teilnimmt
- Leckagen über die Auslassventile

Die zunächst unverbrannten Kohlenwasserstoffe aus den genannten Mechanismen finden sich nicht im gesamten Umfang im Abgas wieder. Eine Teilmenge verbleibt im Brennraum und wird später verbrannt. Außerdem gibt es unverbrannte Kohlenwasserstoffe, die über die Auslasskanäle zurückgesaugt oder im Abgassystem bzw. im Katalysator nachoxidiert werden. Der Anteil unverbrannter Kohlenwasserstoffe im Rohabgas von Ottomotoren liegt in der Regel bei circa 1-2 % des eingesetzten Brennstoffs. Bei einem Dieselmotor sind es prinzipbedingt deutlich weniger [7]. In Abbildung 2-8 sind die Mechanismen der Entstehung unverbrannter Kohlenwasserstoffe sowie eine quantitative Aufteilung auf die einzelnen Pfade für einen Ottomotor exemplarisch dargestellt.

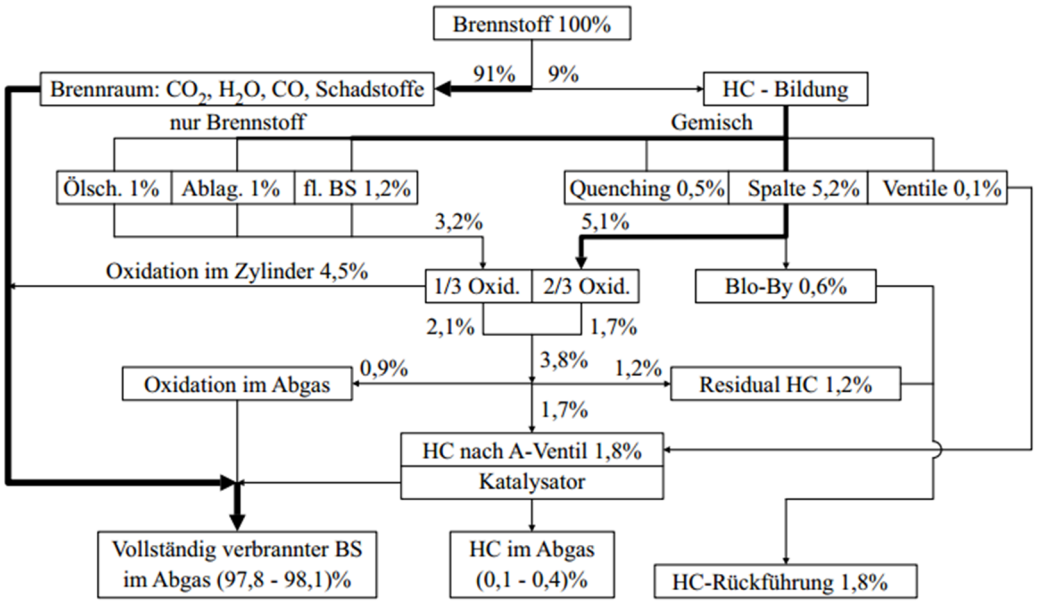


Abbildung 2-8: Bildung von unverbrannten Kohlenwasserstoffen in einem Ottomotor ([13] nach [43])

Abbildung 2-8 demonstriert, dass es viele unterschiedliche Mechanismen und Reaktionspfade bei der Bildung unverbrannter Kohlenwasserstoffe im Rohabgas und nach der Abgasnachbehandlung gibt. Die gezeigte Aufteilung ist nicht allgemeingültig, sondern stellt nur einen möglichen Fall dar. Eine Modellierung der unverbrannten Kohlenwasserstoffe unter Berücksichtigung aller Mechanismen ist sehr aufwändig, da insbesondere auch die Geometrie im Brennraum, das Flame-Quenching und die Verbrennung als solches detailliert dargestellt werden müssen [13].

Lavoie und Blumberg [44] haben 1977 einen Modellansatz für die Vorhersage der Emissionen und des Kraftstoffverbrauchs von Ottomotoren entwickelt, wovon insbesondere die Modellierung der unverbrannten Kohlenwasserstoffe relevant für die vorliegende Arbeit ist. Die *HC*-Emissionen werden in dem Modell berechnet, in dem die Regionen, in welchen die Flammenfront erlischt, modelliert werden und anschließend die Oxidation der darin verbleibenden, bisher unverbrannten Kohlenwasserstoffe während der Expansion und des Ausschiebetaktes kinetisch bestimmt wird. Die kritischen Parameter dabei sind die Abschätzung der Schichtdicken bzw. Volumina, welche von der Flammenfront nicht erreicht werden, und den Variablen, welche die Oxidationsrate nach der Flammenauslöschung beeinflussen.

In einer darauf aufbauenden Arbeit hat Lavoie [45] verfügbare Daten aus Experimenten zusammengefasst und analysiert, um Korrelationen für die zu bestimmenden Größen zu finden. Dazu ist es entscheidend, die Flammenauslöschung in Spalten und an der Zylinderwand zu berücksichtigen, was in der Modellvorstellung durch das sogenannte „2-plate quenching“ und „Single wall quenching“ gelöst werden kann.

**2-plate quenching:** Beim „2-plate quenching“ wird untersucht, bis zu welchem minimalen Abstand zweier paralleler Platten eine Flammenausbreitung möglich ist. Diese Größe wird als „2-plate quench distance“ bezeichnet und wenn sie unterschritten wird, ist die Wärmeabgabe an die Platten höher als die von der Verbrennung freigesetzte Wärme. Die Relation zwischen Wärmefreisetzung und der aufgenommenen Wärme durch die Umgebung wird durch die Péclet-Zahl ausgedrückt. Die Bestimmung des minimalen Plattenabstandes ist in der Modellierung wichtig, um abzuschätzen, ob sich die Flammenfront in Brennraumspalten (bspw. Feuersteg) ausbreiten kann. [45]

$$Pe_2 = \frac{9.5}{\lambda} * \frac{P}{3}^{0.26\min\,(1;\frac{1}{\lambda^2})} \tag{26}$$

$Pe_2$ ist die Péclet-Zahl, wobei der Zusatz „2“ beschreibt, dass die Gültigkeit für das „2-plate quenching“ gegeben ist. $P$ ist der herrschende Druck; hierbei muss beachtet werden, dass die Formel generell für Drücke zwischen 3 und 40 bar aufgestellt worden ist [45]. Lavoie hat Gleichung (26) definiert, indem er nach einer möglichst guten empirischen Übereinstimmung von vorhandenen Messdaten in dem genannten Druckbereich gesucht hat.

**Single wall quenching:** Obwohl das „2-plate quenching“ bei der Modellierung von Verbrennungsmotoren relevant ist, wird die Flammenfront maßgeblicher durch sogenanntes „Single wall quenching“ im Brennraum beeinflusst. Dies kann beispielsweise vorkommen, indem die Flammenfront frontal auf die Zylinderwand trifft, oder sich parallel zu dieser fortbewegt und dadurch ebenfalls eine Wärmeabgabe stattfindet. Für den ersten Fall, welcher nach [45] dominiert, lässt sich die Péclet-Zahl für das „Single wall quenching“ in Abhängigkeit der Péclet-Zahl für das „2-plate quenching“ folgendermaßen ausdrücken: [45]

$$\frac{Pe_1}{Pe_2} = 0.2 \tag{27}$$

Typische Abstandswerte für einen Ottomotor liegen beim „2-plate quenching“ zwischen 0.2 und 0.6 mm, beim „Single wall quenching“ sind es hingegen nur 0.04 bis 0.15 mm [7].

Wie bereits angedeutet, kommt es während der Expansion und dem Ausschiebetakt zu einer Oxidation von Kohlenwasserstoffen, welche nicht an der eigentlichen Verbrennung teilgenommen haben. Basierend auf verschiedenen Messungen kann folgende Gleichung nach [45] hierzu als Grundlage dienen, um die Reaktionsrate von unverbrannten Kohlenwasserstoffen anzunähern:

$$\frac{d[HC]}{dt} = -6.7 * 10^{15} * e * \frac{-37.23}{RT} f_{HC} * f_{O_2} * (\frac{P}{RT})^2 \tag{28}$$

$R$ ist die allgemeine Gaskonstante (circa 8.314 $\frac{J}{mol*K}$), $f_{HC}$ und $f_{O_2}$ stellen die jeweiligen Stoffmengenanteile von Kohlenwasserstoffen und Sauerstoff dar.

Neben der grundlegenden Arbeit von Lavoie gab es auch in den letzten Jahren weitere Anstrengungen, das Erlöschen der Flammenfront und die allgemeine Reaktionsgleichung für unverbrannte Kohlenwasserstoffe außerhalb der regulären Verbrennung vorherzusagen und auf spezifische Umgebungsbedingungen anzupassen. Als weiterführende Informationsquellen sind hier unter anderem Yoshimura et al. [46] und Boust et al. [47] zu nennen.

## 2.3 Maschinelles Lernen

In der heutigen Zeit, geprägt von Fortschritten in der Datenwissenschaft und Computertechnologie, hat sich Maschinelles Lernen (ML) als ein wichtiges Werkzeug in zahlreichen wissenschaftlichen und ingenieurstechnischen Disziplinen etabliert. Dieses Kapitel widmet sich den Grundlagen des Maschinellen Lernens, welches neben der physikalisch-phänomenologischen Modellierung als fundamentales Element der vorliegenden Arbeit zu bewerten ist. Dabei werden zunächst eine Definition des Maschinellen Lernens und eine Einordnung in die Künstliche Intelligenz (KI) vorgestellt, die dessen Kernprinzipien und die Fähigkeit, Muster in Daten zu erkennen und daraus zu lernen, umreißt.

Anschließend werden weitere Unterkategorien des Maschinellen Lernens behandelt, beginnend mit den „Shallow Learning“ Algorithmen, die die Basis für traditionelle ML-Methoden bilden und in zahlreichen Anwendungen ihre Effektivität unter Beweis gestellt haben. Diese Algorithmen sind entscheidend für das Verständnis der evolutionären Entwicklung hin zu aufwändigeren Modellen.

Der darauffolgende Abschnitt ist den „Deep Learning“ Algorithmen gewidmet, die durch ihre Fähigkeit, aus Daten mit mehrschichtigen abstrakten Repräsentationen zu lernen, neue Möglichkeiten bei der Darstellung komplexer Vorgänge bieten.

### 2.3.1 Definition

Maschinelles Lernen lässt sich als Teilgebiet der Künstlichen Intelligenz einordnen. Der Begriff der Künstlichen Intelligenz wurde erstmals 1955 von John McCarthy geprägt. Er definierte Künstliche Intelligenz unter anderem dadurch, dass eine Maschine in der Lage sein kann, jeden Aspekt von Intelligenz und den Prozess des Lernens nachzuahmen, wenn diese ausreichend genau beschrieben werden können. [48], [49]

Die Definition von Maschinellem Lernen ist nicht trivial und es gibt eine Vielzahl unterschiedlicher Interpretationen. Unter anderem kann Maschinelles Lernen als eine Zusammenfassung von Methoden verstanden werden, welche darauf abzielen, Muster in einer vorhandenen Datenmenge zu erkennen und anschließend diese Informationen zu verwenden, um Daten in der Zukunft vorherzusagen oder dadurch eine Entscheidungsfindung zu ermöglichen [50].

Weiterhin ermöglicht Maschinelles Lernen die Optimierung eines (oder mehrerer) Leistungsparameter anhand vorhandener Daten und kann sowohl prädiktiv als auch deskriptiv sein. Unabhängig vom Einsatzzweck des ML basieren die Modelle auf Grundlagen der Statistik. [51]

Die unterschiedlichen Interpretationen des Begriffs „Maschinelles Lernen“ sind auch der Tatsache geschuldet, dass es mehrere Unterteilungen des Fachgebietes gibt, unter anderem anhand der Art der Trainingsüberwachung, was nachfolgend erläutert wird.

**Überwachtes Lernen:** Überwachtes Lernen (englisch „supervised learning“) ermöglicht die Schaffung prädiktiver Modelle, welche darauf abzielen, einen (oder mehrere) Ausgangswert(e) Y anhand anderer Größen X – in diesem Kontext oft auch als „Feature“ benannt – vorherzusagen. Das Überwachte Lernen beschreibt dabei den Prozess, die Relation zwischen der Ausgangsvariablen und den Features zu erkennen bzw. zu erlernen. Während des Trainingsprozesses werden dem Algorithmus neben den Eingangswerten auch die Ausgangswerte zur Verfügung gestellt, wodurch der Algorithmus das gewünschte Verhalten „aufgeprägt“ bekommt, daher auch der Begriff „Überwachtes Lernen“. [52]

Der Trainingsprozess eines Modells aus der Kategorie des Überwachten Lernens ist in Abbildung 2-9 nach [53] dargestellt. Während des Trainings wird eine Teilmenge der Features bzw. Attribute aus der Umgebung, welche das zu erlernende Verhalten widerspiegelt, dem Modell zur Verfügung gestellt. Das Modell prädiziert Ausgangswerte, wobei diese anfangs aufgrund oft zufällig initialisierter Modellparameter stark fehlerbehaftet sind. Gleichzeitig werden die Features auch den korrekten Ausgangswerten zugeordnet, was schematisch durch einen Überwacher geschieht. Die korrekten Ausgangswerte werden den prädizierten Ausgangswerten gegenübergestellt, woraus ein Fehler resultiert, welcher an das Modell zurückgemeldet wird. Anhand eines Optimierungsalgorithmus kann das Modellverhalten dadurch iterativ verbessert werden.

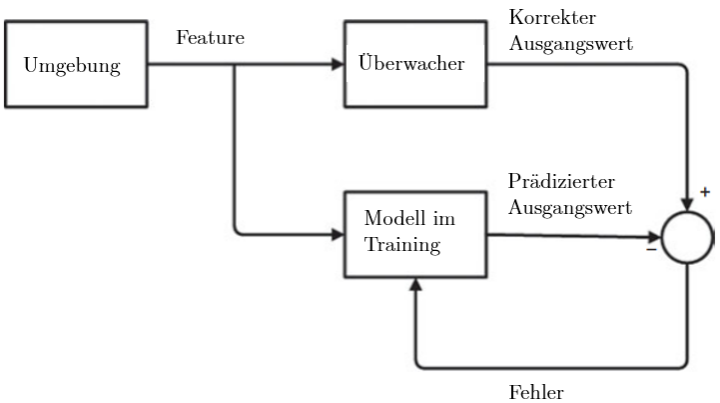


Abbildung 2-9: Schematische Darstellung des überwachten Lernens im Training nach [53]

Das Überwachte Lernen lässt sich anhand der zu beschreibenden Relation in die Gebiete „Regression“ und „Klassifizierung“ unterteilen. Beide können mathematisch über den Zusammenhang

$$y = g(x|\theta) \tag{29}$$

formuliert werden, wobei $g(\cdot)$ das gesuchte Modell, $x$ die Eingangsdaten, $\theta$ die Modellattribute und $y$ die Ausgangswerte/Zielgrößen darstellen [54].

Modellierungsaufgaben aus dem Bereich der Regression zeichnen sich dadurch aus, dass der Algorithmus einen numerischen Ausgangswert $y$ liefert. Die Ausgangswerte sind zudem Teil einer kontinuierlichen, stufenlosen Skala. [52]

Als Beispiele für Regressionsprobleme können die Vorhersage von Immobilienpreisen anhand der Nähe zum Zentrum und der Wohnfläche oder auch – wie in der vorliegenden Arbeit – die Prädiktion von Emissionskonzentrationen anhand motorischer Parameter genannt werden.

Klassifizierung im Kontext des Maschinellen Lernens bedeutet, dass der Ausgangswert $y$ eine endliche Menge diskreter Werte annehmen kann, beispielsweise binäre Werte oder Wahrscheinlichkeitsverteilungen. Diesen Wahrscheinlichkeiten werden häufig qualitative Klassen zugeordnet, z. B., ob es sich bei einer Reihe von Produkten um Hersteller A oder Hersteller B handelt. Da dies dennoch meist über Wahrscheinlichkeiten und damit numerische Werte dargestellt wird, gibt es viele Algorithmen, welche sowohl für Regressions- als auch für Klassifizierungsprobleme genutzt werden können. [52]

**Unüberwachtes Lernen:** Im Gegensatz zum Überwachten Lernen, zielt das Unüberwachte Lernen (englisch „unsupervised learning“) darauf ab, Strukturen in Daten zu finden. Diese sind dabei in der Regel nicht markiert (engl. „unlabeled“) und somit sind beim Unüberwachten Lernen keine korrekten Ausgangswerte vorhanden, wodurch das Fehlersignal nicht berechnet werden kann. Ein häufiges Anwendungsgebiet für das Unüberwachte Lernen ist das sogenannte „Clustering“, bei dem der Algorithmus lernt, vorhandenen Daten verschiedene Gruppen zuzuordnen, ohne dass die Benennung der Gruppen eine entscheidende Rolle spielt. Als Beispiel kann die Empfehlung ähnlicher Artikel zu einem in den Warenkorb gelegten Artikel aus dem Bereich des Online-Shopping genannt werden. Der große Vorteil des Unüberwachten Lernens ist dabei, dass die oft zeitintensive Arbeit für das Markieren bzw. Beschriften der Daten entfällt, was besonders bei großen Datenmengen zum Tragen kommt. Herausfordernd ist jedoch, dass im Vorfeld nur schwer abschätzbar ist, anhand welcher Merkmale die Daten in welche Gruppenanzahl unterteilt werden. Darauf kann durch die Auswahl der Algorithmen und die Aufbereitung der Daten Einfluss genommen werden. [55]

In Abbildung 2-10 ist der Ablauf für das Clustering schematisch dargestellt, womit der Algorithmus erklärt werden kann.

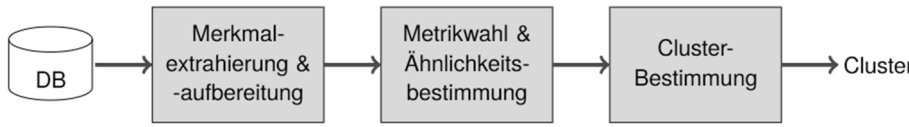


Abbildung 2-10: Schematischer Ablauf der Clusterbildung beim Unüberwachten Lernen [55]

Da es beim Unüberwachten Lernen keine Fehlerrückmeldung gibt, müssen andere Parameter als Zielgrößen für die Optimierung des Modells definiert werden. Daher wird einerseits die Maximierung der Ähnlichkeit von Daten innerhalb eines Clusters angestrebt, andererseits sollen die Unterschiede zwischen den Clustern möglichst groß sein. Ausgehend von einer Datenbasis (DB) werden Merkmale der Einzeldaten extrahiert und aufbereitet (beispielsweise normiert). Dies können je nach Aufgabe unter anderem geometrische Daten oder auch Helligkeitswerte von Fotos sein. Anschließend erfolgt eine Ähnlichkeitsbestimmung anhand der Merkmale und abschließend die Cluster-Bestimmung. [55]

Als weiterführende Literatur für das Clustering und die Vielzahl vorhandener Algorithmen kann das Buch von Xu und Wunsch [56] herangezogen werden.

**Bestärkendes Lernen:** Das Bestärkende Lernen (englisch „Reinforcement Learning“) zielt darauf ab, mit Hilfe von Künstlicher Intelligenz einen Agenten (Entscheider) zu kreieren, der auf Grundlage einer auf maximalem Nutzen basierenden Strategie Entscheidungen in einer komplexen Umgebung trifft. Der Lernprozess orientiert sich dabei am menschlichen Verhalten. Das anfangs unbekannte Umfeld wird im Trainingsprozess erforscht, der Agent kann damit interagieren und Erfahrungen sammeln. Diese Erfahrungen werden bewertet, sodass mit ausreichendem Training eine Strategie entsteht, um in der jeweiligen Situation möglichst gute Entscheidungen treffen zu können. Je nach Art des Problems liegen dem Agenten alle notwendigen Informationen vor, um die Situation (auch Zustand genannt oder englisch „state“) eindeutig beschreiben zu können, oder aber es besteht eine Ungewissheit in der Beobachtung des Zustands, was bei realitätsnahen und damit oft komplexen Problemen häufig vorkommt. Das Bestärkende Lernen nutzt sowohl Methoden des Überwachten als auch des Unüberwachten Lernens und grenzt sich hauptsächlich durch den Zweck (Entscheidungsfindung) als drittes Hauptgebiet des Maschinellen Lernens ab. [57]

Die grundlegende Interaktion zwischen dem Agenten und seiner Umgebung ist in Abbildung 2-11 dargestellt und wird anschließend erörtert.

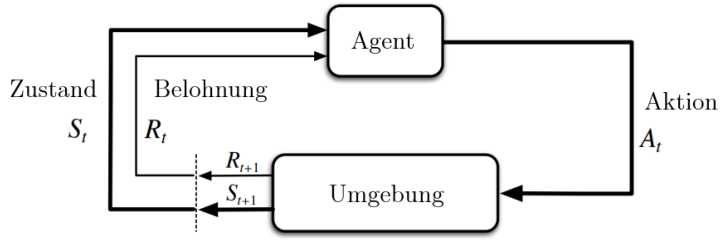


Abbildung 2-11: Bestärkendes Lernen - Interaktion zwischen Agent und Umgebung nach [58]

Zum Zeitpunkt $t$ observiert der Agent seine Umgebung, welche durch den Zustand $S_t$ repräsentiert wird. Ebenfalls erhält er eine Belohnung $R_t$, welche aus seinen Aktionen im Zeitschritt $t-1$ resultiert. Basierend auf seiner bisher erlernten Strategie führt er eine Aktion $A_t$ mit der Erwartungshaltung einer Belohnung $R_{t+1}$ aus und sorgt durch seine Entscheidung für die Erreichung des Zustands $S_{t+1}$.

Neben den bereits beschriebenen Elementen des Bestärkenden Lernens (Agent, Umgebung, Strategie, Aktion, Belohnung) gibt es nach Sutton und Barto [58] noch die Zustandswertfunktion und das Umgebungsmodell als wichtige Säulen dieser Methode. Die Zustandswertfunktion kann als langfristiges Ziel betrachtet werden, im Gegensatz zur Belohnung, welche sich auf einen Zeitschritt bezieht. Sie beschreibt die Summe der Belohnungen, die ausgehend von einem Zustand durch zukünftige Aktionen und dem damit verbundenen Durchlaufen weiterer Zustände vom Agenten erwartet werden. Folgerichtig wird ein trainierter Agent versuchen, Zustände mit einer möglichst hohen Zustandswertfunktion anzustreben, auch wenn die initiale, inkrementale Belohnung nicht maximal sein sollte. Damit ein Agent abschätzen kann, welchen Zustand seine Umgebung durch eine Aktion annimmt, ist ein Umgebungsmodell erforderlich. Algorithmen des Bestärkenden Lernens, die über diese Fähigkeit verfügen, werden „modellbasiert“ genannt. Es gibt jedoch auch Methoden, welche ohne Umgebungsmodell (modellfrei) auskommen, der Agent also weniger planungsbasiert, sondern eher über „Versuch und Irrtum“ definiert werden kann. Unter den modernen Methoden des Bestärkenden Lernens findet sich zwischen diesen Extremen ein breites Spektrum von Zwischenstufen. Beispielsweise kann ein Agent durch „Versuch und Irrtum“ lernen, dabei ein Umgebungsmodell erstellen und dieses anschließend zur Planung verwenden. [58]

### 2.3.2 Neuronale Netze und Deep Learning Algorithmen

Unabhängig von den in den vorherigen Kapiteln betrachteten Lernmethoden gibt es verschiedene Algorithmen des Maschinellen Lernens, die sich grob in zwei Kategorien einteilen lassen – Shallow Learning und Deep Learning. Shallow Learning Algorithmen, wie lineare Regression oder Support Vector Machines (SVMs), zeichnen sich in der Regel dadurch aus, dass zwischen dem Ein- und Ausgang nur eine weitere datenverarbeitende Ebene vorhanden ist. Demgegenüber verfügen Deep Learning Algorithmen über mehrere Verarbeitungsschichten, was bei der Erfassung und Abstrahierung komplexer Muster in großen Datenmengen vorteilhaft sein kann. Beispiele für Deep Learning Algorithmen sind Feed Forward Neural Networks (FNNs – Vorwärtsgerichtete Neuronale Netze) und Recurrent Neural Networks (RNNs – Rekurrente Neuronale Netze).

Shallow Learning Algorithmen erlauben durch die einfache Architektur eine Interpretierbarkeit und Verständlichkeit der Modelle, was in Abhängigkeit der Aufgabe von Vorteil sein kann. Darüber hinaus sind sie effizient bezüglich der notwendigen Datenmenge, die erforderlich ist, um eine ausreichend genaue Abbildung des betrachtenden Sachverhaltes zu erreichen. Ihre Nachteile liegen jedoch in der begrenzten Fähigkeit, mit komplexen oder hochdimensionalen Daten umzugehen. Im Gegensatz dazu liegen die Vorteile von Deep Learning Algorithmen in ihrer Fähigkeit, sehr komplexe Muster und Zusammenhänge aus großen, mehrdimensionalen Datenmengen zu erkennen. Dies ist jedoch mit einer höheren Komplexität, einer verminderten Interpretierbarkeit sowie einem größeren Rechen- und Datenbedarf verbunden.

Deep Learning Algorithmen werden in praktischen Anwendungen vermehrt bevorzugt. Begünstigt durch den Aufbau aus mehreren Verarbeitungsschichten, deren Verhalten nicht vorgegeben, sondern erlernt wird, ist der manuelle Aufwand bei der Modellerstellung gering. Dadurch profitieren Deep Learning Algorithmen unmittelbar von der wachsenden Datenmenge und der (exponentiell) steigenden Rechenleistung. [59]

In einer Vorarbeit [60] wurden verschiedene Algorithmen aus dem Bereich des Shallow und Deep Learning untersucht. Besonders für die Emissionsvorhersage im Kontext einer optimierten Betriebsstrategie haben sich die betrachteten Deep Learning Algorithmen als praktikabel erwiesen, weshalb auch in der vorliegenden Arbeit der Fokus auf diesen Methoden liegt.

**Vorwärtsgerichtete Neuronale Netze:** Bevor die komplexeren Deep Learning Algorithmen beschrieben werden können, ist es notwendig, den grundlegenden Aufbau von Neuronalen Netzen – beginnend mit dem kleinsten Element, dem Neuron oder Perzeptron – darzustellen. Bereits 1943 wurde das Neuron von McCulloch und Pitts [61] als Einzelelement in einem größeren Netzwerk, welches die Aktivitäten in einem Nervensystem darstellen kann, beschrieben. Durch die Verschaltung mehrerer Neuronen wurden logische Operationen (Und, Oder, etc.) ermöglicht, obwohl in der Modellvorstellung mehrere Eingänge, aber nur ein Ausgang

mit zwei möglichen Zuständen (aktiviert oder nicht aktiviert) existierten. Es wurde angenommen, dass die Aktivierung des Neurons von einem fixierten Schwellenwert abhängt. Rosenblatt [62] hat das Konzept des Neurons weiterentwickelt und den Begriff des Perzeptrons geprägt, was als Grundlage der heutigen Neuronalen Netze angesehen werden kann. Abbildung 2-12 [63] zeigt ein einlagiges Perzeptron mit zwei Input-Neuronen.

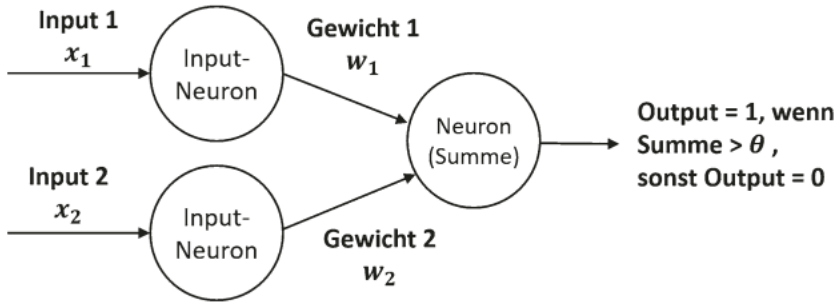


Abbildung 2-12: Einlagiges Perzeptron mit zwei Eingängen und einem Ausgang [63]

Das Perzeptron aus dem Beispiel hat zwei Eingänge $x_1$ und $x_2$. Diese werden intern mit den Gewichtungen $w_1$ und $w_2$ multipliziert und beide Produkte addiert. Diese Summe wird mit einem Schwellenwert $\theta$ verglichen und je nachdem, ob dieser über- oder unterschritten wird, ist die Ausgabe des vorliegenden Perzeptrons eins oder null. Da die Gewichtungen $w_1$ und $w_2$ und der Schwellenwert $\theta$ variabel sind, kann sich bereits dieses einfache Perzeptron an die jeweilige Aufgabe anpassen und dadurch ein Verhalten „erlernen“.

Neben der in Abbildung 2-12 gezeigten Schwellenfunktion gibt es noch weitere Aktivierungsfunktionen, die im Kontext der Neuronalen Netze eingesetzt werden. Hierzu zählen beispielsweise die Sigmoidfunktion oder die ReLU-Funktion (engl. „Rectified Linear Unit“, was sich als „Gleichrichter“ übersetzen lässt). Die Aktivierungsfunktionen sind dabei auf die jeweilige Anwendung anzupassen und ermöglichen es, ein nicht-lineares Verhalten abzubilden. Die Schwellen- oder Sprungfunktion eignet sich beispielsweise für eine binäre Klassifizierung, da letztere jedoch nicht überall stetig ist, kann es dadurch je nach Trainingsalgorithmen zu Problemen kommen [64].

Die Weiterentwicklung des einlagigen Perzeptrons ist das Mehrlagige Perzeptron (MLP). Das MLP (das zu den künstlichen Neuronalen Netzen gehört) verfügt neben bereits bekannten Ein- und Ausgabeschichten des Perzeptrons über mindestens eine weitere, verdeckte Schicht (hidden layer). Durch die Einführung der mehrlagigen Perzeptrons konnten bereits viele der bis dato bestehenden Limitierungen – beispielsweise die Abbildung der ausschließenden Disjunktion (XOR), welche von Minsky und Papert [65] aufgezeigt wurde – überwunden und deutlich komplexere Systeme abgebildet werden.

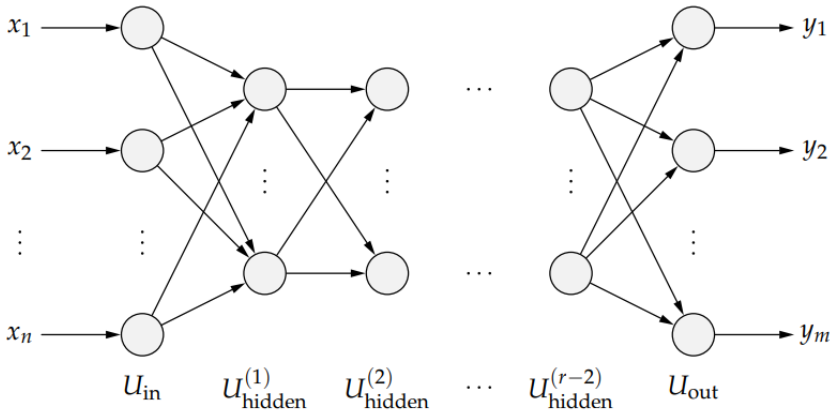


Abbildung 2-13: Aufbau eines mehrlagigen Perzeptrons [66]

Das in Abbildung 2-13 dargestellte $r$-lagige Perzeptron verfügt neben der Eingangs- und Ausgabeschicht ($U_{in}$ und $U_{out}$) über verdeckte Schichten ($U_{hidden}$). Die dargestellten Neuronen sind mit jeweils allen Neuronen aus den benachbarten Schichten verbunden, was als „fully connected“ bezeichnet wird. Die Pfeilrichtung symbolisiert den Informationsfluss, welcher strikt von den Eingängen ($x_1$ bis $x_n$) über alle Schichten bis zu den Ausgängen ($y_1$ bis $x_m$) stattfindet. Dies erklärt den Ursprung des Begriffes „Vorwärtsgerichtetes Neuronales Netz“. Aufgrund dieser Architektur wird das Ausgabeverhalten des (trainierten) Netzes nur von den aktuellen Eingangswerten und nicht von vergangenen Durchläufen beeinflusst.

**Rekurrente Neuronale Netze:** Wenn zeitlich basierte beziehungsweise sich dynamisch ändernde Daten verarbeitet werden sollen, werden häufig Rekurrente Neuronale Netze (engl. „Recurrent Neural Networks“ RNN) eingesetzt.

Die ursprünglichen RNNs haben einen ähnlichen Aufbau wie ein mehrlagiges Perzeptron (Abbildung 2-13), besitzen jedoch einen internen Zustand $H_t$, welcher sich nach jedem Zeitschritt in Abhängigkeit des aktuellen Modelleingangs und des vorherigen Zustands $H_{t-1}$ aktualisiert. Diese Architektur kann in Abbildung 2-14 nachvollzogen werden. [64]

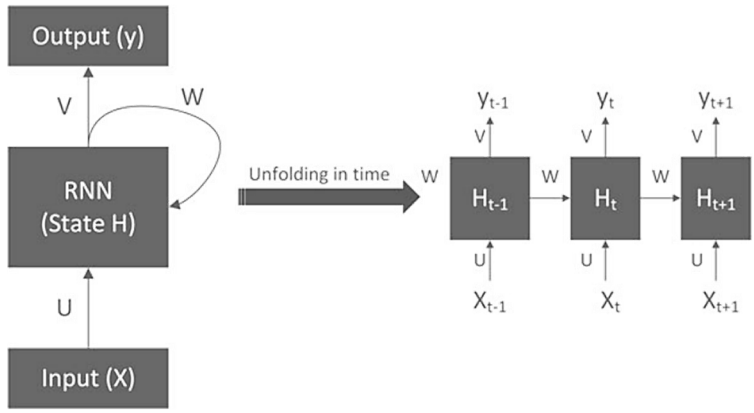


Abbildung 2-14: Architektur eines Rekurrenten Neuronalen Netzes nach [64]

Dargestellt ist ein einfaches Rekurrentes Neuronales Netzwerk mit dem Zustand $H$, einem Eingang $x$ und dem Ausgang $y$. Dem zeitlichen Verlauf $t-1 \rightarrow t+1$ kann entnommen werden, wie sich der interne Zustand ändert und jeweils Einfluss auf den darauffolgenden Zeitschritt nimmt. Dadurch lassen sich mit RNN zeitlich basierte Vorgänge in der Regel besser als mit einem FNN modellieren. Eine Herausforderung der bisher besprochenen klassischen Rekurrenten Neuronalen Netze liegt jedoch im Trainingsprozess.

Obwohl das Training von Neuronalen Netzen in der Tiefe in dieser Arbeit nicht behandelt wird (siehe hierzu z. B. [67]), ist das Verständnis grundlegender Elemente hilfreich für die Herleitung und Anwendung nachfolgender Algorithmen. Hierzu wird vom Überwachten Lernen (Kap. 2.3.1) ausgegangen, bei welchem während des Trainings der reale Ausgangswert bekannt ist und mit dem vom Modell prädizierten Wert verglichen wird. Die dabei auftretenden Abweichungen werden häufig durch Fehlerrückführung (engl. „Backpropagation“) im Verlauf des Trainings verringert, indem die Gewichtungen des Netzwerks angepasst werden.

Die Fehlerrückführung kann bei Rekurrenten Neuronalen Netzen problematisch sein, da das Fehlersignal beim temporal rückwärts gerichteten Durchlauf je nach Tiefe und Verlauf der Datensequenz „explodieren“ (zu groß werden) oder verschwinden kann [68]. Im ersten Fall kann sich dies in einem Oszillieren der Gewichtungen äußern, wohingegen im zweiten Fall der Lernprozess sehr lange dauert oder unter Umständen nicht zum Erfolg führt [68]. Dies liegt unter anderem daran, dass je weiter bei der Fehlerrückführung „zurückgerechnet“ wird, desto öfter kommt es dazu, dass ein aufkommender Fehlerterm mit einem Skalierungsfaktor multipliziert wird. Je nachdem, ob letzterer kleiner oder größer ist, kann das Produkt gegen null oder gegen unendlich streben.

**Long Short-Term Memory:** Hochreiter und Schmidhuber [68] haben ein Konzept vorgestellt, um die genannten Limitierungen der klassischen Rekurrenten Neuronalen Netzwerke (detailliert in Hochreiter [69]) zu überwinden. Der Begriff Long Short-Term Memory (LSTM,

„langes Kurzzeitgedächtnis“) wird unter anderem dadurch geprägt, dass das Modell Vorgänge von mehr als 1000 Zeitschritten berücksichtigen kann, ohne dabei an Dynamik bezüglich zeitlich aktuellerer Vorgänge einzubüßen. Dies wird durch einen gradientenbasierten Algorithmus und eine Architektur erreicht, welche gemeinsam darauf abzielen, eine konstante Fehlerrückführung zu garantieren und dadurch das beschriebene „Explodieren“ und Verschwinden zu vermeiden. Hierzu werden interne Zustände spezieller Zellen verwendet, welche in der nachfolgenden Abbildung 2-15 erläutert werden. [68]

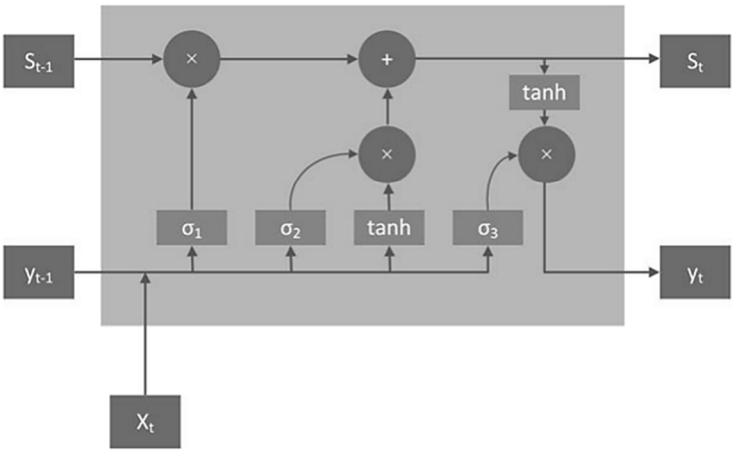


Abbildung 2-15: Architektur einer LSTM-Zelle nach [64]

Der gezeigte Aufbau hat bezüglich der Ein- und Ausgänge Ähnlichkeiten mit der in Abbildung 2-14 dargestellten Architektur eines klassischen Rekurrenten Neuronalen Netzes. Der Zustand der LSTM-Zelle zum Zeitpunkt $t$ ist jedoch mit $S_t$ definiert. Die Eingänge in die Zelle sind erneut der vergangene Ausgang (zum Zeitpunkt $t-1$), der vergangene Zustand $S_{t-1}$ und der aktuelle Eingang $x_t$. Die beiden Ausgänge sind der aktuelle Zustand $S_t$ und die Modellvorhersage $y_t$. Die Neuheit der LSTM-Zelle zeigt sich im Inneren, vor allem durch die Gates $\sigma_1$, $\sigma_2$ und $\sigma_3$. $\sigma_1$ ist das „forget gate“, an welchem der aktuelle Eingang mit dem vorherigen Ausgang kombiniert wird (Wertebereich zwischen 0 und 1) und anschließend mit dem vorherigen Zustand $S_{t-1}$ multipliziert wird. Dadurch entscheidet $\sigma_1$ darüber, wie relevant der vergangene Zustand für die Berechnung des aktuellen Zustands $S_t$ und Ausgabewertes $y_t$ ist. Das „input gate“ $\sigma_2$ bestimmt darüber, welche neuen Informationen in die Berechnung des aktuellen Zustands zum Zeitpunkt $t$ eingehen. Die Eingänge in $\sigma_2$ gleichen zwar denen von $\sigma_1$, dort wird jedoch der kombinierte Ausgang mit dem *tanh* des „input gate“ Eingangs multipliziert. Dieses Produkt wird zu dem vorherigen Zustand $S_{t-1}$ addiert, woraus der aktuelle Zustand $S_t$ resultiert. Der Ausgang des „output gates“ $\sigma_3$ (gleiche Eingangsgrößen und Wertebereiche wie $\sigma_1$ und $\sigma_2$) wird mit dem *tanh* des aktuellen Zustands $S_t$ multipliziert, wodurch der Modellausgang $y_t$ berechnet wird. [64]

Das spezifische Verhalten der gates $\sigma_1$, $\sigma_2$ und $\sigma_3$ wird in Abhängigkeit der jeweiligen Aufgabe (bedingt durch die Architektur und die Trainingsdaten) erlernt, was die Nutzung von LSTM-Zellen sehr flexibel macht [68]. Neben dem *tanh* als interne Aktivierungsfunktion wird für die Funktion der gates häufig eine sigmoid-Funktion (Spezialform der logistischen Funktion) eingesetzt.

Zusammengefasst eignet sich die Long Short-Term Memory-Architektur sehr gut für die Darstellung von zeitlich basierten Abläufen und dadurch potenziell für die Modellierung physikalischer Vorgänge.

### 2.3.3 Verknüpfung von Methoden des Maschinellen Lernens

Die Kombination aus Modellen oder allgemeiner aus Methoden des Maschinellen Lernens wird oft als „Ensemble Learning“ bezeichnet. Ensemble Learning zielt darauf ab, die Stärken mehrerer einfacher Modelle zu vereinigen, um ein leistungsfähigeres Gesamtmodell zu erhalten. Zu den bekanntesten Ensemble Learning Methoden zählen „Bagging“, „Boosting“ und „Stacking“. [70]

**Bagging** (abgeleitet aus dem englischen Begriff „**b**ootstrap **agg**regat**ing**“), zielt darauf ab, die Varianz der Modellprädiktion zu reduzieren. Hierfür wird ein vorhandener Trainingsdatensatz $Z$ in mehrere Datensätze $Z^{*b}, b = \{1, .., B\}$ (auch „bootstrap sample“ genannt), in welchen Elemente aus $Z$ enthalten sind, die jedoch neu kombiniert (Reihenfolge) oder mehrfach enthalten sein können, aufgeteilt. Auf jeden der bootstrap sample Datensätze wird ein Modell trainiert und die finale Prädiktion des Modells $\hat{f}_{bag}(x)$ setzt sich im einfachsten Fall aus dem Mittelwert der Einzelprädiktionen $\hat{f}^{*b}$ anhand folgender Gleichung zusammen. [70]

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{30}$$

Der Begriff **Boosting** wurde von Schapire [71] geprägt und versucht, ausgehend von schwachen Klassifikatoren – welche einzeln betrachtet keine hohe Prädiktionsgenauigkeit besitzen (dafür einfach aufgebaut und schnell ausgewertet sind) – ein Gesamtmodell mit hoher Präzision zu schaffen. Hierzu sind verschiedene Algorithmen entwickelt worden, wobei die Grundlagen anhand des sehr erfolgreichen AdaBoost Algorithmus von Freund und Schapire [72] erläutert werden.

Nachfolgend wird von einem Datensatz $X$ x $Y$ ausgegangen, wobei $Y$ zur Vereinfachung nur die Werte 0 und 1 annehmen kann. Aus dem Datensatz $X$ x $Y$ werden $N$ Trainingsdaten $(x_1, y_1), ..., (x_N, y_N)$ zufällig nach einer definierten, jedoch unbekannten Verteilung $P$ extrahiert. Als nächstes wird aus der Menge der Trainingsdaten eine Verteilung $D$ definiert und

das Ziel des AdaBoost-Algorithmus ist es, eine finale Hypothese $h_f(x)$ zu erhalten, welche eine geringe Abweichung über $D$ besitzt. Die finale Hypothese ist dabei aus mehreren einzelnen Hypothesen $h_t(x)$ aufgebaut, wobei $t$ für die jeweilige Iteration aus der Menge 1 bis $T$ steht. Die einzelnen Hypothesen sind schwache Klassifikatoren und werden beispielsweise mit dem PAC (engl. „Probably Approximately Correct“) Lernalgorithmus (siehe [73]) aufgebaut, wodurch Einfachheit und Geschwindigkeit sichergestellt und Überanpassungen (engl. „overfitting“) vermieden werden. [72]

Während jeder Iteration werden den Trainingsdaten Gewichtungen zugewiesen mit dem Ziel, bisher falsch zugeordnete Modellvorhersagen stärker zu berücksichtigen und bereits korrekt dargestellte Datenpaare zu ignorieren. Dadurch werden bei fortlaufendem Trainingsschritt vermehrt Hypothesen erstellt, welche auf bisher schwierig zu prädizierenden Daten basieren. Am Ende werden alle im Prozess erstellten Hypothesen $h_t(x)$ über eine gewichtete Mehrheitsentscheidung zur finalen Hypothese $h_f(x)$ kombiniert. [70]

**Stacking** oder „Stacked Generalization“ wurde von Wolpert [74] eingeführt. Es zielt ebenfalls darauf ab, die Vorhersagegenauigkeit eines oder mehrerer Modelle zu verbessern, nutzt jedoch im Vergleich zum Bagging oder Boosting einen anderen Ansatz. Ausgehend von einem oder mehreren Basismodellen, welche anhand eines Trainingsdatensatzes $\theta$ oder jeweils nur einer Teilmenge davon trainiert worden sind, wird beim Stacking davon ausgegangen, dass diese Modelle sich in einem „Level-0“ bewegen und in diesem vektoriell Eingangs- mit Prädiktionswerten verbinden. Beim Stacking bildet der entstehende Datensatz die Trainingsgrundlage für ein weiteres Modell, welches sich auf der „Level-1“ Ebene bewegt. Um das dadurch entstehende Modell für die ursprüngliche Fragestellung nutzen zu können, muss die Prädiktion wieder auf die „Level-0“ Ebene rücktransformiert werden. Neben dem beschriebenen Ansatz können ebenfalls nach Bedarf weitere Ebenen eingeführt werden, um das gewünschte Modellverhalten zu erreichen. [74]

# 3 Stand der Technik

## 3.1 Rolle der Emissionsmodellierung in der Antriebsentwicklung

Emissionsmodelle spielen eine wichtige Rolle in der Auslegung moderner Antriebssysteme und können für verschiedene Entwicklungsziele eingesetzt werden. Durch die stetige Verschärfung der Emissionsvorschriften – wie etwa die kommende Euro 7-Norm im PKW-Bereich – ist es für die Hersteller wichtig, bereits früh im Entwicklungsprozess abschätzen zu können, ob ihre Fahrzeuge diese Grenzwerte in unterschiedlichen Betriebsbedingungen einhalten können. Da besonders in dieser Phase die Versuchsträger limitiert sind und noch mit Änderungen (beispielsweise Ausstattungsvarianten im Fahrzeug und dadurch veränderliche Gesamtgewichte) gerechnet werden muss, kann durch die virtuelle Erprobung schnell und kosteneffizient reagiert werden.

Emissionsmodelle können auch die Entwicklung effektiverer Abgasnachbehandlungssysteme unterstützen. Durch das Verständnis, wie verschiedene Betriebsbedingungen die Emissionen beeinflussen, können Systeme entwickelt werden, die diese Emissionen effizienter und kostengünstiger reduzieren können. So haben Ericson et al. [75] bereits 2006 ein Simulationsmodell für die Vorhersage von Stickstoffoxiden von Dieselmotoren vorgestellt, welches sich sowohl für die Auslegung von Abgasnachbehandlungssystemen als auch in vereinfachter Form mit der damaligen Rechenleistung für modellbasierte Regelungsaufgaben eignete.

Neben der reinen Vorhersage der Emissionen für die Entwicklung beziehungsweise Verbesserung nachgeschalteter Systeme eröffnet die Entwicklung und zunehmende Verbreitung von Hybridfahrzeugen signifikante Möglichkeiten, den Betrieb des Verbrennungsmotors auf einen emissionsreduzierten Betrieb zu optimieren. Durch die flexible Aufteilung der vom Fahrer gewünschten Last auf zwei oder mehr Antriebsysteme besteht die Möglichkeit, den Verbrennungsmotor hauptsächlich in Bereichen des Kennfeldes zu betreiben, welche sich durch niedrige spezifische Emissionen auszeichnen. Dies lässt sich unter anderem mittels Lastpunkverschiebung erreichen. Durch den Einsatz fortschrittlicher präziser Vorhersagemodelle für Emissionen und Optimierungsmethoden kann eine Betriebsstrategie entwickelt werden, die diesen Ansatz unterstützt. So können unter anderem Methoden der modellprädiktiven Regelung oder des Reinforcement Learning eingesetzt werden, um in kurzer Zeit gegenüber einer auf statischen Regeln basierten Betriebsstrategie weitere Potenziale in der Emissionsreduzierung zu eröffnen [76].

Für all diese Anwendungsgebiete sind gezielt optimierte Emissionsmodelle erforderlich, welche sowohl hinsichtlich Genauigkeit als auch Rechengeschwindigkeit für den Einsatzzweck geeignet sind. Wie bereits erwähnt wurde die Emissionsmodellierung bis vor wenigen Jahre primär auf Basis von physikalisch-phänomenologischen Ansätzen durchgeführt. Diese sind in der Regel auf stationäre Betriebspunkte optimiert und haben den Vorteil, dass sie durch das Vorhandensein eines vorgeschalteten Verbrennungsmodells auch auf Veränderungen am Verbrennungsmotor/-prozess reagieren können. Wegen der Optimierung der Modelle auf den stationären Bereich können sie jedoch den transienten Bereich – welcher zumeist den Großteil realer Straßenfahrten umfasst – nur ungenügend abbilden.

Datenbasierte Ansätze können den letztgenannten Nachteil ausgleichen, wenn sie mit entsprechenden transienten Messdaten erstellt werden. Durch die steigende Verfügbarkeit kostengünstiger Rechenleistung lassen sich die Modelle schnell erstellen. Eine Änderung am Versuchsträger beziehungsweise an dessen Betriebsweise, welche bisher nicht in den Daten erfasst wurde, führt jedoch zumeist zu einer ungenügenden Vorhersagequalität. Diese muss durch eine erneute Datenaufnahme und weiteres Training ausgeglichen werden.

Eine Kombination von physikalisch-phänomenologischen und datenbasierten Ansätzen kann sich dementsprechend als vorteilhaft erweisen, um ein im transienten Betrieb präzises und gleichzeitig physikinformiertes Modell zu erschaffen, das auf Änderungen am Grundsystem (Verbrennungsmotor) oder auch den Randbedingungen (beispielsweise der Ansaugtemperatur) reagieren kann, ohne dass eine neue Datengrundlage erforderlich ist.

Zusammengefasst werden zwar die Vorteile der datenbasierten Modellierung vermehrt genutzt, jedoch gibt es ebenfalls noch Argumente, die physikalisch-phänomenologische Modellierung einzusetzen. Nachfolgend wird der Stand der Technik der einzelnen Methoden weiter ausgeführt.

## 3.2 Physikalisch-phänomenologisch basierte Modellierung von Emissionen

Physikalisch-phänomenologisch basierte Modelle erfordern ein Verständnis darüber, wie sich die betrachtete Emissionskomponente im Brennraum in Abhängigkeit der äußeren Umgebungsbedingungen wie etwa Stoffkonzentrationen, Temperaturen oder Drücke verhält. Die Grundlage hierfür bilden physikalische Gleichungen und aus experimentellen Beobachtungen abgeleitete phänomenologische Gesetzmäßigkeiten. Für den Einsatz dieser Emissionsmodelle sind auf den Anwendungsfall optimierte Verbrennungsmodelle erforderlich, welche schwer oder nicht messbare, aber für die Emissionsberechnung zwingend erforderliche Größen unter Zuhilfenahme von Mess-, Stoff- und Geometriedaten bestimmen können. Hierzu zählen beispielsweise die Aufteilung und die Zustandsgrößen der verbrannten und unverbrannten Zonen.

Solche Verbrennungsmodelle wurden beispielsweise von Pischinger et al. [17] und Merker et al. [13] entwickelt.

Die Auswahl des Verbrennungsmodells und der erforderlichen Komplexität hängt von der Emissionsspezies ab, die für die Untersuchung von Interesse ist. Exemplarisch verwenden Esposito et al. [77] für die Modellierung von Kohlenstoffmonoxid einen reduzierten Kraftstoffoxidationsmechanismus basierend auf Cai et al. [78], wofür ein 0D-1D-Verbrennungsmodell eingesetzt wird. In der genannten Arbeit wird die Modellierung von gasförmigen Emissionen (*HC*, *CO*, *NO*) an einem Ottomotor untersucht. Unter Variation von Drehzahl, Last, Luft-Kraftstoff-Verhältnis und Steuerzeiten werden Simulationsrechnungen durchgeführt und die Ergebnisse mit den Messungen an über 30 stationären Betriebspunkten verglichen. Dabei liegt die Vorhersagegenauigkeit bezüglich *NO* und *HC* in der Regel innerhalb von 20 %, die Prädiktion von *CO* hingegen liegt oft um 30 % oder mehr außerhalb der Messergebnisse, was unter anderem auf lokale Inhomogenitäten der Mischung zurückgeführt werden kann. Obwohl Ansätze beschrieben werden, wie dieser Herausforderung begegnet werden könnte (beispielsweise durch eine präzise Vorhersage der Kraftstoffoxidation im späten Brennverlauf und einer Kopplung des *CO*- und *HC*-Modells), wird ersichtlich, dass trotz hohem Aufwand auch im stationären Betrieb zusätzliche Untersuchungen und Entwicklungen für eine präzise Vorhersage der Emissionen erforderlich sind. Als weiterführende Literatur zur physikalisch-phänomenologischen Modellierung sind unter anderem die Arbeiten von Janssen [79], Frommater [80] und Bajwa et al. [81] zu nennen.

## 3.3 Datenbasierte Modellierung in der Antriebsentwicklung

Datenbasierte Modelle, die Maschinelles Lernen und allgemein „Künstliche Intelligenz“ nutzen, fanden in der Antriebsentwicklung verglichen mit physikalisch-phänomenologischen Ansätzen in der Vergangenheit weniger Beachtung. In den letzten Jahren stellen sie hingegen eine effiziente Alternative dar, um komplexe Systeme wie Verbrennungsmotoren zu modellieren und zu optimieren.

Die Vorteile datenbasierter Ansätze liegen insbesondere in ihrer Fähigkeit, aus großen Datenmengen Muster und Beziehungen zu extrahieren, die für die Steuerungs- und Regelungsaufgaben in Motorsteuergeräten von Bedeutung sind. Diese Methoden ermöglichen eine effiziente und genaue Abbildung von Motorcharakteristiken, selbst wenn die zugrunde liegenden physikalischen Prozesse komplex oder nicht vollständig verstanden sind. Datenbasierte Modelle werden häufig zur Modellierung von Teilaspekten des Motors eingesetzt, wie beispielsweise zur Regelung des Luft-Kraftstoff-Verhältnisses [82] oder der Prädiktion der AGR-Rate [83]. Im Rahmen der Kommunikation bzw. des Datenaustausches in der Fahrzeugtechnik wird der

Einsatz von Künstlicher Intelligenz ebenfalls erforscht, Lutchen et al. [84] beschäftigen sich beispielsweise mit der Analyse von CAN-Botschaften.

Angesichts der steigenden Anforderungen an die Reduktion von Schadstoffemissionen und der zunehmenden Komplexität von Abgasnachbehandlungssystemen sind diese Methoden auch für die Modellierung von Emissionen interessant. Unter den datenbasierten Methoden gibt es eine Vielzahl möglicher Ansätze, wobei die genaue Unterteilung bereits in Kapitel 2.3 weiter ausgeführt wurde. Papaioannou et al. [85] stellen in ihrer Arbeit eine Modellierung basierend auf einem Random Forest Algorithmus vor, um die Partikelanzahl und -größe in einem hochaufgeladenen Ottomotor mit Benzindirekteinspritzung abzubilden. Shin et al. [86] vergleichen die Eignung zweier Deep Learning Algorithmen (FNN und LSTM), um die $NO_x$-Emissionen eines Dieselmotors im WLTP-Zyklus vorherzusagen. Es zeigt sich, dass mit beiden Modellansätzen eine hohe Genauigkeit ($R^2 > 0.96$) erreicht werden kann. Huang et al. [87] nutzen datenbasierte Methoden, um mit drei Eingangsparametern (Zündzeitpunkt, Drehzahl und Luft-Kraftstoff-Verhältnis) eine Vielzahl an Größen, welche die Verbrennung charakterisieren, eines Nutzfahrzeug Erdgasmotors zu modellieren. Darunter zählen unter anderem der Spitzendruck, der maximale Druckgradient, der indizierte Mitteldruck, aber auch Rohemissionswerte wie beispielsweise Stickstoffoxide, Kohlenstoffoxide und unverbrannte Kohlenwasserstoffe. Fang et al. [88] fokussieren sich bei der Modellierung der Stickstoffoxidemissionen eines Dieselmotors auf den Einfluss der Eingangsparameter auf die Prädiktionsgenauigkeit und die Effizienz zweier Optimierungsalgorithmen.

Zusammengefasst gibt es eine hohe Bandbreite an Anwendungsmöglichkeiten für datenbasierte Methoden in der Antriebsentwicklung, was sich besonders in den letzten Jahren an einer Vielzahl von Veröffentlichungen zeigt. Neben der Flexibilität hinsichtlich des Untersuchungsschwerpunktes bieten auch die unterschiedlichen Methodiken (Shallow Learning, Neuronale Netze etc.), welche wiederum regelmäßig durch neue Entwicklungen erweitert werden, ein ständiges Verbesserungspotenzial.

## 3.4 Hybride Modellierungsansätze

In den vorangehenden Kapiteln wurden die Vorteile und Limitationen sowohl physikalisch-phänomenologischer als auch datenbasierter Ansätze erörtert. Daraus ergibt sich die logische Schlussfolgerung, dass die Entwicklung hybrider Modelle potenzielle Vorteile bieten könnte.

Diese kombinierten Modelle zielen darauf ab, die Stärken beider Ansätze zu vereinen und gleichzeitig deren individuelle Schwächen zu minimieren. Durch die Integration der robusten, physikalischen Grundlagen der phänomenologischen Modelle mit der Flexibilität und Effizienz der datenbasierten Methoden können hybride Modelle eine verbesserte Genauigkeit und Zuverlässigkeit in der Vorhersage und Analyse komplexer Systeme erreichen.

Hinsichtlich der hybriden Modellierungsansätze gibt es zwei Hauptrichtungen, welche für den Fokus dieser Arbeit relevant sind. Dies sind die physikalisch informierten Neuronalen Netze (PINN) und die sogenannten Grey-Box-Modelle. Erstgenannte sind Neuronale Netze, welche an einer oder verschiedenen Stellen physikalische Informationen, beispielsweise in Form von Energieerhaltungsgleichungen in der Optimierungsfunktion, implementiert haben und dadurch die Informationsdichte erhöhen können. Die Grey-Box-Modelle hingegen können aus zwei klar getrennten und auch strukturell verschiedenen Modellen bestehen, deren Ausgänge beziehungsweise Prädiktionen auf mehrere Arten kombiniert werden. Dabei wird der Name aus der Kombination aus Black-Box (datenbasiertes Modell) und White-Box (physikalisch-phänomenologisches Modell, dessen Verhalten stets interpretierbar ist) gebildet.

Auch in der Antriebstechnik werden beide Ansätze bereits genutzt. Nazoktabar et al. [89] präsentieren eine Kombination aus physikalisch-phänomenologischen und datenbasierten Modellen, um einen Regler für einen Motor mit homogener Kompressionszündung (HCCI) zu erstellen und zu optimieren. Dieser ist darauf ausgelegt, die Parameter, welche die Verbrennung beeinflussen, zu bestimmen, um eine Verbrennungsschwerpunktlage zu erreichen, die sowohl eine hohe thermische Effizienz als auch einen niedrigen Emissionsausstoß kombiniert. Kernkomponente dieses Reglers ist ein HCCI-Modell, das physikalisch-phänomenologische Modelle nutzt, um die Verbrennungsschwerpunktlage zu bestimmen. Diese fungiert (unter anderem) als Eingangsgröße für eine Neuronales Netzwerk, welches die Emissionsentstehung prädiziert. Dadurch werden die Stärken beider Methoden kombiniert. Das physikalisch-phänomenologische Modell liefert fundierte Informationen über den Verbrennungsvorgang, wohingegen das auf maschinellen Methoden basierte Teilmodell die komplexe und hochnichtlineare Emissionsentstehung abbildet.

Zhao et al. [90] demonstrieren in ihrer Untersuchung, wie physikalisch informierte Neuronale Netze genutzt werden können, um das Verhalten einer Asynchronmaschine im hochfrequenten Betrieb zu beschreiben. Jede Phase der E-Maschine wird dabei über ein Ersatzschaltbild mit je 18 Elementen abgebildet und die Parametrisierung dieser Elemente über ein Neuronales Netzwerk dargestellt. Die Verlustfunktion ist so gestaltet, dass die vom Neuronalen Netzwerk errechneten Ausgänge in die Ersatzschaltung eingesetzt und damit physikalische Größen – wie etwa die Impedanz – errechnet werden, welche mit Messungen an der realen Maschine in einem Frequenzbereich bis 30 MHz abgeglichen werden. Der beschriebene Aufbau ist in Abbildung 3-1 dargestellt.

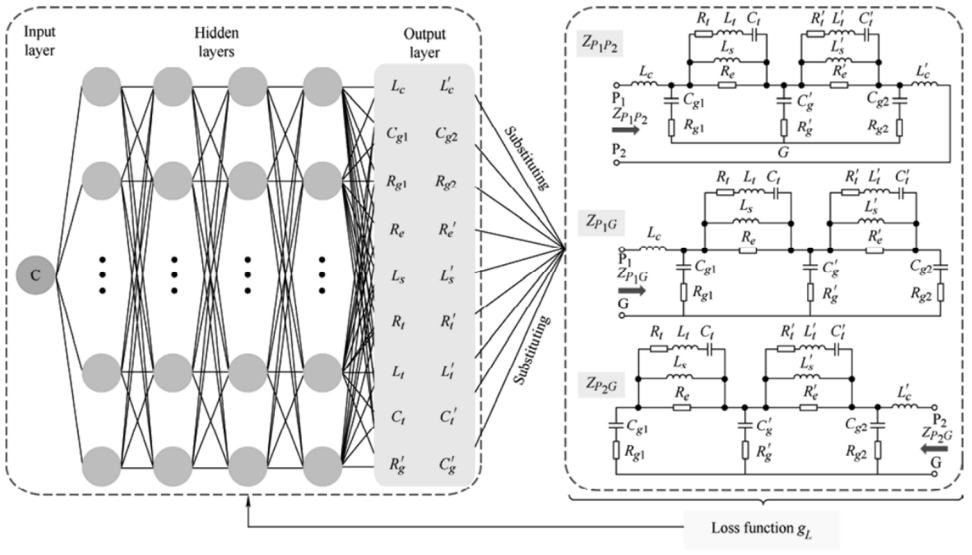


Abbildung 3-1: Physikalisch informiertes Neuronales Netzwerk nach [90]

Dadurch lässt sich ein komplexes physikalisches Modell sehr effizient parametrieren. Dies stellt einen weiteren Anwendungsfall hybrider Modellierungsmethoden dar und verdeutlicht das Potenzial dieser Methodik. Einen guten Überblick über die jüngere Entwicklung der physikinformierten Neuronalen Netze liefert die Arbeit von Cuomo et al. [91]. Dort werden die verschiedenen Ausprägungen und Anwendungsfelder beschrieben und ein Ausblick bezüglich möglicher zukünftiger Entwicklungsschwerpunkte gegeben.

Lang et al. [92] kombinieren physikalisch-phänomenologische und datenbasierte Methoden für die präzise Abbildung des Verbrennungs- und Emissionsverhaltens eines Dieselmotors. Dabei werden der Ladungswechsel und die vorherrschenden Umgebungsbedingungen mit einem physikalisch-phänomenologischen Modell simuliert, der Hochdruckteil des Arbeitsprozesses inklusive der Emissionsbildung wird über einen datenbasierten Ansatz errechnet. Daraus ergeben sich Vorteile gegenüber rein physikalisch-phänomenologischer Simulation, und auch transiente Vorgänge (beispielsweise Lastsprünge) können präzise dargestellt werden.

.

# 4 Ziel der Arbeit

Das vorrangige Ziel dieser Arbeit ist die Entwicklung innovativer Methoden zur Modellierung von Emissionen im hochtransienten Motorbetrieb. Dadurch sollen effiziente Werkzeuge geschaffen werden, welche sowohl für Optimierungs- als auch für Regelungsfunktionen in der modernen Fahrzeugtechnik eingesetzt werden können. Dabei liegt ein projektspezifischer Fokus auf der Nutzung der Emissionsmodelle für die Optimierung der Betriebsstrategie von Hybridfahrzeugen. Ein besonderes Augenmerk wird außerdem auf die Nutzung und Evaluierung neuer Möglichkeiten der datenbasierten Modellierung und des Maschinellen Lernens gelegt. Diese fortschrittlichen Techniken in Zusammenhang mit der exponentiell steigenden Verfügbarkeit von Rechenkapazität bieten das Potenzial, die Genauigkeit und Effizienz der Emissionsmodellierung erheblich zu verbessern.

Zur Erreichung dieses Ziels werden ebenfalls bekannte physikalisch-phänomenologische Ansätze eingesetzt. Sie dienen als Vergleichsbasis und ermöglichen es, die Leistungsfähigkeit und Präzision der neuen datenbasierten Methoden im Kontext bestehender Modellierungspraktiken zu bewerten. Ein wesentlicher Bestandteil und maßgebliches Innovationsmerkmal der Arbeit liegt in der Kombination beider Modellierungsmethoden. Durch die Integration physikalischer Informationen in die datenbasierten Ansätze soll die Informationsdichte erhöht werden, wodurch eine - für den jeweiligen Anwendungsfall spezifizierte - Vorhersagegenauigkeit bei gleichzeitig reduziertem Zeit- und Kostenaufwand ermöglicht werden kann.

Obwohl der Schwerpunkt dieser Arbeit auf der Modellierung von Emissionen aus Verbrennungsmotoren liegt, soll der entwickelte methodische Ansatz nicht darauf beschränkt sein. Die angewandten Techniken und Modelle eröffnen die Möglichkeit einer Übertragung auf andere physikalische, zeitlich schnell veränderliche Vorgänge mit entsprechenden Anpassungen. Mit der Fokussierung auf die Abgasemission soll demonstriert werden, dass der entwickelte Ansatz in der Lage ist, komplexe und dynamische Systeme effektiv zu modellieren, und somit einen bedeutenden Beitrag zur Reduktion von Umweltbelastungen leisten kann.

# 5 Experimentelle Methodik

Das Kapitel „Experimentelle Methodik“ umfasst sowohl die Versuchsaufbauten inklusive der eingesetzten Hardware als auch die Versuchsdurchführung und die dabei gewonnenen Daten. Letztere bilden die Grundlage für die Modellbildung in Kapitel 6.

Ausgehend vom Aufbau des Motors auf dem hochdynamischen Motorenprüfstand und der Bestückung mit Standard- und Sondermesstechnik wurden die Versuchsreihen im Projektverlauf komplexer und die Versuchsumgebung wuchs im gleichen Maße mit, um die Potenziale der datenbasierten Methoden ausnutzen zu können. Entsprechend beginnen die Versuche mit stationären Kennfeldmessungen (Kap. 5.2), worauf auf Straßenfahrten basierende dynamische Untersuchungen (Kap. 5.3) auf dem vollautomatisierten, transienten Prüfstand folgen.

## 5.1 Prüfstandsversuch

Die Versuche am hochdynamischen Motorenprüfstand des LAF bilden die Basis der vorliegenden Arbeit. Dieser wird im weiteren Verlauf näher beschrieben, wobei auch der Versuchsträger und die eingesetzte relevante Messtechnik betrachtet werden.

### 5.1.1 Versuchsträger

Als Versuchsträger wurde ein BMW B48B20M0 Reihenvierzylinder-Ottomotor mit Benzindirekteinspritzung, Twinscroll-Turboaufladung, VALVETRONIC-Ventiltrieb und 2.0 l Hubraum eingesetzt. In Tabelle 5-1 sind wichtige technische Daten nach [93] aufgeführt.

Tabelle 5-1: Technische Daten BMW B48B20M0 nach [93]

| | **Wert** | **Einheit** |
|---|---|---|
| Nennleistung | 141 | kW |
| Nenndrehzahl | 5000 - 6000 | $min^{-1}$ |
| Maximales Drehmoment | 280 | Nm |
| Hubvolumen | 1.998 | $dm^3$ |
| Hub | 94.6 | mm |
| Bohrung | 82 | mm |
| Verdichtungsverhältnis | 11.0:1 | - |

Der Motor ist Teil einer übergeordneten Motorenfamilie, welche aus 3-, 4- und 6-Zylindermotoren besteht und sowohl die ottomotorischen als auch die dieselmotorischen Brennverfahren umfasst. Die primären Entwicklungsziele hierbei waren ein einheitliches konstruktives Grundmotorenkonzept, Weiterentwicklung der Brennverfahren, verbesserte Einspritztechnik, optimiertes Wärmemanagement und die Reibungsreduzierung. Auch der B48B20M0 verfügt über einen markentypischen Zylinderabstand von 91 mm und ein Einzelzylinderhubvolumen von 0.5 dm³. Eine Besonderheit der Twinscroll-Turboaufladung stellt die im Abgaskrümmer integrierte Turbinenseite dar. Das entwickelte Brennverfahren profitiert von einer gesteigerten Ladungsbewegung bei großen Ventilhüben und einer großflächigen Sprayauslegung der Mehrloch-Injektoren. Dadurch wurde die Homogenisierung des Kraftstoffgemisches gegenüber den Vorgängermotoren signifikant verbessert und in weiten Kennfeldbereichen ist ein Betrieb mit stöchiometrischem Luft-Kraftstoff-Verhältnis möglich. [93]

In Abbildung 5-1 ist der einsatzbereite und mit Messtechnik (detailliert in Kapitel 5.1.2) ausgestattete Versuchsmotor (Markierung 1) nach Aufbau auf dem hochdynamischen Motorenprüfstand dargestellt. Über ein Prüfstandsgetriebe (Markierung 2) mit direkter Übersetzung (1:1) und einem Drehmomentmessflansch ist der B48B20M0 mit einer Asynchronmaschine verbunden, welcher in beide Drehrichtungen einen sehr schnellen Drehmomentaufbau und -abbau mit positivem und negativem Vorzeichen ermöglicht. Aufgrund dieser Tatsache können bei realen Straßenfahrten auftretende dynamische Einsatzprofile inklusive der Schaltvorgänge exakt abgebildet werden. Die Kraftstoffversorgung erfolgt durch eine Kraftstoffkonditionieranlage (Markierung 3), welche den gewünschten Vorförderdruck und die Kraftstofftemperatur einregelt. Der Versuchsmotor besitzt einen im Saugrohr integrierten Ladeluftkühler mit einem Luft-Wasser-Wärmetauscher, wodurch die Ladeluftstrecke sehr kurz gehalten werden kann. Für eine signifikante Abkühlung der Ladeluft steht neben dem Hochtemperaturkühlkreislauf (für den Motorblock, Zylinderkopf und die Turboladerlagerung) noch ein Niedertemperaturkühlkreis für diese Anwendung zur Verfügung. Die beiden Kühlkreisläufe werden im Fahrzeug über Wärmetauscher von der Außenluft gekühlt. Im Prüfstandsaufbau sind dahingegen zwei Plattenwärmetauscher (Markierung 4) vorhanden, über welche Wärme an das Kühlsystem des

Labors abgegeben werden kann. Die räumliche Führung der Abgasanlage (Markierung 5) musste zwar an die Platzverhältnisse im Prüfstandsraum angepasst werden, entspricht hinsichtlich der Komponenten und deren Reihenfolge aber der im Gesamtfahrzeug vorhandenen Konfiguration. So wird sichergestellt, dass sich ein vergleichbares Verhalten des Abgasgegendrucks ausbilden kann. Die Temperatur der Abgasnachbehandlung und deren Wirkung spielt dahingegen eine untergeordnete Rolle, da der Fokus dieser Arbeit auf den Rohemissionen liegt. Über einen Bedienraum (Markierung 6) können die Prüfläufe in sicherer Umgebung durchgeführt und überwacht werden.

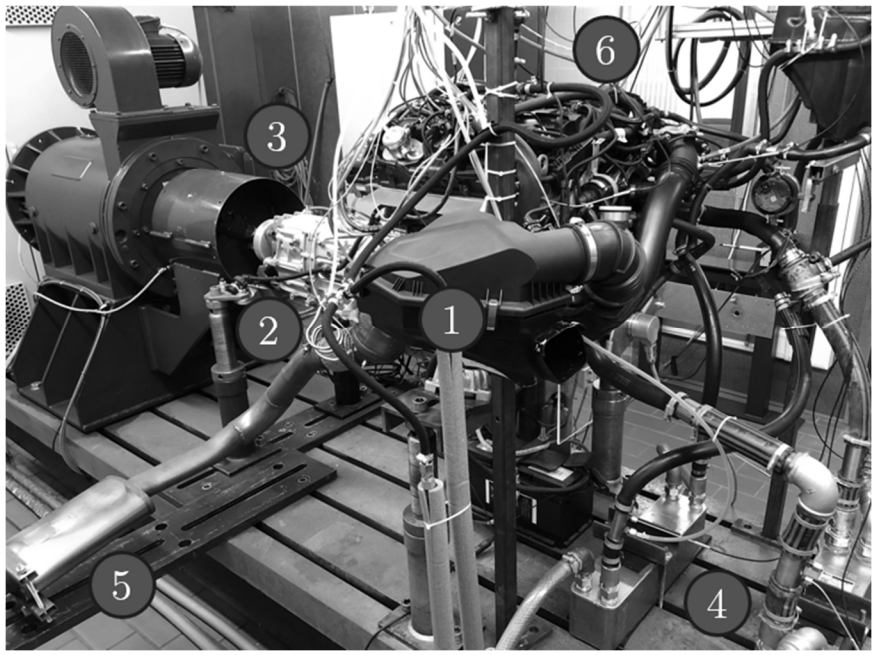


Abbildung 5-1: Versuchsmotor auf dem hochdynamischen Motorenprüfstand

Da im Rahmen des Projektes, in dem wesentliche Teile der vorliegenden Arbeit entstanden sind, auch ein Gesamtfahrzeug mit identischem Verbrennungsmotor (BMW 530e) eingesetzt wurde, um reale Lastprofile von Fahrversuchen zu erhalten (siehe Kapitel 5.3), musste auch beim Prüfstandsmotor darauf geachtet werden, ein Motorsteuergerät (engl. „Engine Control Unit“ ECU) mit Serienstand zu verwenden. Somit kann sichergestellt werden, dass sich beide Motoren in identischen Lastzuständen auch vergleichbar verhalten. Die serienmäßige ECU dieses Motors ist (wie bei vielen weiteren modernen Motoren) über mehrere Datenleitungen (Bussysteme) mit anderen Systemen des Fahrzeuges – im vorliegenden Fall unter anderem mit dem Getriebe – verbunden und auf diese Kommunikation angewiesen, um einwandfrei zu funktionieren. Über die Datenleitungen werden Informationen ausgetauscht, die entweder funktional wichtig sind oder sicherheitskritische Zustände vermeiden sollen. So wird beispielsweise überprüft, dass vor der Startfreigabe kein Kraftschluss zwischen Motor und Antriebsachse(-n) besteht.

Da im Prüfstandsaufbau nicht alle Subsysteme eines Fahrzeuges in physischer Form vorliegen, muss der Datenaustausch inklusive plausibler Sensorwerte simuliert werden. Dies geschieht über eine sogenannte Restbussimulation. Hierzu wurde eine Zustandsmaschine im Prüfstandsautomatisierungssystem erstellt, welche in Abhängigkeit der möglichen Fahrzeugzustände (Geparkt, Zündung ein, Fahren, etc.) die richtigen Signale generiert und an das Motorsteuergerät sendet.

### 5.1.2 Messtechnik

In diesem Kapitel liegt der Fokus auf der am Motorenprüfstand eingesetzten Messtechnik. Diese ist für die Überwachung und Sicherheit des Versuchsmotors inklusive der unmittelbaren Umgebung wichtig und erlaubt die Einhaltung reproduzierbarer Messbedingungen. Ferner liefern die Sensoren die relevanten Daten für die anschließende Modellierung. Abbildung 5-2 stellt schematisch dar, welche Messgeräte und dazugehörigen erfassten Messgrößen für die Messungen am Prüfstand von entscheidender Bedeutung sind.

Zu Gunsten der Übersichtlichkeit wird dabei zwischen Standard- und Sondermesstechnik unterschieden. Zur Ersteren gehören mit regulärer Messfrequenz (5 Hertz in diesem Aufbau) ausgewertete Druck- und Temperatursensoren, die in nahezu jedem Prüfstandsaufbau zu finden sind. Diese sind Teil der Grundausstattung und werden im Folgenden nur am Rande betrachtet und in der Abbildung nicht dargestellt. Insgesamt werden am Motorenprüfstand 146 Messgrößen erfasst. Die Sondermesstechnik wird im Folgenden detaillierter beschrieben.

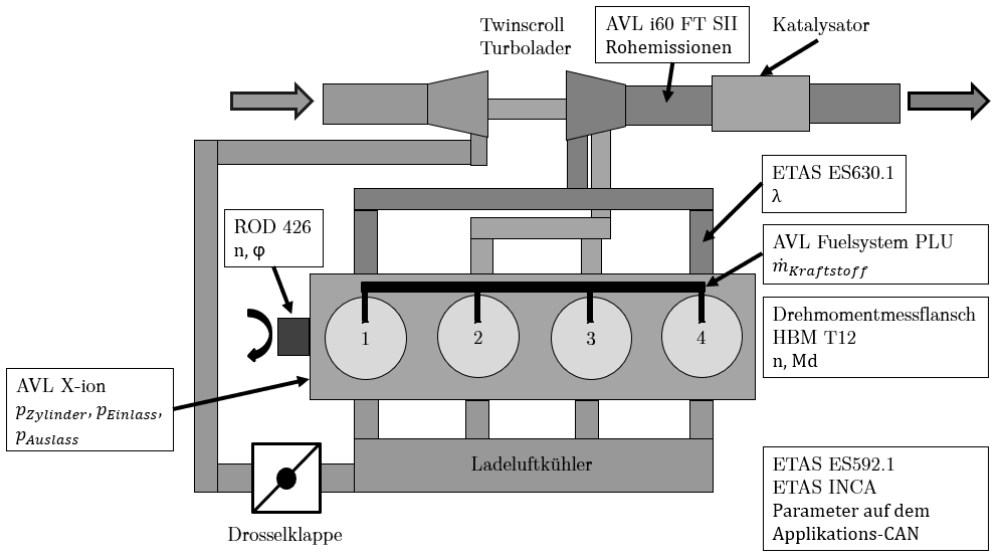


Abbildung 5-2: Auf die Sondermesstechnik reduzierter Messstellenplan

Über ein ETAS ES592.1 Interface kann per Applikations-CAN auf das Motorsteuergerät zugegriffen werden. Mit den dazugehörigen Beschreibungsdateien (.dbc und .a2l Datei) und der Mess-, Kalibrier- und Diagnosesoftware ETAS INCA sind ein Auslesen von Steuergerätegrößen, des Fehlerspeichers und verschiedene Kalibrierungsfunktionen möglich. Da der Fokus dieser Arbeit auf einem seriennahen Betrieb liegt, werden primär Messwerte aus der ECU ausgelesen (Ventilhub, Einspritzzeitpunkt, Ventilspreizung, etc.) und über eine ASAM-Schnittstelle (engl. „Association for Standardisation of Automation and Measuring Systems“ ASAM) an das Prüfstandsautomatisierungssystem AVL Puma 2 gesendet, welches die Zusammenführung und Abspeicherung der Messgrößen zentralisiert. Die Erfassung der Motordrehzahl und des effektiven Drehmoments erfolgt über einen Drehmomentmessflansch HBM T12, welcher zwischen Getriebeausgang und der Asynchronmaschine vom Typ Schorch IEC 250M sitzt. Dieser dient ebenfalls der Prüfstandsautomatisierung zur Regelung der E-Maschine und des Verbrennungsmotors. Zusätzlich werden die Drehzahl und der Drehwinkel der Kurbelwelle in 0.1 °KW - Schritten mit einem Drehwinkelgeber (ROD 426) erfasst. Dieser ist über eine Kupplung an der Riemenscheibe der Kurbelwelle befestigt. Die hochpräzise Erfassung der Kurbelwellenposition ist entscheidend für die Indizierung. Mit dem System AVL X-ion werden die Zylinderdrücke, der Druck im Sammler und der Druck im Abgaskrümmer hochdynamisch gemessen, um die Verbrennung analysieren zu können. Hierzu werden pro Messung 100 Verbrennungszyklen erfasst und anschließend gemittelt, um regulär auftretende Verbrennungsschwankungen ausgleichen zu können. Die Kraftstoffversorgung erfolgt über eine Kraftstoffkonditionieranlage (AVL Fuelsystem PLU), welche den Kraftstoff (Super Plus) unter stets gleichbleibenden Bedingungen (20 °C, 5 bar Relativdruck) zur motoreigenen Hochdruckpumpe fördert. Dieses System ersetzt die im Fahrzeug vorhandene Innentankpumpe und erhöht zusätzlich die Reproduzierbarkeit. Die Rohemissionen werden mit einem AVL i60 FT SII gemessen, welches nach dem Prinzip eines Fourier-Transformations-Infrarotspektrometers funktioniert. Direkt nach dem Turbinenausgang und vor dem Katalysator befindet sich die Entnahmestelle, sodass von einer guten Durchmischung und der Berücksichtigung aller Zylinder ausgegangen werden kann. Das Messgerät leitet einen Teilstrom des Abgases über eine beheizte Leitung zur Messzelle und verfügt über eine Pumpe mit hohem Volumenstrom, weshalb auch dynamische Effekte durch kurze Signalanstiegszeiten berücksichtigt werden können („fast response“). Dennoch ist die Gaslaufzeit ein relevanter Faktor und wird detailliert in Kapitel 6.2.2 beschrieben.

Für die physikalisch-phänomenologische Modellierung in Kapitel 6.1 ist die Druckverlaufsanalyse des Brennraumes und der Ein- und Auslasskanäle entscheidend. Wenn die Betrachtung auf einen einzelnen Zylinder beschränkt werden kann und die Querbeeinflussung durch die anderen Zylinder bezüglich des Ein- und Auslassdrucks so gering wie möglich ist, lässt sich dies effektiv umsetzen. Besonders Zylinder 4 eignet sich durch die Zugänglichkeit auf der Abgasseite für eine solche Einzelbetrachtung. Deshalb wurden (siehe Abbildung 5-3) eine Temperaturmessstelle (Markierung 1), ein gekühlter hochdynamischer Drucksensor (Markierung 2) und eine zusätzliche Lambdasonde (Markierung 3) so nah wie möglich hinter den Auslass von Zylinder 4 in den Krümmer montiert. Durch das Twinscroll-Design und die Zündfolge

werden die Signale an der gewählten Position nur geringfügig von Zylinder 1 beeinflusst, was in der Datenverarbeitung durch die Kenntnis des Drehwinkels korrigiert werden kann.

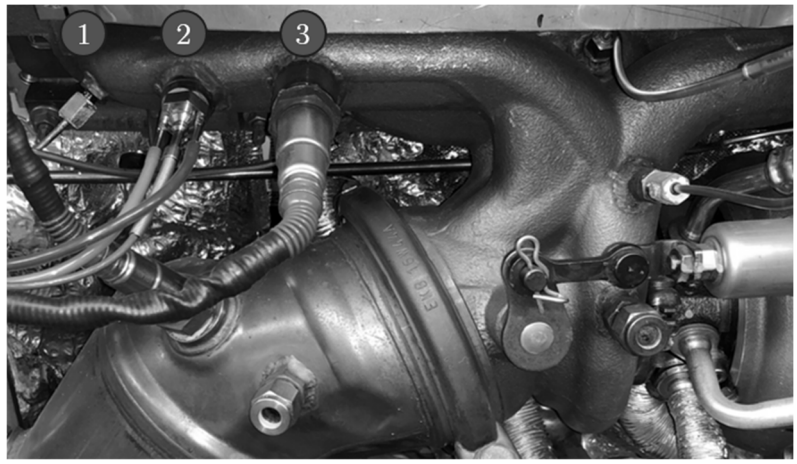


Abbildung 5-3: Messstellen am Krümmer für die Modellbildung

## 5.2 Stationäre Kennfeldmessungen

Der Versuchsmotor wurde dem LAF vom Hersteller als neu und bisher nicht gelaufen zur Verfügung gestellt. Daher musste nach der Inbetriebnahme und umfassenden Funktionstests zuerst ein Einlaufprogramm absolviert werden, um die volle Leistungsfähigkeit des Motors nutzen zu können.

Als erste Messkampagne erfolgte im Anschluss die Vermessung des Kennfeldes unter stationären Bedingungen. Das bedeutet, dass einzelne Drehzahl-/Lastkombinationen über einen längeren Zeitraum stabil gehalten werden, sodass unter anderem physikalischen Rahmenbedingungen, wie beispielsweise die Temperaturen der Komponenten, ein stationäres Zielniveau erreichen. Die Dauer kann je nach Betriebspunkt variieren, als minimale Haltezeit wurden jedoch fünf Minuten festgelegt, auch wenn bereits vorher ein Einschwingen der Messwerte erkennbar war. Die Mindestanzahl der Betriebspunkte wurde, basierend auf den Erfordernissen der Druckverlaufsanalyse (TPA) und der Kalibrierung prädiktiver Verbrennungsmodelle (ab Kapitel 6.1) laut Empfehlungen des Herstellers der verwendeten Software GT-Suite [94] auf 25 festgelegt. Weiterhin lieferten die im Gesamtfahrzeugversuch (Kap. 5.3) gewonnenen Daten einen Aufschluss über die im realen Betrieb häufig vorkommenden Kennfeldbereiche, welche folglich auch im Prüfstandsversuch als besonders relevant eingestuft wurden.

Durch eine feine Rasterung dieser Bereiche und einer Hinzunahme der Volllastlinie ergibt sich ein erweitertes stationäres Messprogramm mit 66 Betriebspunkten, welches in Abbildung 5-4

dargestellt ist. Die Erhöhung der Messpunkte führt zwar zu einer längeren Messdauer, erfahrungsgemäß kann es jedoch bei der Kalibrierung von Verbrennungsmodellen hilfreich sein, die Datenbasis zu erhöhen und bei Bedarf auf nicht benötigte Stützpunkte zu verzichten. Ebenso ist die Volllast ein im Straßenbetrieb nur in Ausnahmesituationen erreichter Betriebszustand, die Aufnahme dieser Daten ist aber gleichzeitig ein essenzieller Schritt, um durch den Abgleich mit Kennwerten des Herstellers den Gesamtaufbau zu validieren.

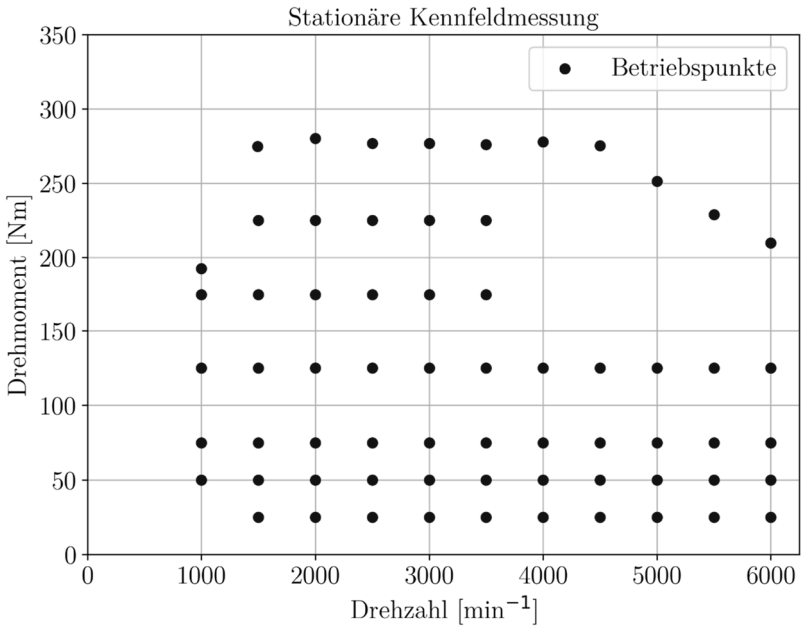


Abbildung 5-4: Betriebspunkte der stationären Kennfeldmessung

In Abbildung 5-4 ist erkennbar, dass es oberhalb von 3500 min$^{-1}$ und mehr als 50 % Drehmoment einen Bereich gibt, in welchem keine stationären Messungen durchgeführt wurden. Dieser Kennfeldbereich wurde in den Straßenfahrten nur in Einzelfällen genutzt und ist dementsprechend als Datengrundlage für Bildung von Emissionsmodellen wenig relevant.

Bei den stationären Kennfeldmessungen wurden die 146 Messgrößen über einen Zeitraum von einer Minute mit 5 Hz Auflösung aufgenommen und anschließend gemittelt. Hierzu zählen unter anderem:

- Emissionen
- Drücke
- Temperaturen
- Steuergerätegrößen
- Drehzahl/Drehmoment

Zusätzlich wurden die Zylinderdruckverläufe und der Druck im Sammler sowie im Krümmer mit dem Indiziersystem über 100 Verbrennungszyklen erfasst und ebenfalls gemittelt. Diese Daten bilden die Grundlage für die physikalisch-phänomenologische Modellierung.

## 5.3 Ableitung und Messung realer Lastprofile am hochdynamischen Motorenprüfstand

Im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Projektes „ML-MoRE“ (Maschinelles Lernen für die Modellierung und Regelung der Emissionen von Hybridfahrzeugen in Realfahrzyklen; Förderkennzeichen 01|S20007A-C) konnten wichtige Datengrundlagen der vorliegenden Arbeit geschaffen werden. Unter anderem stand ein Hybridfahrzeug (BMW 530e) zur Verfügung, welches ebenfalls einen Motor vom Typ B48B20M0 einsetzt, siehe Abbildung 5-5. Die vom Projektpartner KST Motorenversuch GmbH durchgeführten Straßenfahrten nach RDE-Kriterien (engl. „Real Driving Emissions“ RDE [4]) beinhalteten die Messung der Abgasemissionen mit einem portablem Emissionsmesssystem (PEMS). Zusätzlich zum regulären Hybridbetrieb konnte der Fahrmodus so gewählt werden, dass ein rein verbrennungsmotorischer Betrieb möglich war.


Abbildung 5-5: BMW 530e ausgestattet für portable Emissionsmessungen, Bild bereitgestellt durch KST Motorenversuch GmbH

Die Analyse der im Fahrzeug erhobenen Messdaten erfolgte mit dem Ziel, Fahrmanöver mit hohem Emissionspotential zu identifizieren. Die direkte Verwendung der Emissionsmessdaten für die Modellierung wurde jedoch nicht in Betracht gezogen, da hierzu auf die Vorteile der stationären Prüfeinrichtung vertraut wurde. Dazu zählt unter anderem die bessere Reproduzierbarkeit der Versuche im Labor aufgrund der besseren Kontrolle über die Randbedingungen (kein Verkehr, Wettereinflüsse, etc.). Weiterhin kann die stationäre Messtechnik in der Regel

performanter gestaltet werden, da Restriktionen mobiler Messtechnik entfallen (Größe, Gewicht, Vibrationen). Im Vergleich liegt die Messfrequenz des portablen Emissionsmessgerätes bei 1 Hz, das im Labor verwendete AVL i60 FT SII erreicht die fünffache Auflösung und kann mehr Emissionsspezies gleichzeitig erfassen. Die höhere Erfassungsfrequenz ist im Kontext der vorliegenden Arbeit besonders wichtig, da im hochtransienten Betrieb erhöhte Emissionen zu erwarten sind ([95], [96]) und deren modellhafte Darstellung eine Möglichkeit zur Reduktion eröffnen kann.

Um eine exakte Nachbildung der RDE-Lastprofile am Prüfstand zu ermöglichen, sind spezifische Sollwertvorgaben für geeignete Stellgrößen notwendig. Im Versuchsfahrzeug wurden neben den Emissions- und GPS-Daten über das On-Board-Diagnose-System (OBD) auch zusätzliche Parameter wie die Motordrehzahl, das gewünschte Drehmoment, oder die Fahrpedalstellung erfasst. Die Motordrehzahl kann am Prüfstand mithilfe einer Belastungsmaschine sehr präzise und dynamisch geregelt werden, wobei die Regelgeschwindigkeit ausreicht, um auch hochtransiente Schaltvorgänge des Automatik-Getriebes realitätsnah abzubilden. Je nach Regelungsmodus kann die externe Last über ein Solldrehmoment oder resultierend aus einer gewünschten Fahrpedalstellung in Kombination mit der vorgegeben Drehzahl eingestellt werden. Für die Versuche dieser Arbeit wurde die Vorgabe der Fahrpedalstellung präferiert, da dieser Parameter über die im Automatisierungssystem hinterlegte vom Fahrzeughersteller bekannte Fahrpedalkennlinie direkt an das Motorsteuergerät übermittelt werden kann, wodurch sich der Prüfstandsmotor im Vergleich zum Fahrzeugmotor möglichst identisch verhalten sollte. Eine Drehmomentvorgabe könnte durch die damit verbundene Regelung des Fahrpedalsollwerts zu zusätzlichen Ungenauigkeiten führen.

Insgesamt wurden über die beschriebene Methodik mehr als zehn Stunden reale Straßenfahrt am Prüfstand nachgebildet. Die Gesamtdauer setzt sich aus Realfahrten zusammen, welche zwischen 60 und 90 Minuten lang waren, und unterschiedliche Routenverläufe hatten. Die gemessenen Parameter und die Messfrequenz sind identisch mit den stationären Kennfeldmessungen, siehe Kapitel 5.2. In Abbildung 5-6 ist ein Ausschnitt einer am Prüfstand nachgefahrenen RDE-Fahrt dargestellt. Abgebildet sind die Drehzahl des Verbrennungsmotors [$min^{-1}$], das Drehmoment [Nm] und die $NO_x$-Emissionen [ppm] im Rohabgas.

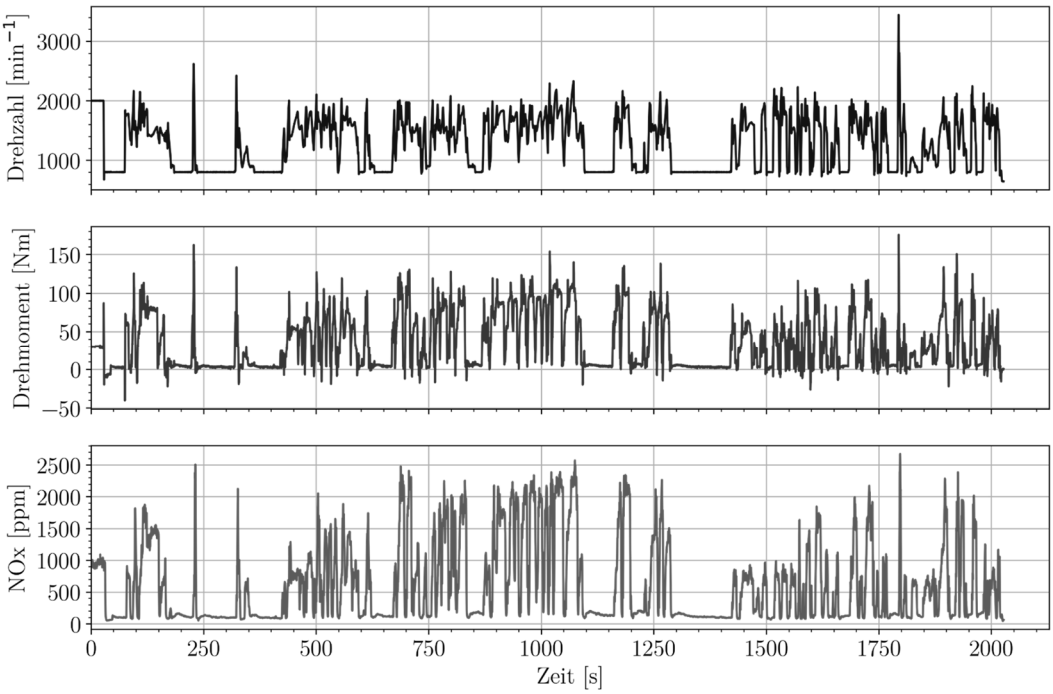


Abbildung 5-6: $NO_x$-Emissionsverlauf während einem nachgefahrenen RDE-Zyklus

Die Abbildung zeigt die ersten circa 2000 Sekunden der Messung. Aufgrund der hohen Anzahl an gemessenen Parametern und der Aufzeichnungsfrequenz von 5 Hertz wurden die Versuche unterteilt, um innerhalb der maximalen Speicherkapazität des Prüfstandsrekorders zu bleiben.

Obwohl die RDE-Fahrten eine Datengrundlage für den zu modellierenden realen Fahrbetrieb liefern, ist es wegen der limitierten Auswahl an erhobenen Messdaten nicht möglich, sämtliche Fahrmanöver (potenziell unendlich viele) umfassend zu repräsentieren. Eine enorme Ausweitung des Umfangs der Messungen und Trainingsdaten wäre zwar denkbar, allerdings auch mit einem beträchtlichen Zusatzaufwand verbunden. Aus diesem Grund wurde ergänzend auf Verfahren der klassischen Systemidentifikation zurückgegriffen (siehe [60], [97], [98]). Diese Methoden basieren auf einer Anregung des Systems mit periodischen Signalen variierender Intensität, um aus den Systemreaktionen Rückschlüsse auf das Systemverhalten zu ziehen. Im vorliegenden Projekt wurde eine Trajektorie generiert, die sich aus 50 Fourier-Koeffizienten zusammensetzt, und zwar sowohl für die Position des Gaspedals (als Laststeller) als auch für die Motordrehzahl. Bei der Konzeption dieser Trajektorie wurde darauf geachtet, ein breites Spektrum innerhalb des Betriebskennfeldes abzudecken, während zugleich kritische Betriebszustände vermieden wurden. Die Anregung des Systems in Form von Drehzahl und Drehmoment und die Systemantwort ($NO_x$) sind in Abbildung 5-7 dargestellt.

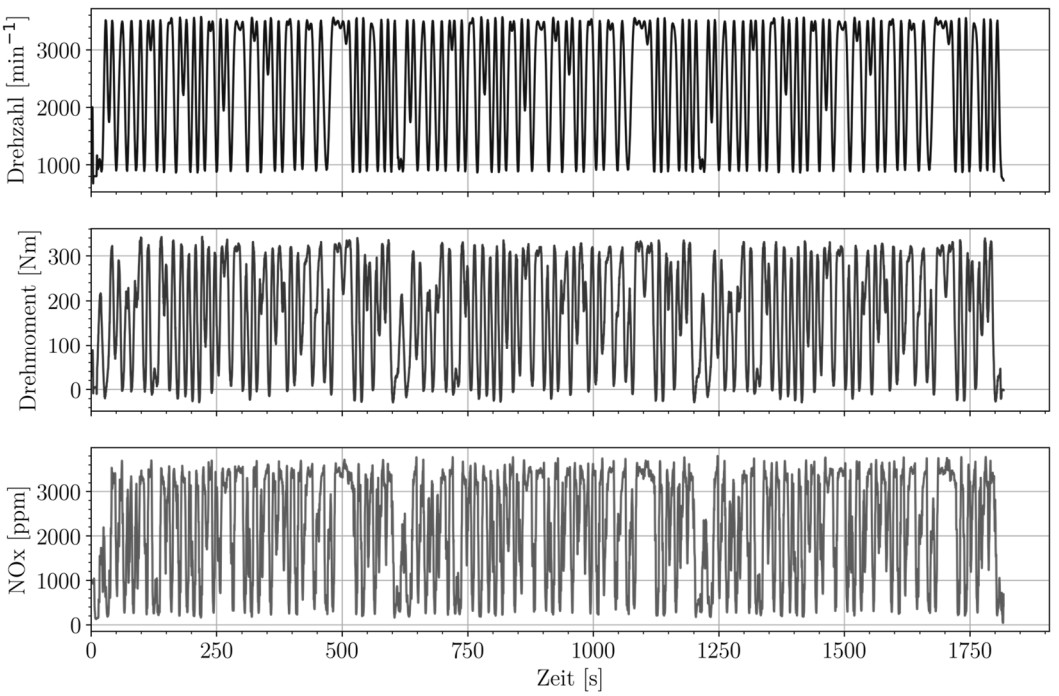


Abbildung 5-7: Lastprofil basierend auf Methoden der klassischen Systemanalyse

Im Lastprofil (Drehzahl und Drehmoment) sind die sich periodisch wiederholenden Sequenzen erkennbar, welche zueinander verschoben und unterschiedlich stark ausgeprägt sind. Dadurch wird erreicht, dass eine hochvariable Systemanregung in möglichst kurzer Zeit erfolgt.

# 6 Modellbildung

In diesem Kapitel werden die Abläufe, Randbedingungen und Herausforderungen bei der Erstellung der drei Haupt-Modellarten „physikalisch-phänomenologisch“ (Kapitel 6.1), „datenbasiert“ (Kapitel 6.2) und „hybrid“ (Kapitel 6.3) erläutert. Dabei werden die notwendigen Daten und Parameter sowie der Optimierungsprozess (bzw. Trainingsprozess bei den datenbasierten Modellen) dargestellt, die Bewertung der Vorhersagegenauigkeit bezüglich der Emissionsentstehung erfolgt im darauffolgenden Kapitel 7.

## 6.1 Physikalisch-phänomenologische Modellierung

Das physikalisch-phänomenologische Modell wurde mit der Software GT-Suite v2022 von Gamma Technologies erstellt. GT-Suite ist als Entwicklungswerkzeug – besonders bei Verbrennungsmotoren – sowohl in der Industrie als auch in der Forschung etabliert und wird seit Jahren am Lehrstuhl für Antriebe in der Fahrzeugtechnik erfolgreich eingesetzt, weshalb es auch für die vorliegende Arbeit ausgewählt wurde.

Für die Berechnungen wurde der lehrstuhleigene Simulationsserver verwendet. Dieser verfügt über einen AMD Ryzen Threadripper 2990WXProzessor mit 32 Kernen und 128 GB Arbeitsspeicher.

Der Ablauf bei der physikalisch-phänomenologischen Modellierung lässt sich in vier Schritte unterteilen:

1. Durchführung einer Druckverlaufsanalyse: Beschränkung auf einen Zylinder - einfache Geometrie - Anpassung des simulierten Druckverlaufes im Brennraum an den gemessenen Druckverlauf → Ableitung des Brennverlaufs und weiterer nicht messtechnisch erfasster Informationen (Restgasgehalt, etc.)
2. Erstellung des prädiktiven Verbrennungsmodells: Grundlage sind die in Schritt 1 analysierten Brennverläufe - Parametrierung eines prädiktiven Verbrennungsmodells (Wärmeübergänge, Turbulenzen, Kraftstoffverdampfung, Gemischverteilung, etc.), sodass sich ein möglichst gut übereinstimmender Brennverlauf einstellt
3. Aufbau und Optimierung der prädiktiven Emissionsmodelle: Basierend auf dem optimierten Verbrennungsmodell und den dadurch gegebenen Randbedingungen (Druck-, Temperaturverlauf, Restgasgehalt, etc.) - Auswahl unterschiedlicher Modelle je Spezies - Abgleich und Parameteroptimierung anhand der Messungen

4. Motormodell: Zusammenführung der Geometrien, des Verbrennungsmodells und der Emissionsmodelle zu einem Gesamtmotormodell mit vier Zylindern

### 6.1.1 Druckverlaufsanalyse

Die Druckverlaufsanalyse wurde auf sämtliche 66 Betriebspunkte der stationären Kennfeldmessung (Kapitel 5.2) angewendet. Das verwendete GT-Suite Modell ist schematisch in Abbildung 6-1 zu sehen. Der Motor wird dabei auf einen Zylinder (Zylinder 4) beschränkt und geometrisch von der Messung des Druckverlaufs im Sammler (ca. 10 cm vor dem Einlassventil) bis zur Messung des Druckverlaufs im Krümmer (Abbildung 5-3) nachgebildet.

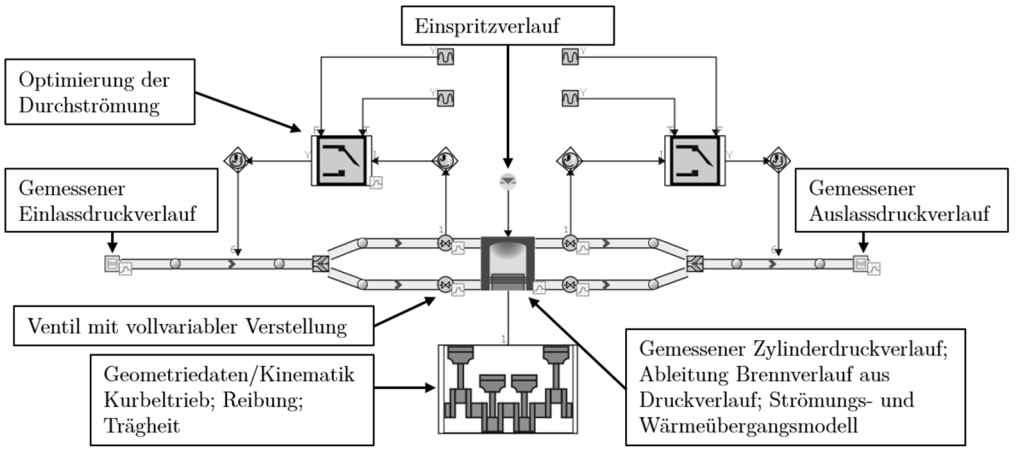


Abbildung 6-1: Aufbau des TPA-Modells

Der Aufbau des Modells wird nachfolgend in Strömungsrichtung (von links nach rechts) beschrieben. Zuerst wird ein Element genutzt, das Luft mit dem gemessenen Druckverlauf und der gemessenen Temperatur als Randbedingung aufprägt. Der Übergang zwischen Sammler und Zylinderkopf ist beim Versuchsmotor noch einflutig und teilt sich erst innerhalb des Zylinderkopfs in zwei Kanäle. Die Kanalgeometrie ist mit Daten des Herstellers und den Ergebnissen eigener Messungen parametriert. Durch die Angabe der Materialien und Radien/Winkel wird zusätzlich Wandreibung berücksichtigt.

Im oberen Bereich in Abbildung 6-1 ist ein vom Softwarehersteller integriertes System zur Optimierung der Durchströmung gekennzeichnet. Wenn ein Ventil schließt, resultiert daraus eine ungewollte zusätzliche Druckschwankung, welche sich dem gemessenen Druckverlauf überlagert und dadurch zu einer Abweichung zwischen Modell und Messung führt. Dies wird verhindert, indem die Kanalreibung in Abhängigkeit der Ventilstellung beeinflusst wird. Bei geöffneten Ventilen wird der Reibungsmultiplikator auf 1 gesetzt, womit die Reibung zwischen

Wand und Medium den parametrierten Material- und Oberflächeneigenschaften entspricht. Wenn das Ventil geschlossen ist, wird die Reibung über den Multiplikator stark erhöht, um die Druckschwankungen zu dämpfen.

Links und rechts des Brennraumes sind die vier Ventile dargestellt. Diese werden über den Referenzdurchmesser, Druckverlustbeiwerte in beide Strömungsrichtungen und die Koeffizienten für den „Swirl“ und „Tumble“ (Rotationsverhalten der Strömung um die vertikale/horizontale Achse) festgelegt. Die Daten stammen zu einem Teil vom Motorhersteller, zu anderen Teilen wurden diese selbst gemessen oder resultieren aus dem Optimierungsprozess. Essenziell bei der Abbildung der Ventile ist der Hubverlauf. Wie in Kapitel 5.1.1 beschrieben, verfügt der Versuchsmotor über einen vollvariablen Ventiltrieb und kann folgende Parameter beeinflussen:

- Maximaler Ventilhub auf der Einlassseite (der Auslasshub ist konstant)
- Spreizung auf der Ein- und Auslassseite
- „Phasing“ auf der Einlassseite: je nach eingestelltem Maximalhub öffnen die beiden Einlassventile unterschiedlich weit, um die Zylinderinnenströmung und damit die Gemischaufbereitung in Abhängigkeit des Betriebspunkts zu optimieren

Die Ventilhubkurven im Kennfeldbereich vollumfänglich abzubilden ist sehr aufwändig, da diese in der Realität in feinen Abstufungen diskretisiert sind. Daher wäre eine umfangreiche Vermessung der Hubkurven auf einem Komponentenprüfstand zeitintensiv und technisch anspruchsvoll. Die Hubverläufe konnten dem LAF vom Hersteller des Motors zur Verfügung gestellt werden. Dies beinhaltet ein eigens entwickeltes Werkzeug, welches die Hubverläufe in Abhängigkeit der vorgegeben Spreizungen und Maximalhübe berechnet und in Abhängigkeit der Kurbelwellenwinkel ausgibt. Der Einlasshub und die Spreizung auf der Einlass- und Auslassseite sind Stellgrößen des Motorsteuergeräts, welche in den gespeicherten Messgrößen enthalten sind. Dadurch lassen sich in jedem erfassten Betriebspunkt die Ventilhubkurven rekonstruieren. Das vorhandene Werkzeug lässt sich nicht unmittelbar in GT-Suite einbinden. Für den stationären Anwendungsfalls wäre es zwar möglich, die jeweilige Ventilhubkurve vorab zu generieren und als festen Verlauf zu hinterlegen, dies ist aber einerseits bereits bei 66 Betriebspunkten aufwändig, andererseits sollte das finale Motormodell auch dynamisch betrieben werden können. Somit wird eine unmittelbare Integration der Berechnung der Ventilhubkurven in die Simulationsumgebung erforderlich.

Die Auslassventile lassen sich einfach steuern, da durch den konstanten Hub eine definierte Ventilhubkurve um den Wert der Spreizung entlang der Achse der Kurbelwellenposition verschoben wird. Das ist in GT-Suite direkt vorgesehen und kann über eine Variable implementiert werden. Die Abbildung der Einlassventilhubkurven wurde im Rahmen der vorliegenden Arbeit entwickelt. Jedes Einlassventil besitzt anstatt einer definierten Hubkurve ein individuelles Kennfeld, aus welchem der Hubverlauf extrahiert wird. Die Kennfelder unterscheiden sich, um das eingangs erwähnte Phasing abbilden zu können. Sie sind ferner bezüglich des maximalen Ventilhubs auf 0.1 mm aufgelöst und stellen die momentane Hubhöhe in 2 °KW-

Schritten dar. Die damit generierten Ventilhubverläufe unterscheiden sich um weniger als 0.00001 mm von den Herstellerwerten. In Abbildung 6-2 ist das implementierte Kennfeld für den Hubverlauf eines Einlassventils dargestellt.

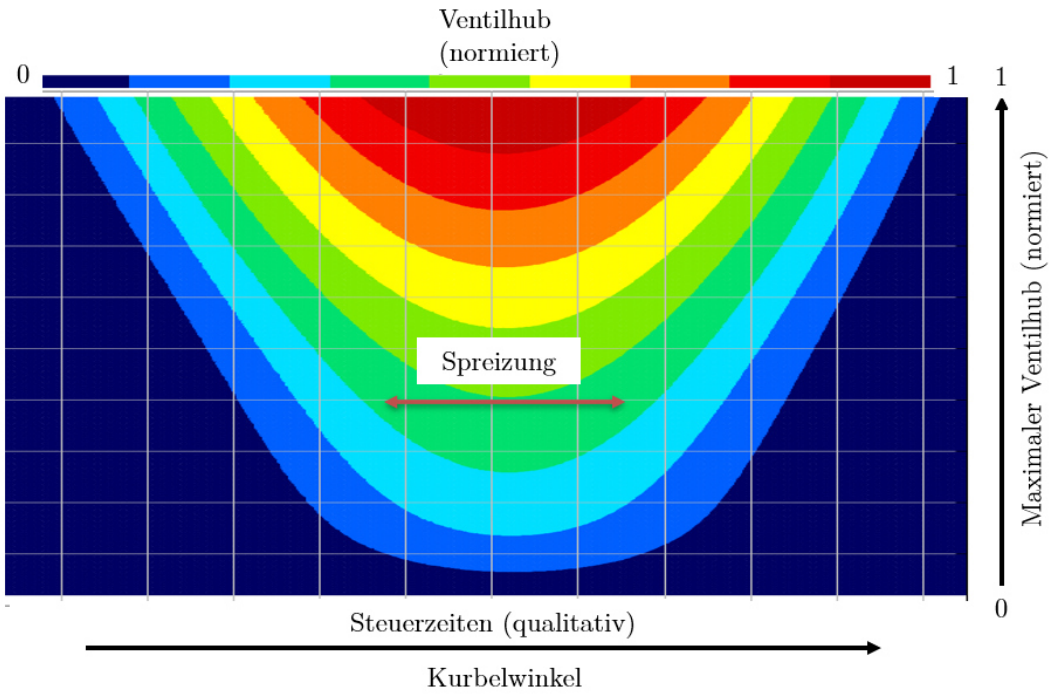


Abbildung 6-2: Kennfeld der Einlassventilhubkurven

Die Inhalte im abgebildeten Konturdiagramm sind normiert (Minimum-Maximum-Methode) beziehungsweise qualitativ dargestellt. Ausgehend von dem gemessenen Maximalhub aus dem Steuergerät (y-Achsenabschnitt) kann aus dem Kennfeld der Hubverlauf ermittelt werden, indem es entlang der x-Achse horizontal von links nach rechts durchlaufen wird. Die momentane Hubhöhe ergibt sich aus den abgebildeten Konturen und der dazugehörigen Farblegende. Der Hubverlauf wird anschließend um den gemessenen Wert der Spreizung entlang der x-Achse verschoben, um den tatsächlichen Hubverlauf aus dem Versuchsmotor nachbilden zu können.

Im Kurbeltriebmodell (Abbildung 6-1 unten) werden grundlegende Parameter bezüglich der Geometrie (Zylinderdurchmesser, Kolbenhub), der Reibung, der Trägheit, der Kinematik (Drehzahl) oder des Motortyps (hier 4-Takt) festgelegt.

Das Injektormodell (in Abbildung 6-1 über dem Brennraum abgebildet) ist über den maximalen Kraftstoffdurchfluss, die Anzahl und Geometrie der Löcher, den Kraftstoff und die volumetrische Durchflusseffizienz parametriert. Aus den Messdaten werden der Einspritzbeginn und das vorherrschende Luft-Kraftstoff-Verhältnis integriert, sodass der Einspritzverlauf durch die Berechnung der notwendigen Kraftstoffmenge festgelegt werden kann.

Das zentrale Element des Modells ist der Brennraum. An dieser Stelle wird der gemessene Zylinderdruckverlauf berücksichtigt. Weiterhin sind Modelle für den Wärmeübergang und die Zylinderinnenströmung vorhanden, welche optimiert bzw. kalibriert werden können. Als Wärmetransfermodell wird „WoschniGT“ (abgeleitet aus den von Woschni [99] formulierten Zusammenhängen) verwendet. Der Brennraum ist dabei in zwei Temperaturzonen unterteilt (verbrannte und unverbrannte Zone).

Das Ziel der Druckverlaufsanalyse ist es, mit den im Modell integrierten Informationen (Messwerte, Geometrien, Randbedingungen, etc.) den Druckverlauf im Zylinder an die realen Messwerte vom Prüfstandsversuch anzugleichen. Der Druckverlauf im Modell hängt von zahlreichen Faktoren ab, unter anderem von der Verdichtung, den Wandwärmeverlusten, aber auch essenziell vom Brennverlauf. Jener resultiert als eines der wichtigsten Ergebnisse aus der Druckverlaufsanalyse. Weiterhin können die Initialbedingungen im Brennraum nach dem Schließen der Einlassventile (Strömung, Gaszusammensetzung, Temperatur, etc.) bestimmt werden.

Das Ergebnis der Druckverlaufsanalyse wird nachfolgend exemplarisch an einem Betriebspunkt (3000 $min^{-1}$; 125 Nm) erläutert. In Abbildung 6-3 ist der gemessene Druckverlauf (Quadrat) zusammen mit dem Druckverlauf der Simulation (Dreieck) im p-V-Diagramm dargestellt. Die Messung und die Simulation stimmen über den gesamten Arbeitszyklus sehr gut überein. Die maximale Abweichung zwischen den beiden Verläufen beträgt während des Spitzendrucks circa 0.4 bar und damit relativ betrachtet weniger als 1 %.

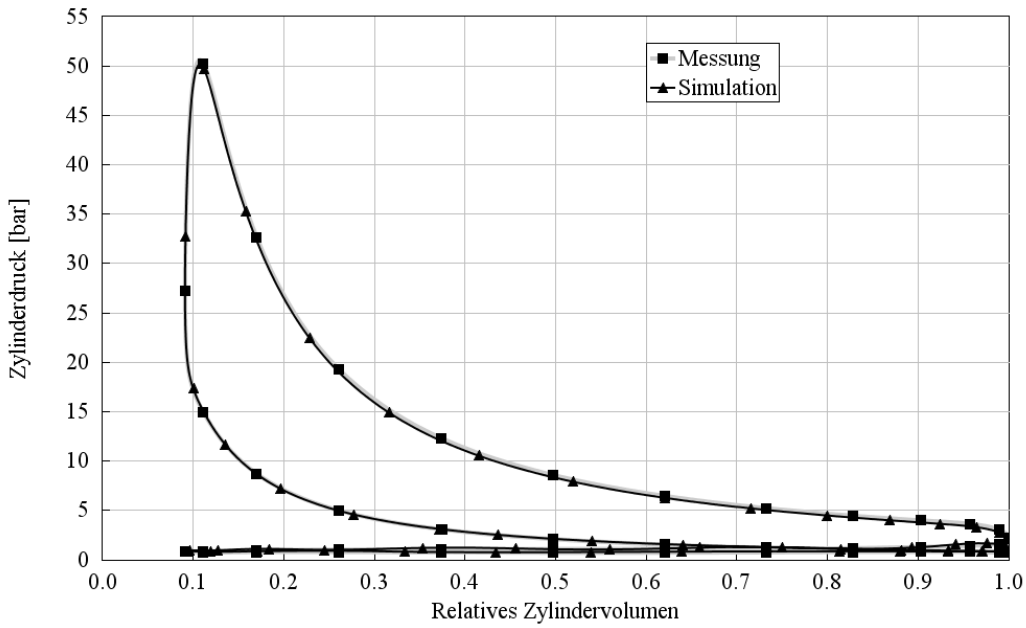


Abbildung 6-3: Vergleich p-V-Diagramm Messung und Simulation (Betriebspunkt: 3000 $min^{-1}$; 125 Nm)

Die Abweichungen zwischen dem gemessenen und dem simulierten Druckverlauf (und so auch dem Brennverlauf) können eine Vielzahl von Ursachen haben, wie beispielsweise nach [94]:

- Fehler in den Eingangsdaten: Verschiebung zwischen dem gemessenen Druck und der Kurbelwellenposition
- Fehler in geometrischen Annahmen: Abweichung des angegebenen vom tatsächlichen Verdichtungsverhältnis
- Fehler in den Annahmen bezüglich des Wärmetransfers im Zylinder: generell problematisch, da Messungen sehr aufwändig und Annahmen zwingend erforderlich sind

Um trotz der Unsicherheiten eine Bewertungsgrundlage für den errechneten Brennverlauf zu haben, wird in GT-Suite ein sogenannter „Fuel Energy Multiplier“ eingeführt, der die Kraftstoffenergie anpasst, um die simulierte Verbrennungseffizienz der gemessenen anzupassen. Eine Abweichung dieses Multiplikators größer 5 % ausgehend von dem Basiswert 1 wird als Fehler gewertet und deutet auf Mängel im Modell oder den Eingangsdaten hin [94]. Im vorliegenden Betriebspunkt liegt der Fuel Energy Multiplier bei circa 2 % und damit deutlich innerhalb der Vorgaben. Der resultierende Summenbrennverlauf ist in Abbildung 6-4 dargestellt.

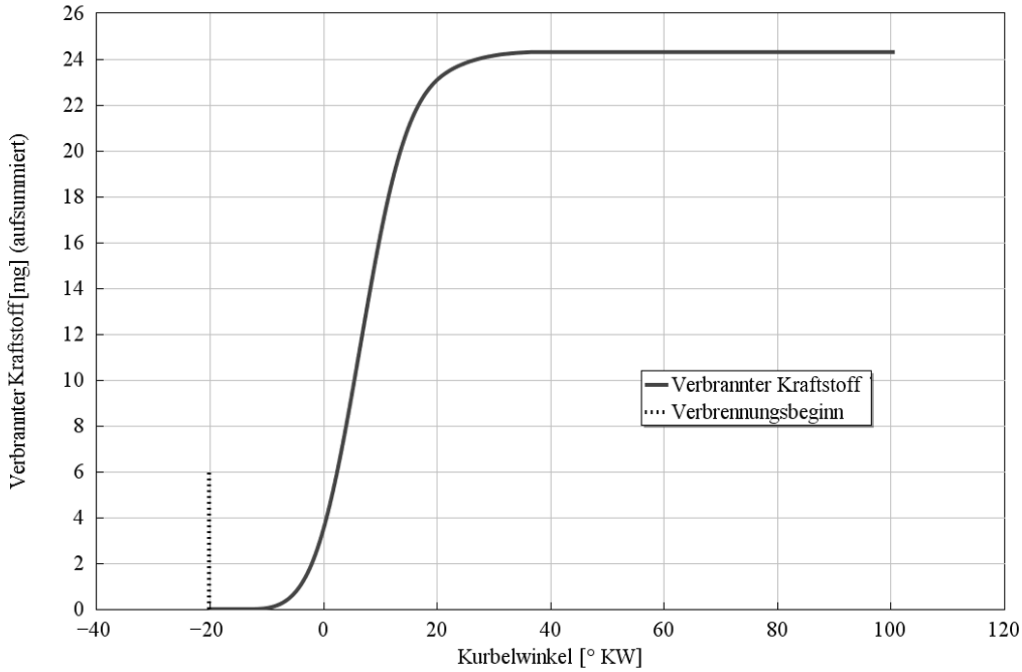


Abbildung 6-4: Summenbrennverlauf (Betriebspunkt: 3000 $min^{-1}$; 125 Nm)

Der Summenbrennverlauf stellt die summierte verbrannte Kraftstoffmasse in Abhängigkeit der Kurbelwinkelposition dar. Im vorliegenden Beispiel werden während des Arbeitsspiels circa 24 mg Kraftstoff verbrannt. Der Verbrennungsschwerpunkt, bei welchem 50 % des Kraftstoffanteils verbrannt sind, liegt bei circa 8 °KW. Die Vorgehensweise bei der Auswertung der

anderen 65 Betriebspunkte ist identisch. Je nach Betriebspunkt gibt es weniger starke oder stärkere Abweichungen verglichen mit dem vorgestellten Beispiel. Insgesamt wurde auf eine Minimierung der Gesamtabweichung – bei stärkerer Gewichtung der Kennfeldbereiche, die im realen Straßenbetrieb vorkommen – geachtet. Perspektivisch könnten die Ergebnisse noch weiter verbessert werden, indem zusätzliche Strömungssimulationen für den Ladungswechsel genutzt, oder statt Steuergeräteparametern eine direkte Messung wichtiger Größen (Zündzeitpunkt, Ansteuerung des Einspritzventils) durchgeführt wird.

### 6.1.2 Verbrennungsmodell

Die Architektur für die Erstellung und Kalibrierung des Verbrennungsmodells ist, verglichen mit dem Modell der Druckverlaufsanalyse, sehr viel einfacher, was daraus resultiert, dass das Verbrennungsmodell auf Randbedingungen bzw. Initialbedingungen aus der Druckverlaufsanalyse basiert. Dazu gehören unter anderem der Liefergrad, der Restgasgehalt, Wandtemperaturen und die Turbulenz im Zylinder. Es findet in dieser Phase des Prozesses kein Gaswechsel statt. Das Verbrennungsmodell wird daher in einem geschlossenen Volumen kalibriert, weshalb alle Luftpfad-Verbindungen ausgehend vom Brennraum/Zylinder entfallen, siehe Abbildung 6-5.

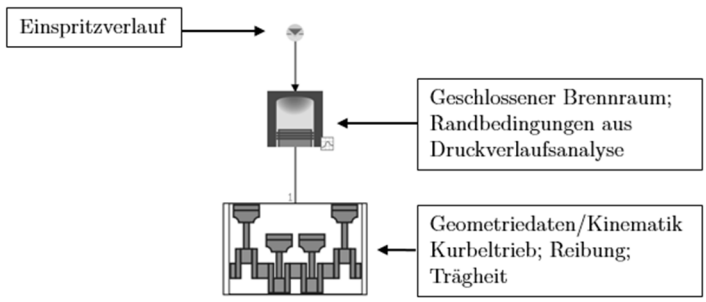


Abbildung 6-5: Aufbau zur Kalibrierung des Verbrennungsmodells

Während der Kalibrierung werden der Druck- und Brennverlauf aus der Druckverlaufsanalyse mit den äquivalenten Verläufen aus dem prädiktiven Verbrennungsmodell verglichen. Das Verbrennungsmodell ist ein zusätzliches (in der Druckverlaufsanalyse nicht vorhandenes) Objekt im Brennraummodell, welches Parameter bezüglich der Flammengeometrie, der laminaren und turbulenten Flammenausbreitung, des Klopfverhaltens und der Emissionsbildung (siehe 6.1.3) enthält. Das Hauptziel im Optimierungsprozess ist es, einen Parametersatz zu finden, welcher die Wurzel der mittleren quadratischen Abweichung (engl. „Root Mean Square Error“ RMSE) zwischen analysiertem und prädiziertem Brennverlauf reduziert. Eine Optimierung unter

Berücksichtigung mehrerer Zielfunktionen (Betriebspunkte) stellt in der Regel – und definitiv im vorliegenden Fall – einen Kompromiss dar. Aus diesem Grund wurden die Betriebspunkte im für den Straßenverkehr relevanten Kennfeldbereich stärker gewichtet. Da das Verbrennungsmodell die Grundlage für die Emissionsmodelle darstellt, können diese ebenfalls besser an die Messungen angepasst werden. In Abbildung 6-6 ist der erwähnte besonders relevante Kennfeldbereich eingezeichnet.

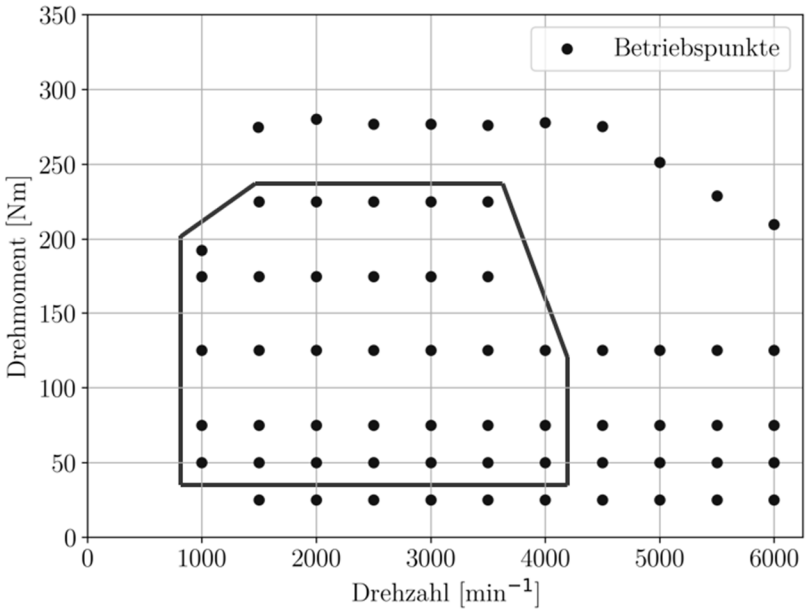


Abbildung 6-6: Kennfeldbereich (Rahmen) mit stärkerer Gewichtung bei der Optimierung der Verbrennungsparameter

Die eingerahmten Betriebspunkte stellen etwa die Hälfte der erfassten Betriebspunkte im Kennfeld dar. Es ist zu erkennen, dass besonders hohe Lasten, sehr niedrige Lasten (unter 10 % der im Drehzahlbereich verfügbaren Last) und Drehzahlen ab 4000 $min^{-1}$ weniger stark gewichtet sind.

Für die Optimierung wurde ein genetischer Algorithmus verwendet, welcher ursprünglich auf dem multikriteriellen evolutionären Algorithmus NSGA-III (engl. „Non-dominated Sorting Genetic Algorithm Version III“) [100] basiert. Die wichtigsten Parameter des genetischen Algorithmus sind die Populationsgröße und die Anzahl der Generationen. Erstere bestimmt die Anzahl an verschiedenen Designs, welche in einer Generation erstellt werden und hat Einfluss auf die Chance, durch Diversifizierung ein Optimum zu finden. In jeder Generation entwickelt sich die Population durch Rekombinationen von erfolgsversprechenden Parameterkombinationen und Mutationen weiter. Letztere dienen dazu, um weiterhin außerhalb vermeintlich bekannter globaler Optima zu explorieren, welche unter Umständen nur lokale Optima darstellen. [101]

Basierend auf gesammelten Erfahrungswerten und Empfehlungen des Softwareherstellers wurde eine Populationsgröße von 90 gewählt und die Anzahl der Generationen auf 40 festgelegt. Daraus resultieren insgesamt 3600 Designs – Parameterkombinationen – die für alle 66 Betriebspunkte getestet wurden. Die wichtigsten Parameter, welche die höchste Sensitivität bezüglich des RMSE des Brennverlaufs aufweisen, sind [94]:

- Dilution Effect Multiplier: Beschreibt den Einfluss des Restgasgehaltes (Dilution) in der unverbrannten Zone auf die laminare Flammengeschwindigkeit
- Flame Kernel Growth Multiplier: Bestimmt die initiale Ausbreitung des Flammenzentrums; höhere Werte reduzieren den Zündverzug
- Turbulent Flame Speed Multiplier: Modifiziert die turbulente Flammengeschwindigkeit
- Taylor Length Scale Multiplier: Skaliert den berechneten Wert der Taylor-Mikroskala (siehe [102])

Neben den genannten Parametern wurden während des Optimierungsprozesses noch weitere berücksichtigt, welche aufgrund fehlender Geometrie- oder Strömungsinformationen nicht fest definiert waren (Tumble-Zerfall, initiale Zündfunkengröße, etc.). Da diese zu einem Teil auch die Druckverlaufsanalyse beeinflussen, welche wiederum Initialbedingungen für die Optimierung des Verbrennungsmodells liefert, wurde eine iterative Herangehensweise gewählt. In Abbildung 6-7 ist der finale Optimierungsprozess dargestellt. Dieser beginnt bereits mit einem optimierten Basisdesign (RMSE ca. 0.0045).

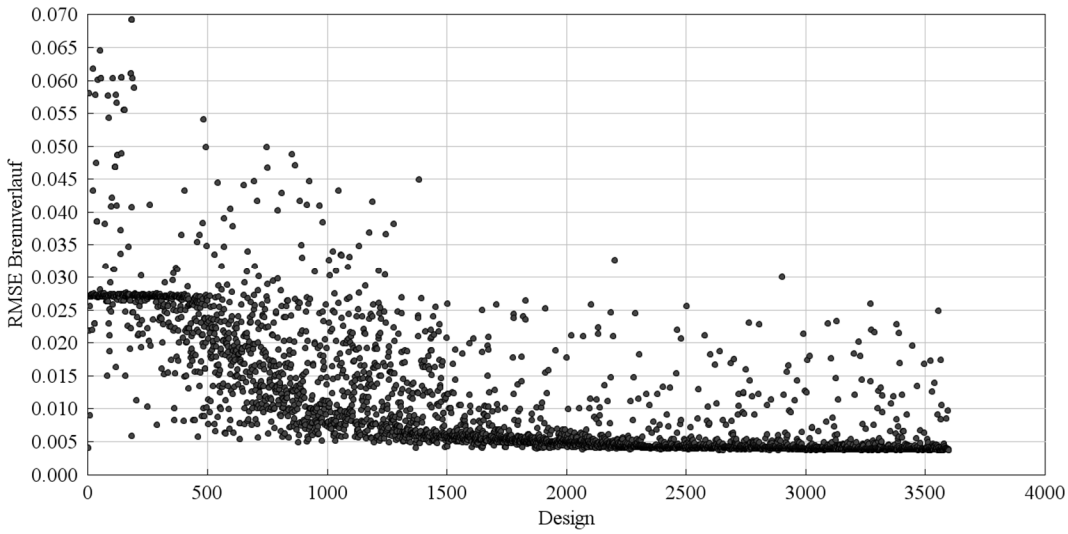


Abbildung 6-7: Optimierung der Verbrennungsmodellparameter bezüglich minimalem Brennverlauf RMSE mit einem genetischen Algorithmus

Es ist zu erkennen, dass die Population der ersten Generation breit gestreut ist und alle Designs schlechter als die Basis sind. Auch die dominanten Rekombinationen der ersten 5-6

Generationen (circa 500 Designs) konzentrieren sich um RMSE-Werte von 0.0275. Erst danach ist eine Verbesserung erkennbar, es dauert jedoch noch weitere circa 1000 Designs, bis sich ein neues Optimum einstellt. Dies liegt jedoch nicht an einer mangelnden Performance des Algorithmus, sondern an der angesprochenen iterativen Vorgehensweise und dem guten Startpunkt. Ab Design 1500 wird eine deutliche Konzentration von Designs um das Optimum von knapp unter 0.004 RMSE ersichtlich. Die Rekombinationen häufen sich um diesen Bereich und fallen bis in Generation 25 (2250 Designs) leicht ab. Danach ist keine Verbesserung mehr erkennbar. Bei den Ausreißern nach oben handelt es sich um Mutationen, welche jedoch ebenfalls kein neues Optimum liefern. Der finale RMSE-Wert für den Brennverlauf liegt bei 0.0036, die dazugehörigen Verbrennungsparameter bilden die Basis für das Verbrennungsmodell.

### 6.1.3 Optimierung der Emissionsmodelle

Die Kalibrierung der Emissionsmodelle basiert auf den beiden vorangehenden Kapiteln, was sich auch im Modellaufbau nachvollziehen lässt, siehe Abbildung 6-8.

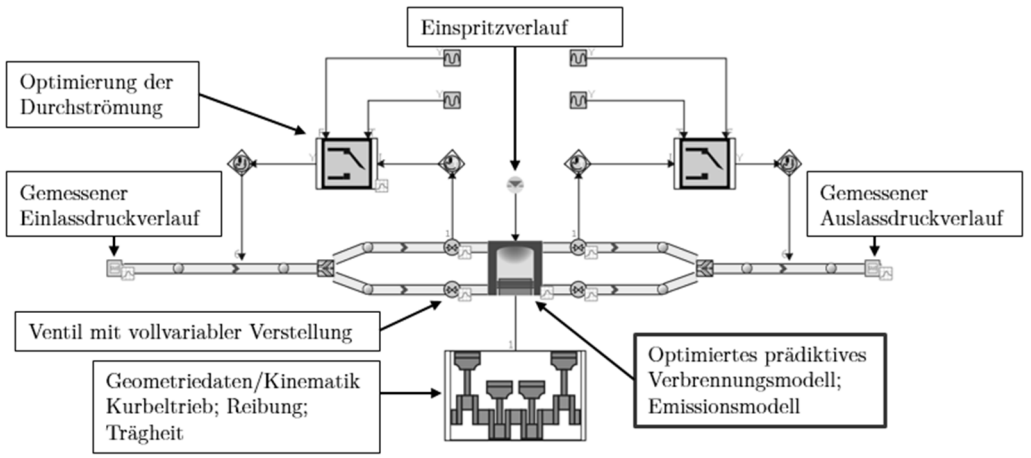


Abbildung 6-8: Aufbau zur Kalibrierung der Emissionsmodelle

Das Modell ist grundsätzlich analog zu dem Modell der Druckverlaufsanalyse aufgebaut. Der Unterschied liegt im Zylinder bzw. Brennraum. Bei der Druckverlaufsanalyse wird an dieser Stelle die Analyse und Anpassung des Druckverlaufes über einen gesamten Verbrennungszyklus durchgeführt. Für die Optimierung der Emissionsmodelle ist es jedoch sinnvoll, den Druck- und Brennverlauf durch das kalibrierte Verbrennungsmodell zu prädizieren. Dadurch können auch die Emissionsmodelle entsprechend den Zielen der Arbeit prädiktiv eingesetzt werden. Sowohl die Emissionsmodelle als auch die darin enthaltenen Parameter sind in der Grundform in GT-Suite enthalten und können daher als effiziente Basis genutzt werden. Die Druckverläufe

außerhalb des Brennraumes (Einlassdruck und Auslassdruck) werden erneut aus Messungen entnommen. Die Modellstruktur ist demzufolge vereinfacht, was die Rechenzeit verringert. Ein weiterer Vorteil besteht darin, dass Komponenten (Ansaugstrecke inklusive Aufladung; Abgasstrecke mit Turbine und Abgasnachbehandlung) außerhalb der gemessenen Druckverläufe nicht modelliert werden müssen und dadurch die zwangsläufig damit verbundenen Unsicherheiten entfallen.

Nachfolgend werden die verwendeten Emissionsmodelle, die wichtigsten Parameter und die jeweiligen Optimierungsprozesse/-ergebnisse dargestellt. Bei der Optimierung wurde das gleiche Vorgehen bezüglich der Betriebspunkte wie in Kapitel 6.1.2 gewählt. Das Optimierungsziel je Emissionsspezies ist die Minimierung der Wurzel der mittleren quadratischen Abweichung zwischen den am Motorenprüfstand mit Hilfe der Abgasmesstechnik aufgenommenen Messwerten und den prädizierten Emissionswerten.

**Stickstoffoxide:** Für die Berechnung der Stickstoffoxide wird ein erweiterter Zeldovich-Mechanismus (siehe Kap. 2.2.1) eingesetzt, welcher mit sechs Parametern kalibriert werden kann [103]:

- $NO_x$ Calibration Multiplier: Multiplikator für die Netto (Entstehung – Dissoziation) $NO_x$-Bildungsrate
- $N_2$ Oxidation Rate Multiplier: Multiplikator für die $N_2$-Oxidationsgleichung ($O + N_2 = NO + N$)
- $N_2$ Oxidation Activation Energy Multiplier: Multiplikator für die Aktivierungsenergie für die $N_2$-Oxidationsgleichung
- $N$ Oxidation Rate Multiplier: Multiplikator für die $N$-Oxidationsgleichung ($N + O_2 = NO + O$)
- $N$ Oxidation Activation Energy Multiplier: Multiplikator für die Aktivierungsenergie für die $N$-Oxidationsgleichung
- $OH$ Reduction Rate Multiplier: Multiplikator für die $OH$-Reduktionsgleichung ($N + OH = NO + H$)

In Abbildung 6-9 ist eine Gegenüberstellung der prädizierten Emissionswerte des Basismodells (Dreieck), des optimierten Stickstoffoxid-Emissionsmodells (Quadrat) und der Messwerte (Kreis) dargestellt. Zu Gunsten der Übersichtlichkeit ist je Drehzahlstufe ein Betriebspunkt (50 Nm) abgebildet. Das Basismodell berechnet die $NO_x$-Emissionen über den gesamten Drehzahlbereich zu niedrig. Die gemessenen Werte liegen bis zu 200 % über dem Prädiktionswert. Durch die Kalibrierung der oben beschriebenen Parameter kann der Verlauf der Stickstoffoxide deutlich besser angenähert werden. Die maximale Abweichung in den dargestellten Betriebspunkten beträgt 35 %, im Minimum sind es 2.2 %. Bezüglich der Abweichungen lässt sich jedoch keine allgemeine Aussage (wie es beim Basismodell der Fall ist) treffen. Das optimierte Modell überschätzt die Stickstoffoxid-Emissionen tendenziell im oberen Drehzahlbereich, liegt dahingegen am unteren Drehzahlende unter den Messwerten. Insgesamt wurden während des

Optimierungsprozesses 2400 Designs getestet, wobei im letzten Fünftel des Prozesses keine merklichen Verbesserungen mehr auftraten.

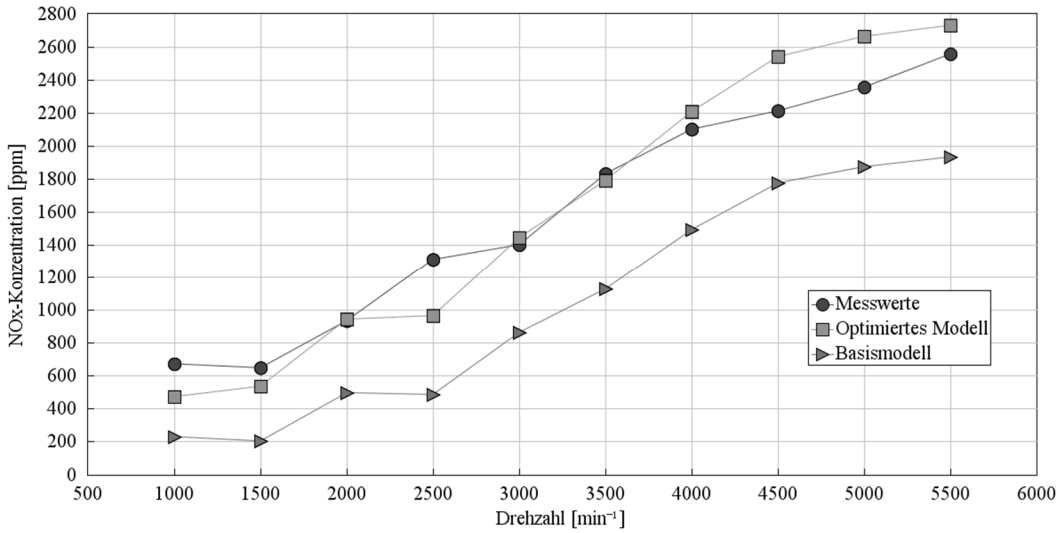


Abbildung 6-9: $NO_x$-Emissionen - Messwerte (Kreis), optimiertes Modell (Quadrat), Basismodell (Dreieck); je ein Betriebspunkt je Drehzahlstufe

**Kohlenstoffmonoxid:** Die Prädiktion von Kohlenstoffmonoxid erfolgt mit Hilfe eines kinetischen *CO*-Modells, welches in seiner Grundform auf dem in Kap. 2.2.2 (Gl. 20) beschriebenen Mechanismus – $CO + OH <=> CO_2 + H$ – basiert. Für die Anpassung des Emissionsmodells an die gemessenen Werte können zwei Parameter kalibriert werden [103]:

- Pre-exponent Multiplier: Multiplikator für die Reaktionsrate
- Activation Temperature Multiplier: Multiplikator für die notwendige Aktivierungstemperatur

Abbildung 6-10 zeigt einen Vergleich der *CO*-Konzentrationsverläufe zwischen der Messung (Kreis), dem optimierten prädiktiven Modell (Quadrat) und dem Ausgangsmodell (Dreieck). Das Basismodell stellt die Kohlenstoffmonoxid-Konzentration in jedem Betriebspunkt zu gering dar. Gegenüber den Messwerten betragen die Prädiktionswerte im Mittel zwischen 10 und 20 %. Im zweiten Betriebspunkt (Drehzahlstufe 1500 $min^{-1}$) sind sowohl im Basismodell als auch im optimierten Modell die berechneten Konzentrationen negativ, was im Widerspruch zur Realität steht. Dies wird im weiteren Verlauf der Arbeit berücksichtigt – beispielsweise erlernen die hybriden Modelle diese offensichtlich falschen Werte korrekt zu interpretieren, um schlussendlich positive Konzentrationen zu berechnen. Das kalibrierte kinetische *CO*-Modell kann den Verlauf der Messwerte besser abbilden, liegt bis auf einen Ausreißer bei der Drehzahlstufe 2500 $min^{-1}$ bezüglich der Konzentrationswerte jedoch ebenfalls zu niedrig. Der Optimierungsprozess wurde mit 1200 Designs bei lediglich zwei Parametern stark ausgedehnt, dennoch konnte das prädizierte Konzentrationsniveau nicht weiter angehoben werden. Dies spricht

für eine im vorhandenen Modell nicht berücksichtigte kinetische Limitierung der Kohlenstoffmonoxid-Oxidation gegen Ende des Expansionsvorganges (vergleiche die Erkenntnisse von Newhall [42], s. Kapitel 2.2.2). So könnten die modellierten Kohlenstoffmonoxid-Konzentrationen zu gering ausfallen, da im Modell mehr Kohlenstoffmonoxid zu Kohlenstoffdioxid oxidiert wird, als dies in der Realität der Fall wäre.
Eine Möglichkeit zur Verbesserung der physikalisch-phänomenologischen Modellierung der Bildung von Kohlenstoffmonoxid kann die zusätzliche Berücksichtigung der in den Gleichungen 21-23 beschriebenen Reaktionsabläufe sein. Dies konnte im Rahmen der vorliegenden Arbeit jedoch nicht umgesetzt werden.

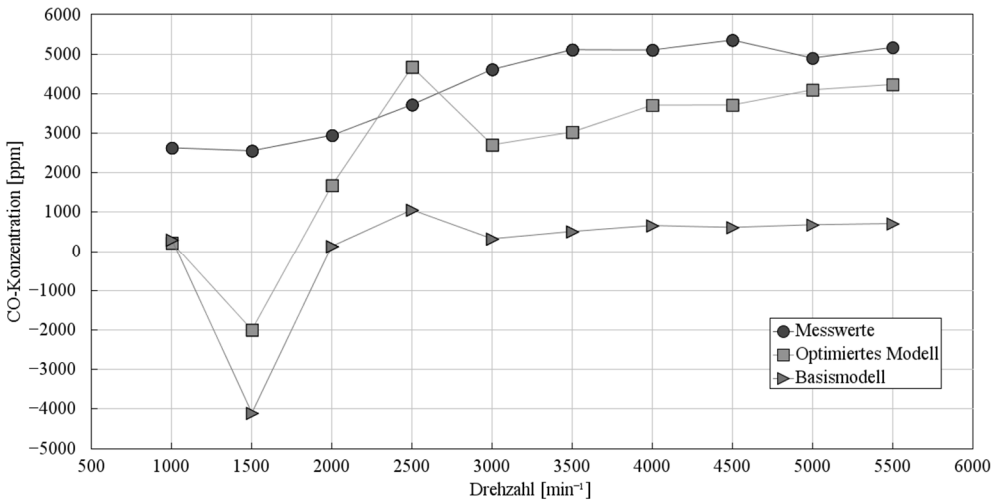


Abbildung 6-10: *CO*-Emissionen - Messwerte (Kreis), optimiertes Modell (Quadrat), Basismodell (Dreieck); je ein Betriebspunkt je Drehzahlstufe

**Unverbrannte Kohlenwasserstoffe:** Für die Simulation der Konzentration von unverbrannten Kohlenwasserstoffen wird ein „2-plate quenching“ Modell für das Erlöschen der Flammenfront und anschließend eine Reaktionskinetik auf Basis der Arbeit von Lavoie [45] eingesetzt. Insgesamt gibt es vier Parameter, welche für den Abgleich der *HC*-Emissionen genutzt werden können [103]:

- Piston-Liner Crevice Volume: Volumen, welches von der Kolbenlauffläche, der Zylinderlaufbahn, dem obersten Kolbenring und dem Kolbenboden eingeschlossen wird (sog. Feuersteg)
- Piston-Liner Clearance: Vorhandenes Kolbenspiel zur Zylinderlaufbahn
- Pre-Exponent Multiplier: Multiplikator für die Reaktionskinetik
- Activation Temperature Multiplier: Multiplikator für die Aktivierungstemperatur in der Reaktionskinetik

In Abbildung 6-11 sind die Konzentrationsverläufe der Messung (Kreis), des optimierten prädiktiven Modells (Quadrat) und des Ausgangsmodells (Dreieck) dargestellt. Es zeigt sich, dass das Basismodell die *HC*-Emissionen im Mittel um circa 100 % zu hoch berechnet. Dies kann einerseits an zu groß angenommenen Spalten liegen, in welchen sich unverbranntes Gemisch ansammeln kann. Andererseits kann die Oxidationsrate von unverbrannten Kohlenwasserstoffen nach der primären Verbrennung zu gering angenommen sein. Das optimierte Modell hingegen korreliert abgesehen von der ersten Drehzahlstufe mit durchschnittlichen Abweichungen im Bereich von 5 % gut mit den Messwerten.

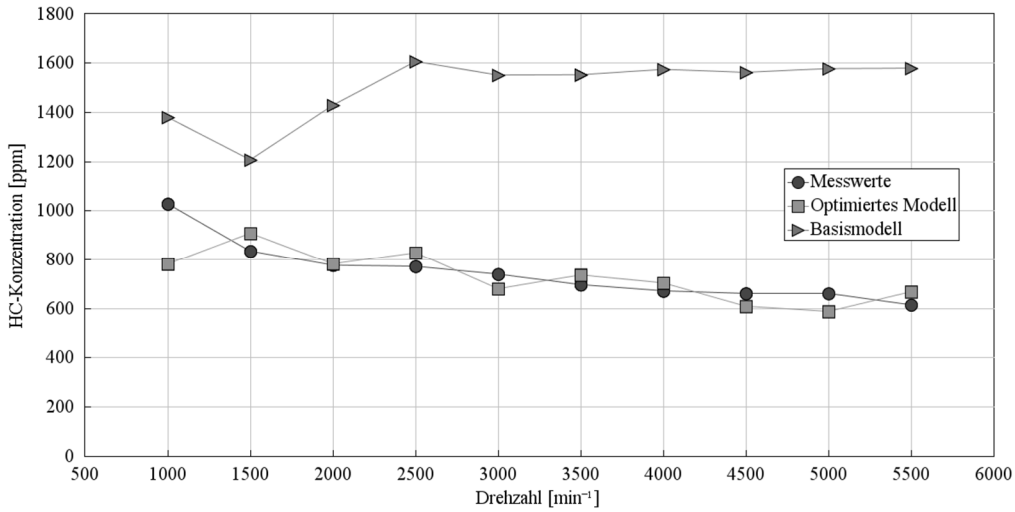


Abbildung 6-11: *HC*-Emissionen - Messwerte (Kreis), optimiertes Modell (Quadrat), Basismodell (Dreieck); je ein Betriebspunkt je Drehzahlstufe

### 6.1.4 Motormodell

Die primäre Motivation für die Erstellung des Motormodells ist die Abbildung des transienten Betriebs und der dazugehörigen Emissionsprädiktionen. Hierzu müssen reale Lastprofile, welche beispielsweise aus den Straßenversuchen stammen, simulativ „nachgefahren“ werden. In Abbildung 6-12 ist der Aufbau des Motormodells schematisch dargestellt.

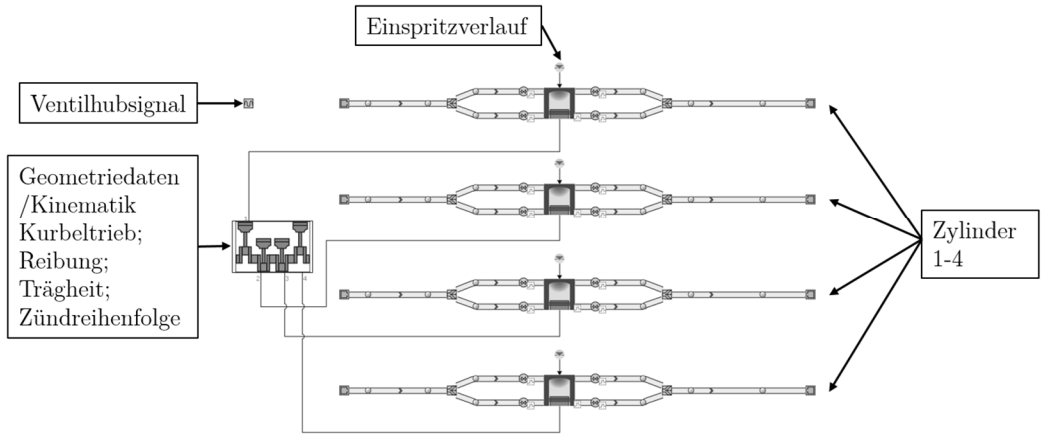


Abbildung 6-12: Transientes Motormodell B48B20M0

Im Gegensatz zu den bisher beschriebenen Modellen (Kap. 6.1.1-6.1.3), sind im Motormodell (wie beim realen Motor B48B20M0) alle vier Zylinder abgebildet. Dadurch können die Betriebspunkte des Vollmotors auf dem Prüfstand ohne weitere Umrechnung dargestellt werden. Durch die Verknüpfung der einzelnen Zylinder mit dem modellierten Kurbeltrieb sind die folgenden Motoreigenschaften/-systeme ebenfalls darstellbar beziehungsweise zu berücksichtigen:

- Zündfolge
- Trägheit (oszillierende und rotierende Massen)
- Reibung (diese kann aus Schleppmessungen bestimmt werden)
- Realistischer Drehmomentverlauf durch Überlagerung der Drehmomentverläufe der Einzelzylinder

Die Vorgänge im Einzelzylinder basieren auf den zuvor kalibrierten Verbrennungs- und Emissionsmodellen. Das Vorgehen, die Verbrennungsmodelle an einem Einzylinder zu erstellen und danach ein Mehrzylindermodell aufzubauen, wurde präferiert, da sich besonders die essenzielle Druckverlaufsanalyse als Basis für das prädiktive Verbrennungsmodell besser mit dem Bezugssystem eines Brennraums durchführen lässt.

Außerhalb der Brennräume ist das Modell geometrisch stark vereinfacht und basiert auf Messdaten, mit denen die Umgebungsbedingungen (hauptsächlich Druck und Temperatur) an den entsprechenden Stellen vor und nach dem Brennraum eingestellt und damit das reale Systemverhalten außerhalb dieses Bereiches nachgebildet werden kann. Dadurch müssen Komponenten wie beispielsweise der Turbolader oder der Ladeluftkühler nicht modelliert werden, was Fehlerquellen ausschließt und eine effizientere Modellerstellung erlaubt.

Ein Nachteil dieser Methode ist, dass das Modell außerhalb der gemessenen Betriebsbereiche nicht verlässlich eingesetzt werden kann. In diesem Anwendungsfall würden Informationen

über das nicht modelltechnisch dargestellte System fehlen und besonders bei einer Extrapolation ist mit Fehlern zu rechnen. Außerdem kann das Modell nicht genutzt werden, um beispielsweise die Auswirkungen von geometrischen Anpassungen (Ladeluftstrecke, Verdichterrad, etc.) abzuschätzen. Dies ist im Rahmen dieser Arbeit jedoch unkritisch, da das Modell einen anderen Zweck verfolgt und primär in einem klar definierten und durch Daten hinreichend abgesicherten Betriebsbereich genutzt wird.

Für den dynamischen Betrieb des Motormodells müssen gegenüber den bisherigen Modellen einige Anpassungen gemacht werden. Zuvor fixierte Werte für den stationären Betrieb (Zündzeitpunkt, Ventilhub, etc.) müssen als zeitlich hochaufgelöste Verläufe integriert werden, um den realen Betrieb des Motors zu gewährleisten. Hierzu werden die mit 5 Hz Messfrequenz aufgenommenen Daten des Prüfstandsmotors genutzt und integriert. In Abbildung 6-13 ist als Beispiel einer solchen Größe der zeitliche Verlauf des maximalen Einlassventilhubs dargestellt. Dieser verhält sich über den 2000 Sekunden langen RDE-Ausschnitt sehr dynamisch, was die Notwendigkeit einer schnellen Signalerfassung unterstreicht. Zu den weiteren dynamisch eingebundenen Parametern gehören:

- Drehzahl
- Druck und Temperatur im Sammler
- Druck und Temperatur im Abgaskrümmer
- Lambda
- Einlassspreizung
- Auslassspreizung
- Einspritzbeginn
- Zündzeitpunkt

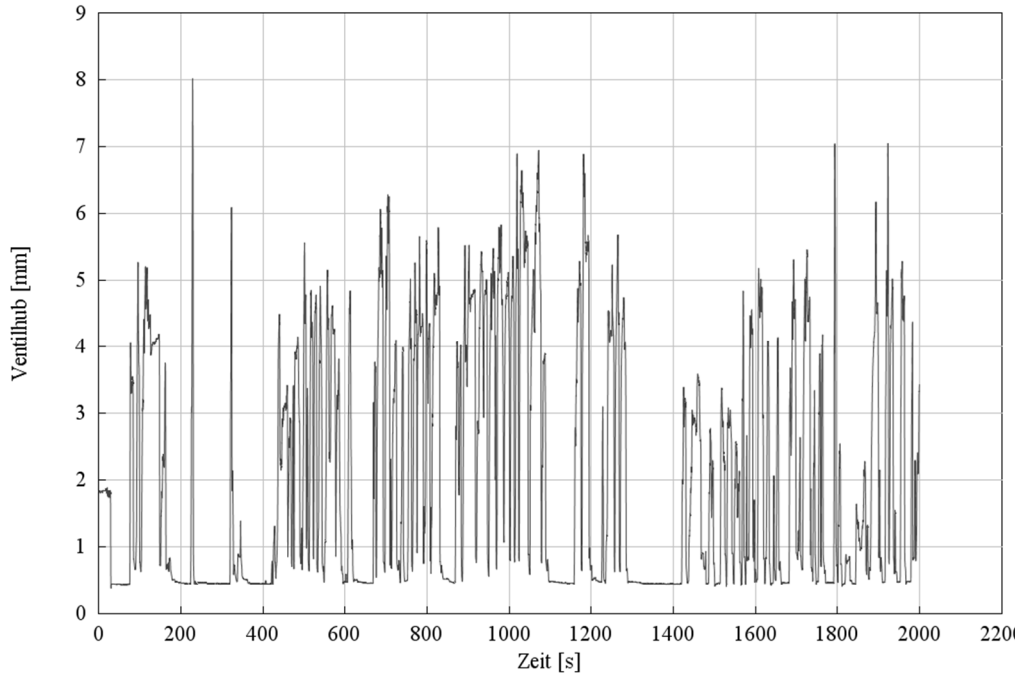


Abbildung 6-13: Verlauf des maximalen Ventilhubs während einer RDE-Fahrt

Durch die Vorgabe der Druckverhältnisse vor und nach dem Brennraum, der Steuerzeiten, der Drehzahl und dem in vorherigen Modellierungsschritten abgeglichenen Liefergrad, kann die experimentell ermittelte Luftmenge des Prüfstandsmotors angenähert werden. Dadurch ergibt sich mit dem gewünschten Luft-Kraftstoff-Verhältnis, dem aufgezeichneten Zündzeitpunkt und dem kalibrierten Verbrennungsmodell ein realitätsnaher Drehmomentverlauf. Das ist in Abbildung 6-14 zu erkennen. Dargestellt ist ein 900 Sekunden langer Ausschnitt aus einer RDE-Fahrt. Die RDE-Fahrt wurde nach dem in Kapitel 5.3 beschriebenen Vorgehen am Prüfstand nachgefahren und vermessen. Die durchgängige Kurve stellt das am Drehmomentmessflansch aufgezeichnete Drehmoment dar. Die oben genannten Parameterverläufe wurden dem Motormodell zur Verfügung gestellt, wodurch der als gestrichelte Linie dargestellte Verlauf entstand. Es wird deutlich, dass das Verhalten des Prüfstandsmotors simulativ auch im hochdynamischen Betrieb gut abgebildet werden kann.

Durch die Integration der in Kapitel 6.1.3 optimierten Emissionsmodelle können die Emissionen transient berechnet werden, was in Kapitel 7.1 analysiert wird.

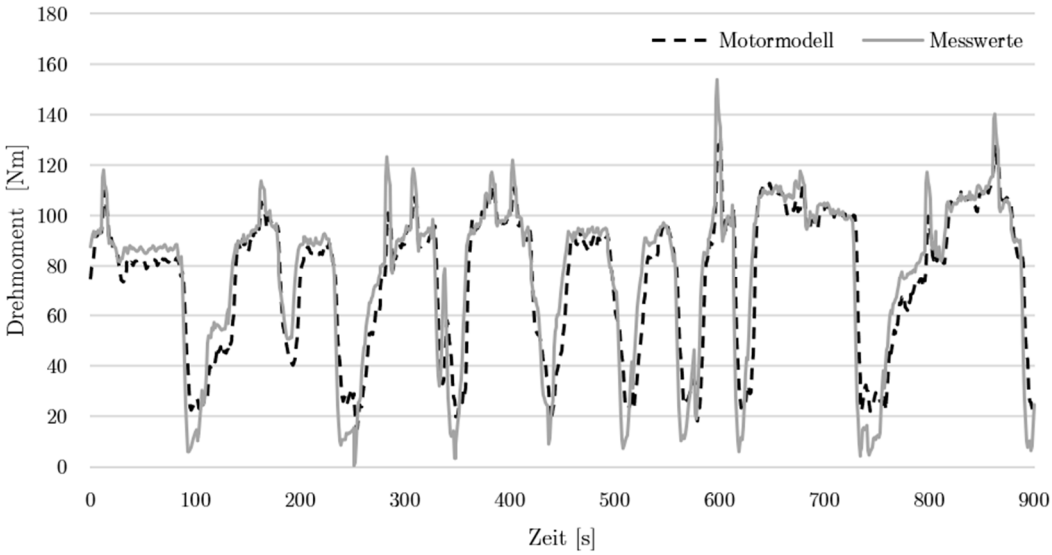


Abbildung 6-14: Drehmomentverlauf Motormodell und Messwerte

## 6.2 Datenbasierte Modellierung

Für die datenbasierte Modellierung wurde das Deep Learning Framework PyTorch eingesetzt. PyTorch kann als Machine Learning Bibliothek für das Open Source Programm Python beschrieben werden und zeichnet sich neben der stetigen Weiterentwicklung aufgrund des Open Source Charakters durch den gleichzeitigen Fokus auf Geschwindigkeit und Bedienfreundlichkeit aus. Es ist array-basiert und für die Verwendung auf Grafikprozessoren (GPUs engl. „Graphics Processing Unit“) und die Nutzung der CUDA-Architektur (engl. „Compute Unified Device Architecture“) optimiert. CUDA ermöglicht die Nutzung der GPU für generelle (nicht Grafik-) Berechnungen, wodurch besonders bei der parallelen Verarbeitung von Vorgängen gegenüber einem Standard-Prozessor ein erheblicher Zeitvorteil entstehen kann. Obwohl PyTorch in Python eingebettet ist, ist es in der Grundstruktur in C++ geschrieben, um die Leistungsfähigkeit weiter zu optimieren. [104], [105]

Für das Training und die Anwendung (Inferenz) der verschiedenen Modelle wurde je nach Leistungsbedarf auf zwei verschiedene Hardwaresysteme zurückgegriffen. Die Inferenz und kleinere Trainingsumfänge (kompakte Modellstruktur, Anzahl der Trainingsepochen $< 1000$) konnten mit einem lokalen Simulationsrechner erfolgen. Dieser verfügt über einen Ryzen 5 5600X 6-Kern Prozessor, 16 GB Arbeitsspeicher und eine Geforce 3060 Grafikkarte mit 16 GB VRAM und CUDA-Unterstützung.

Für das Training größerer Modelle und besonders das Hyperparametertuning (siehe Kapitel 6.2.4) stand der Hochleistungsrechner „Elwetritsch“ der Allianz für Hochleistungsrechnen Rheinland-Pfalz (AHRP) zur Verfügung. Dieses Cluster vereint mehr als 17000 Rechenkerne, 136 TB Arbeitsspeicher und auf KI-Berechnungen optimierte GPUs mit Tensor-Recheneinheiten [106].

In den nachfolgenden Unterkapiteln wird die Konzeption, die Erstellung und die Optimierung der datenbasierten Modelle für die hochdynamische Emissionsmodellierung beschrieben. Hierzu werden zuerst relevante Methoden des Maschinellen Lernens betrachtet und eine Auswahl basierend auf den höchsten Erfolgsaussichten und dem geplanten Modelleinsatz getroffen. Anschließend werden die zur Verfügung stehenden Messdaten aus unterschiedlichen Quellen betrachtet und für die weitere Verarbeitung aufbereitet. Dies zielt darauf ab, im späteren Prozess einen möglichst effizienten Trainingsvorgang zu ermöglichen. In Kapitel 6.2.3 werden die Modelleingangsparameter bestimmt. Hierzu stehen grundsätzlich die in Kap. 5.1.2 genannten 146 Messgrößen zur Verfügung. Um jedoch einen effizienten Modellaufbau darstellen zu können, sollte die Anzahl an Eingängen auf wenige, dafür stark mit den Ausgangsgrößen korrelierende Parameter reduziert werden.

Der prinzipielle Ablauf für die Erstellung eines optimierten Machine Learning Modells für die Prädiktion von Emissionen lässt sich wie folgt beschreiben:

1. Generierung einer Datengrundlage
2. Auswahl möglicher, auf den Anwendungsfall passender Machine Learning Methoden bzw. Architekturen
3. Aufbereitung der Messdaten
4. Festlegung der Eingangsgrößen
5. Festlegung der Modellstruktur (Startpunkt: Erfahrungswerte basierend auf Ein- und Ausgangskonfiguration und erwarteter Komplexität der zu modellierenden Zusammenhänge)
6. Training der Modelle inklusive Hyperparametertuning (Anpassung der Modellstruktur zur Optimierung des Trainings- und Validierungserfolges); Validierung mit zusätzlichen Daten
7. Testen der trainierten Modelle – Verifikation durch den Abgleich der Prädiktionswerte mit Messwerten aus einem weiteren, nicht im Training enthaltenen Datensatz, dem sogenannten Testdatensatz
8. Iteration: Je nach erreichter Vorhersagequalität Wiederholung des Ablaufes. Wenn die Datengrundlage nicht ausreichend ist, muss ein erneuter Durchlauf ab Schritt 1 erfolgen

Der dargestellte Ablauf wird mit Ausnahme des ersten Schritts, welcher bereits in Kap. 5 beschrieben wurde, im Anschluss erläutert.

### 6.2.1 Anwendungsspezifisch geeignete Methoden des Maschinellen Lernens

Wie in Kapitel 2.3.2 erwähnt, können aus den Vorarbeiten zu dieser Arbeit ([60]) bereits wichtige Erkenntnisse bezüglich der Auswahl anwendungsspezifisch geeigneter Methoden des Maschinellen Lernens gezogen werden. Hierfür ist es relevant, zuerst den Begriff „anwendungsspezifisch“ genauer zu definieren. Aus den Zielen der vorliegenden Arbeit (siehe Kapitel 4) lassen sich folgende Bewertungskriterien ableiten:

- Vorhersagequalität (allgemeingültig)
- Modelleffizienz
- Prädiktion der Emissionen für eine unbestimmte Anzahl an Zeitschritten auf Basis verfügbarer Eingangsgrößen und/oder vergangener Prädiktionen (→ essenziell für die Optimierung der Betriebsstrategie)

Zu den in [60] getesteten Modellen gehören drei Architekturen aus dem Bereich des Shallow Learnings und zwei Deep Learning-Algorithmen. Zur ersten Kategorie gehören die „Random Forest Regression“, „XGBoost Regression“ (XGB) und „Support Vector Regression“ (SVR). Diese Modelle wurden autoregressiv und auf exogenen Eingängen basierend aufgebaut, was bedeutet, dass der Modellausgang von vergangenen Ausgängen, internen Zuständen und zusätzlich von nicht direkt beobachtbaren Informationen abhängt. Letztere sind im konkreten Anwendungsfall reale Emissionsmesswerte, welche die Vorhersage weniger Zeitschritte erheblich verbessern. Auf diese Weise konnte gegenüber den betrachteten Deep Learning Algorithmen eine im Schnitt circa sechsfach höhere Prädiktionsgenauigkeit erreicht werden (RMSE-Wert). Da das Feedback mit realen Messwerten jedoch im Kontext einer Betriebsstrategieoptimierung und auch im Alltagsbetrieb eines Fahrzeuges für Regelfunktionen nicht möglich ist, werden die Shallow Learning Algorithmen in dieser Arbeit nicht weiter analysiert.

Bei den Deep Learning Algorithmen wurden ein Vorwärtsgerichtetes Neuronales Netz (FNN) und ein Rekurrentes Neuronales Netz, spezifisch ein Long Short-Term Neuronales Netz, für die Modellierung eingesetzt. Beide Modelle konnten in Bezug auf die Architektur so gestaltet und trainiert werden, dass die Emissionswerte für beliebig viele Zeitschritte berechnet werden konnten und hierzu lediglich verfügbare Eingangsgrößen und beim Rekurrenten Neuronalen Netz zusätzlich interne Informationen (Zustände, vergangene Ausgänge) notwendig waren.

Die in Kapitel 2.3.2 angeführten potenziellen Vorteile des RNN gegenüber dem FNN bei der Darstellung zeitlich basierter Vorgänge konnten auch in [60] bei der Modellierung transienter Emissionen nachgewiesen werden. Dies zeigt sich in der Wurzel der mittleren quadratischen Abweichung der Prädiktionswerte gegenüber den Messwerten, welche bei dem RNN circa 10 % niedriger im Vergleich mit dem FNN ist. Daher wird im weiteren Verlauf auf die Long Short-Term Memory Architektur als vielversprechendstem Ansatz hinsichtlich der in Kapitel 4 formulierten Ziele und der zeitlich basierten Vorgänge im Verbrennungsmotor fokussiert.

### 6.2.2 Aufbereitung der Messdaten

Bei der Aufbereitung der Messdaten sind für die datenbasierte Modellierung zwei grundsätzliche Aspekte relevant:

1. Die Größenordnung der Messwerte
2. Die zeitliche Synchronisierung der Messsignale

Dies wurde bei der bisherigen, physikalisch-phänomenologischen Modellierung nicht berücksichtigt, da die grundlegenden Berechnungen auf die Magnitude realer (SI-) Größen ausgelegt sind und die Einheiten bei Bedarf angeglichen werden können. Weiterhin entfällt die zeitliche Synchronisierung der Messsignale aus verschiedenen Messsystemen, da in Kapitel 6.1 für die Kalibrierung und Optimierung stationäre Messdaten eingesetzt werden. Dies bedeutet, dass die Einschwingphase für die Realisierung (theoretisch) gleichbleibender Bedingungen die angenommene Verzögerungszeit deutlich übersteigt.

**Größenordnung der Messwerte:** Beim Einsatz von Machine Learning Modellen ist die Normalisierung von Eingangsgrößen weitverbreitet, da durch die Angleichung der Größenordnung der Trainingsvorgang effizienter ablaufen kann [107]. Dies lässt sich unter anderem durch die Betrachtung der in Kapitel 2.3.2 eingeführten Aktivierungsfunktionen nachvollziehen. Wenn die Eingangswerte in diesen Funktionen entsprechend groß und die Gewichtungen im frühen Trainingsverlauf nicht darauf angepasst sind, wird das betreffende Neuron als gesättigt bezeichnet und der Trainingsprozess (Anpassung der Gewichtung in ein passendes Fenster) kann langsamer ablaufen.

Unabhängig von der Wahl der Modelleingangsparameter (Kapitel 6.2.3) zeigen sich bereits an den Emissionsspezies als Modellausgangsgrößen Unterschiede in den Größenordnungen der Werte. Diese liegen beispielsweise für unverbrannte Kohlenwasserstoffe im Bereich von Hunderten bis (mehreren) Tausenden ppm, wohingegen Kohlenstoffdioxid in der Regel fünfstellige ppm-Werte annimmt. Um bereits im Vorfeld zusätzliche Herausforderungen bei der Fehlerrückführung zu vermeiden, werden neben den Eingangsparametern auch die Ausgangsparameter normalisiert. Hierzu wird eine Min-Max Normalisierung eingesetzt, die die Datensignale isoliert betrachtet linear in einen Wertebereich zwischen 0 und 1 skaliert [108], was sich formal nach [108] mit folgender Gleichung ausdrücken lässt:

$$x'_{i,n} = \frac{x_{i,n} - \min\,(x_i)}{\max(x_i) - \min\,(x_i)} * (nMax - nMin) + nMin \tag{31}$$

In obiger Gleichung zeigen min und max die jeweiligen Minimal- und Maximalwerte der Größe $i$ (beispielsweise die Kohlenstoffmonoxid-Konzentration). $nMax$ und $nMin$ sind die Reskalierungsfaktoren 1 und 0. $x_{i,n}$ beschreibt den ursprünglichen Wert und $x'_{i,n}$ den Min-Max normalisierten Wert.

Das gewählte Vorgehen im Training und in der Anwendung des trainierten Modells (Inferenz) ist in Abbildung 6-15 dargestellt. Im Trainingsprozess werden die Trainingsdaten zusammengefasst und entsprechend Gleichung 31 in Bezug auf die jeweilige Messgröße Min-Max normalisiert. Die normalisierten Trainingsdaten werden dem Basismodell übergeben und der Trainingsprozess kann gestartet werden. Dabei werden die Gewichtungen im Modell entsprechend den bereitgestellten Daten angepasst, was bedeutet, dass das Training des Modells neben den ursprünglichen Daten auch von deren Normalisierung abhängt und diese für die Anwendung des Modells mit neuen Daten reproduziert werden muss. Um dies zu gewährleisten, werden die Normalisierungsparameter (Minimal- und Maximalwerte) während des Normalisierungsvorgangs in einer strukturierten Textdatei (.json – JavaScript Object Notation) gespeichert.

Jene Parameter können bei der Inferenz genutzt werden, um eine äquivalente Skalierung der Testdaten zu erhalten. Daher müssen die Testdaten nicht zwingend den Größenbereich von 0 bis 1 ausfüllen, vielmehr wird eine korrekte Repräsentation der Datenverläufe im trainierten Wertebereich gewährleistet.

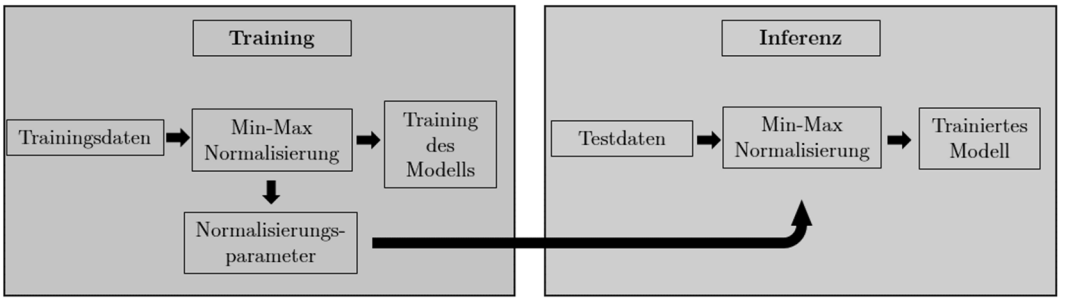


Abbildung 6-15: Normalisierung der Daten im Training und in der Inferenz

Durch den Normalisierungsvorgang liefert das trainierte Modell die Emissionsprädiktionen ebenfalls als skalierte Modellausgänge. Um ein Ergebnis im ursprünglichen ppm-Bereich zu erhalten, werden die Normalisierungsparameter erneut eingesetzt und die Normalisierung umgekehrt.

**Zeitliche Synchronisierung der Messsignale:** Wie in Kapitel 5.1.2 beschrieben, stammen die Messsignale aus verschiedenen Quellen/Messsystemen. Diese sind über Datenaustauschsysteme (CAN-Bus, TCP/IP) mit den Prüfstandsrechner verbunden, sodass die Messsignale gebündelt und gespeichert werden können. Durch die eingesetzte Sensorik, die Auswertung und Übertragung kommt es zwangsläufig zu einem zeitlichen Versatz zwischen einem Ereignis

und dessen Erfassung im Datenspeicher. Der zeitliche Versatz ist zudem aufgrund der unterschiedlichen Messsysteme nicht identisch für alle Messwerte.

Die genaue Bestimmung der Signalverzögerung ist nicht immer direkt möglich, beim vorliegenden Messaufbau lassen sich jedoch grundsätzlich zwei Fälle unterscheiden und einordnen. Die meisten Messwerte entstammen den Sensoren (Druck, Temperatur, Durchfluss, etc.), welche unmittelbar an der Messstelle angebracht sind. Die Signale werden verarbeitet (elektronisch) und digital an den Prüfstandsrechner gesendet. Eine Quantifizierung dieser Dauer wurde nicht vorgenommen, es kann jedoch davon ausgegangen werden, dass Sie deutlich geringer als ein Zeitschritt der allgemeinen Messfrequenz (5 Hz, entsprechend 0.2 s) ist. Im Kontrast dazu steht die Messung der Emissionen, bei der die Entnahmestelle am Prüfstandsmotor und die Analysatoren räumlich getrennt sind. Das Abgas wird in einem Teilvolumenstrom mittels einer beheizten Messgasleitung über eine circa 10 m lange Strecke zum Messgerät AVL i60 FT SII transportiert, wo die eigentliche Messung stattfindet. Die Distanz resultiert aus den erforderlichen Umgebungsbedingungen für den optimalen Betrieb des Messgerätes und den räumlichen Gegebenheiten des Prüfstands. Obwohl das Messsystem über eine Messgaspumpe mit erhöhtem Volumenstrom verfügt, ist davon auszugehen, dass die Gaslaufzeit von der Messstelle zu den Analysatoren relevant ist und deutlich über einem Zeitschritt von 0.2 s liegen kann.

Die gemessenen Emissionswerte werden im weiteren Verlauf für das Training der datenbasierten Emissionsmodelle genutzt, welche durch Optimierung der internen Parameter die Beziehung zwischen den gewählten Eingangsgrößen und den Emissionsspezies erlernen. Wenn jedoch ein erheblicher zeitlicher Versatz in der Messung besteht, ordnet das Modell die Ausgänge (Wirkung; Emissionswert) den falschen Eingängen (Ursache; Motorparameter) zu. Dies kann besonders bei der Verwendung des Modells für spätere Optimierungs- oder Regelungsfunktionen problematisch sein, weshalb die Gaslaufzeit bestimmt werden muss.

Hierzu wurde eine Parameterkombination aus einer schnell veränderlichen, unmittelbar messbaren Größe und einer damit korrelierenden Emissionsspezies gesucht, um durch ein darauf abgestimmtes Lastprofil des Versuchsmotors die Verzögerung bestimmen zu können. Ein solcher Zusammenhang kann in positiver Korrelation zwischen dem gemessenen Motordrehmoment (der Last) und der Stickstoffoxidkonzentration beobachtet werden. Dies lässt sich in der Theorie durch die Steigerung der Verbrennungsspitzentemperatur und die damit verbundenen Auswirkungen auf den Zeldovich-Mechanismus (thermisches *NO*) und die Prompt-*NO* Bildung begründen (siehe Kapitel 2.2.1). In der Theorie wurde ebenfalls erläutert, dass die *NO*-Bildung in der primären Reaktionszone unmittelbar stattfindet und damit keine weitere Zeitverzögerung angenommen wird.

Um den zeitlichen Versatz (interpretiert als Gaslaufzeit) zwischen dem Motordrehmoment- und dem Stickstoffoxidsignal zu untersuchen, wurde ein Messprogramm entwickelt und durchgeführt, welches bei gleichbleibender Drehzahl (2000 $min^{-1}$) verschiedene, lineare Lastrampen zwischen dem Leerlauf (Nulllast) und 200 Nm abfährt. Dies ist in Abbildung 6-16 dargestellt.

Das Drehmoment ist im oberen Verlauf dargestellt, die gemessene Stickstoffoxidkonzentration im Unteren. Der Drehmomentverlauf besteht aus verschiedenen Rampenhöhen und -steigungen. Qualitativ ist eine starke, positive Korrelation der beiden Verläufe zu erkennen. Bei der Betrachtung der Startpunkte der positiven Drehmomentrampen im Vergleich mit dem dazugehörigen Anstieg der Stickstoffoxidkonzentration lässt sich unabhängig von der Ausgangslast und Steigung ein konstanter Zeitversatz erkennen. Dies ist in der folgenden Abbildung durch die vertikalen Balken dargestellt.

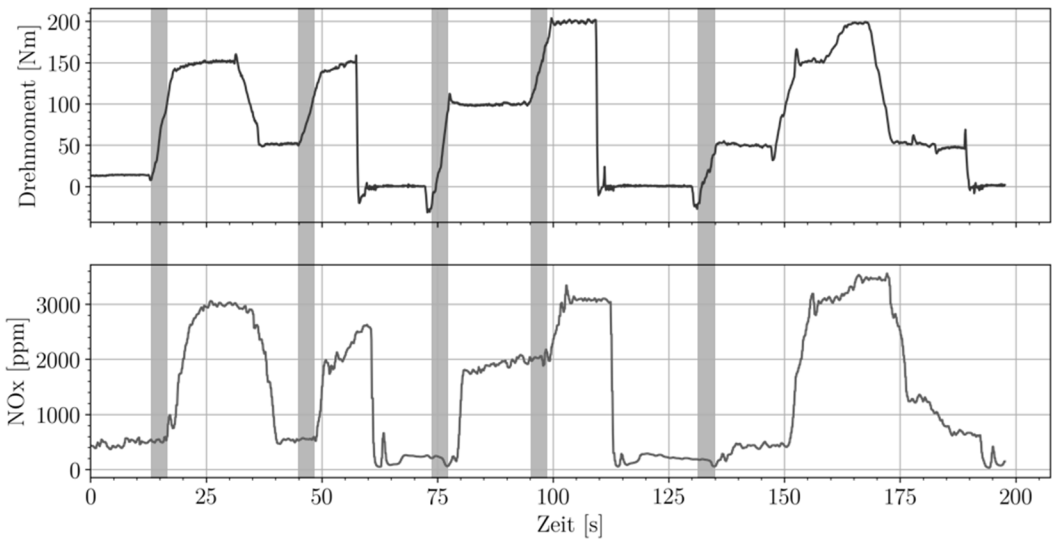


Abbildung 6-16: Drehmomentrampenprogramm bei 2000 $min^{-1}$. Dargestellt ist das Drehmoment, die Stickstoffoxidkonzentration und die Signalverzögerung als vertikale Balken

Der Zeitversatz beträgt knapp über 3 Sekunden. Um diesen Wert genauer bestimmen zu können, kann eine Korrelationsanalyse durchgeführt werden, in welcher das Stickstoffoxidsignal über mehrere Zeitschritte nach vorne verschoben wird, bis die Korrelation zwischen dem Drehmoment und der Stickstoffoxidkonzentration maximal ist. Dies wurde speziell für die positiven Drehmomentrampen durchgeführt, was zu einer Verschiebung von 16 Zeitschritten führt. Bei einer Messfrequenz von 5 Hz entspricht das 3.2 s und bestätigt damit die Beobachtungen aus Abbildung 6-16.

### 6.2.3 Bestimmung der Modelleingänge

Nach der Wahl des Machine Learning Algorithmus (Long Short-Term Memory Neuronales Netzwerk) und den gewünschten Prädiktionswerten bzw. Modellausgängen (Emissionswerte)

ist die Bestimmung der Modelleingänge ein essenzieller Schritt für die spätere Modellperformance in Berücksichtigung des jeweiligen Anwendungsfalles.

Wie in Kapitel 4 angedeutet, sollte das Emissionsmodell im zugrundeliegenden Projekt als Entscheidungskriterium für die Optimierung einer Betriebsstrategie für ein Hybridfahrzeug dienen. Dazu sind Eingangswerte erforderlich, welche sich einfach vorhersagen lassen, um ausgehend von einem Zeitpunkt mehrere Trajektorien für die Betriebsweise des Hybridfahrzeuges zu prädizieren und diese hinsichtlich des Emissionsverhaltens zu bewerten. Dabei kommen Motorparameter in Frage, die sich aus einem Betriebspunkt direkt ableiten lassen, wie etwa der Motordrehzahl, ausgehend von einer Fahrzeuggeschwindigkeit und der momentanen Übersetzung. Weiterhin können Größen betrachtet werden, die sich mit ausreichender Genauigkeit bestimmen lassen. Hierzu zählt beispielsweise der Zündzeitpunkt, der aus verschiedenen Kennfeldern aus dem Steuergerät ausgelesen werden kann und ebenso präzise und dynamisch real umgesetzt wird.

Eine weitere Überlegung für die Bestimmung der Eingangswerte ist die Relevanz für die zu prädizierenden Ausgangswerte. An dieser Stelle können theoretische Überlegungen, wie etwa die Berücksichtigung der physikalisch-phänomenologischen Zusammenhänge in Kapitel 2.2 hilfreich sein, weil die experimentell ermittelten Zusammenhänge auch für die datenbasierte Modellierung relevant sind. Da diese Grundlagen jedoch nicht immer vorausgesetzt werden können, wurde im Zuge der datenbasierten Modellierung eine mathematische Vorgehensweise für die Quantifizierung der Relevanz gewählt, siehe hierzu auch [60].

In Abbildung 6-17 ist eine Korrelationsmatrix ausgewählter Messgrößen dargestellt. Darunter finden sich einerseits Emissionswerte als Modellausgänge und andererseits mögliche Modelleingänge, welche die bisher definierten Kriterien erfüllen. In einem ersten Schritt wird der lineare Zusammenhang der skalierten (Min-Max normalisierten) Messgrößen mit Hilfe des Pearsonschen Maßkorrelationskoeffizienten untersucht. Dieser liefert Werte zwischen 1 (bei einer perfekt linearen Korrelation zwischen zwei Messgrößen), 0 (wenn kein linearer Zusammenhang besteht) und -1 (bei einem vollständig negativen linearen Verhalten). Die Werte in Abbildung 6-17 sind auf die zweite Nachkommastelle gerundet. Es ist zu beachten, dass der Pearsonsche Maßkorrelationskoeffizient nicht ausreichend ist, um eine Auswahl im vorliegenden Anwendungsfall zu treffen, da für die Emissionsentstehung ebenfalls nichtlineare Zusammenhänge angenommen werden. Hierzu kann zusätzlich der Spearmansche Rangkorrelationskoeffizient berechnet werden, siehe [109]. Die beschriebene statistische Auswertung der Zusammenhänge wurde für alle Messwerte durchgeführt, wobei die Darstellung in der folgenden Abbildung auf die im späteren Modellaufbau relevanten Größen reduziert ist.

In Abbildung 6-17 kann die Korrelation am Beispiel des Drehmoments (MD) auf der x-Achse beschrieben werden. In Zusammenhang mit den unverbrannten Kohlenwasserstoffen (THC) beträgt der Pearsonsche Maßkorrelationskoeffizient -0.49. Dies bedeutet, dass bei steigendem Drehmoment im Durchschnitt von einer sinkenden Konzentration der unverbrannten Kohlenwasserstoffe ausgegangen werden kann, wobei der absolute Wert eine mittlere lineare

Korrelation darstellt. Demgegenüber führt ein steigendes Drehmoment zu einem Anstieg der Stickstoffoxide. Der Maßkorrelationskoeffizient von 0.68 deutet auf eine starke lineare Korrelation hin, was beispielsweise in Abbildung 6-16 nachvollzogen werden kann.

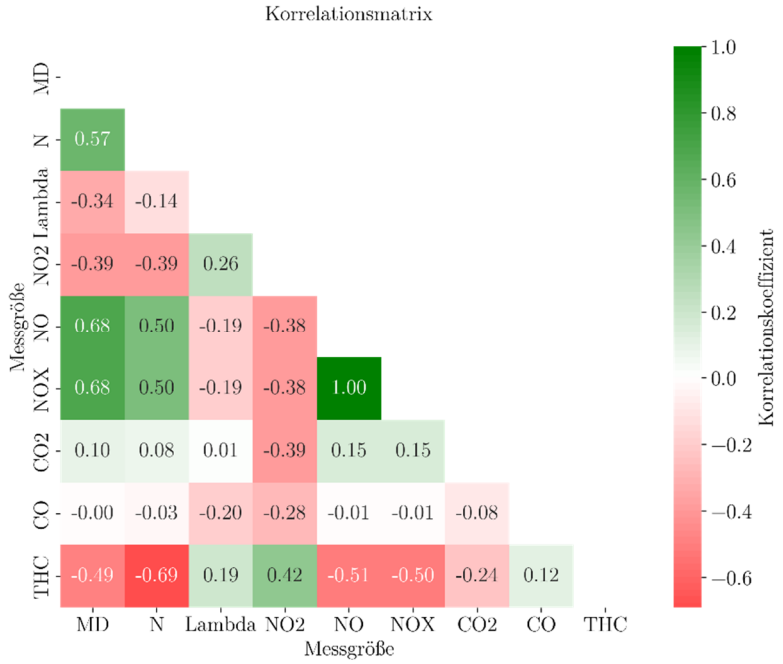


Abbildung 6-17: Korrelationsmatrix ausgewählter Eingangswerte und Prädiktionswerte nach [60]

Mit den Vorüberlegungen bezüglich der Modellverwendung und den Korrelationsanalysen wurden die Parameter Drehmoment (MD), Drehzahl (N) und das Luft-Kraftstoff-Verhältnis (Lambda) als Eingangsgrößen (Feature) für die datenbasierte Modellierung gewählt. Diese lassen sich in einem Optimierungsszenario hinreichend genau prädizieren und sind relevant in Bezug auf die untersuchten Emissionsspezies. Eine zu diesem Zeitpunkt noch offene Frage ist, ob die drei Eingangsparameter ausreichend für eine transiente Abbildung sind und in welchem Maße sich die Vorhersagegenauigkeit mit weiteren Features verbessern würde. Diese beiden Punkte werden in Kapitel 7 behandelt, wobei besonders in der Wahl möglicher Eingangsgrößen noch erhebliches Forschungspotenzial steckt.

### 6.2.4 Modellaufbau und Hyperparametertuning

Aufbauend auf den Erkenntnissen der vorherigen Kapitel, wird im Folgenden der Aufbau, das Training und das darin enthaltene Hyperparametertuning erläutert. Unter Hyperparametertuning wird das Optimieren der sog. Hyperparameter (z. B. Anzahl der Zellen pro Schicht,

Anzahl der Schichten) zur Steigerung der Prädiktionsqualität des Modells verstanden [110]. Um die Relevanz der Hyperparameter zu verdeutlichen, werden zuerst die schematische Modellstruktur (siehe Abbildung 6-18) und der Informationsfluss durch das Modell beschrieben und einige Parameter eingeführt (Dropout, Sequenz, etc.), welche im weiteren Verlauf des Kapitels definiert werden.

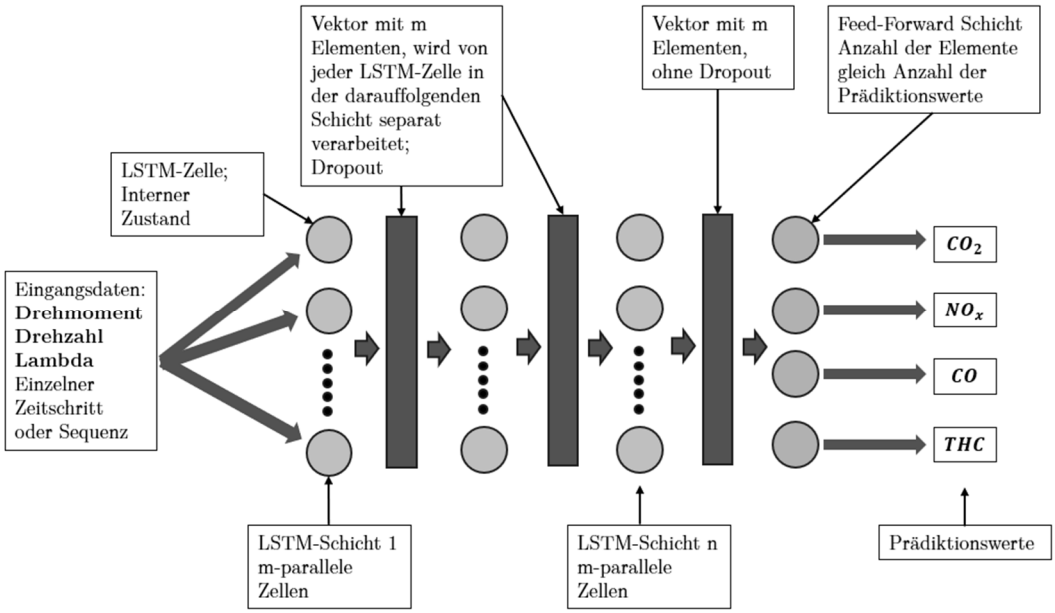


Abbildung 6-18: Schematischer Aufbau des datenbasierten Emissionsmodells

Der generelle Informationsfluss verläuft von den Eingangsdaten (links) über das Neuronale Netz (Kreise) zu den Prädiktionswerten auf der rechten Seite. Die Informationen werden dabei jedoch nicht strikt nur in eine Richtung verarbeitet. Durch die Integration der Long Short-Term Memory Zellen sind auch rekursive Schritte und die Verwendung interner Zustände enthalten.

Die Eingangsdaten Drehmoment, Drehzahl und Lambda werden dem Modell schrittweise (Zeitschritt gleich 0.2 s entsprechend der Messfrequenz) oder als Sequenz bestehend aus einer beliebigen Anzahl von Messschritten jeweils über einen Dataloader zur Verfügung gestellt. Die Wahl der Sequenzlänge ist dabei ebenfalls ein Parameter, welcher die Genauigkeit des Modells beeinflussen kann, und Optimierungspotenzial bietet. Die Eingangsdaten sind an dieser Stelle bereits normalisiert und werden in der ersten Schicht des Neuronalen Netzwerks verarbeitet. Diese besteht aus parallel angeordneten LSTM-Elementen der Anzahl $m$, von denen jedes die gleichen Informationen in Form der Features verarbeitet. Die Art der internen Verarbeitung ist jedoch von den im Training optimierten Eigenschaften (siehe Kapitel 2.3.2) der jeweiligen Zelle abhängig, wodurch unterschiedliche Aspekte der Eingangsdaten berücksichtigt werden. Die LSTM-Zellen können wie bereits erläutert nicht nur den aktuellen Zeitschritt beachten,

sondern haben über den internen Zustand auch Informationen über vergangene Ein- und Ausgabewerte, was die Ausgabe der Zelle beeinflussen kann.

Das Resultat der ersten Schicht ist ein Vektor (oder eine Sequenz von Vektoren) mit $m$ Elementen, welcher sich aus den Ausgaben der $m$ LSTM-Zellen zusammensetzt. In Abbildung 6-18 ist ebenfalls das sogenannte „Dropout“ enthalten, wobei es sich um eine Methode der Regularisierung handelt [111]. Diese wird ausschließlich im Training des datenbasierten Modells eingesetzt. Beim Dropout werden je nach gewählter Intensität (bspw. 0.3) eine gewisse Anzahl an Ausgabewerten der jeweiligen LSTM-Schicht gleich null gesetzt. Im vorliegenden Beispiel hat jede Zelle hierzu eine Chance von 30 %. Somit wird die Relevanz der einzelnen Zelle auf die abschließende Prädiktion herabgestuft und es wird verhindert, dass einzelne Zellen durch hohe Gewichtungen den Prädiktionswert dominieren, was zu Überanpassung führen kann.

Das beschriebene Verfahren wird fortgeführt, bis die letzte der insgesamt $n$ LSTM-Schichten erreicht wird. Diese verarbeitet die Informationen identisch bis auf die Tatsache, dass für ihren Ausgabevektor keine Regularisierung angewendet wird, da der Einfluss am „Netzende“ sehr groß wäre. Die Prädiktion der $n$-ten Schicht wird an eine Feed-Forward Schicht übergeben. Diese besteht aus exakt vier Elementen, was der Anzahl der Prädiktionswerte entspricht. Neben einer weiteren Informationsverarbeitung über interne Gewichtungen ist die Informationsbündelung der vorherigen Schichten eine zentrale Aufgabe der Feed-Forward Schicht. Je nach Wahl der Hyperparameter (in diesem Fall als Beispiel $m = 10$) ist zwischen der letzten und vorletzten Schicht ein Vektor mit zehn Elementen zu verarbeiten, was bedeutet, dass dieser auf vier Elemente reduziert werden muss. Der Vektor wird dabei parallel an jede der vier Zellen übergeben, welche über die eigenen zehn optimierten Gewichtungen und die gewählte Aktivierungsfunktion (bei Regressionsaufgaben häufig linear bzw. keine oder ReLU) den normalisierten Wert der jeweiligen Emissionsspezies berechnen. Dieser wird abschließend in die ppm-Konzentration übersetzt.

**Hyperparametertuning und Training:** Bisher wurden bereits einige Parameter eingeführt, welche als übergeordnete Modellgrößen und damit als Hyperparameter definiert werden können. Dazu gehören die Anzahl der LSTM-Schichten $n$, die Menge der parallelen LSTM-Zellen pro Schicht $m$ oder die Dropout-Intensität. Die Wahl der Hyperparameter bestimmt die Modellgröße und -komplexität, was sich sowohl beim Training als auch bei der Inferenz auf die Rechenzeit auswirkt. Ferner können die Hyperparameter einen Einfluss auf die Prädiktionsgenauigkeit beim Training und vor allem bei der Inferenz nehmen. Ein sehr komplexes Modell kann durch das Vorhandensein zahlreicher Gewichtungen den Verlauf der Trainingsdaten sehr genau abbilden, was jedoch zu einer hohen Varianz und einer schlechten Prädiktion der Testdaten führen kann (Überanpassung). Im Gegensatz dazu, kann ein sehr einfacher Modellaufbau dazu führen, dass bereits im Trainingsprozess keine gute Übereinstimmung zwischen Prädiktion und Messwerten gefunden wird, da das Modell die Vorgänge zu stark vereinfacht (Unteranpassung).

Die Wahl geeigneter Hyperparameter, die weder zu Über- oder Unteranpassung führen und zudem eine performante Inferenz ermöglichen, ist eine zentrale Herausforderung der datenbasierten Modellierung. Bei bekannten Problemen des Maschinellen Lernens (etwa der Klassifizierung mit logistischer Regression) kann in Abhängigkeit der Datentiefe und der erwarteten Komplexität mit Erfahrungswerten vergleichbarer Anwendungen gestartet werden. Dies ist bei der vorliegenden Problemstellung allerdings nicht der Fall. Dafür können neben den Gewichtungen auch bei den Hyperparametern Optimierungen durchgeführt werden, um die Wahl der bestmöglichen Parameterkombinationen zu automatisieren. Hierzu zählen beispielsweise die Rastersuche oder Bayessche Optimierung (siehe [112]).

Als Bewertungskriterium für die Modellgenauigkeit wurde der mittlere RMSE-Wert der vier prädizierten Emissionskonzentrationen in Bezug auf die Messwerte betrachtet. An dieser Stelle sind auch Gewichtungen denkbar, um die Modellgenauigkeit je nach Anwendung gezielt auf einzelne Emissionsspezies zu optimieren.

Für die Bestimmung der optimierten datenbasierten Modellstruktur wurden folgende Hyperparameter untersucht:

- Anzahl der Neuronen pro Schicht $m$ (hidden_size): Die Anzahl der Neuronen pro Schicht beeinflusst die Modellgröße/-komplexität über die Modellbreite und die damit verbundene Vektorlänge der ausgetauschten Informationen zwischen den Schichten.
    - Untersuchter Bereich: 2 - 30
- Anzahl der LSTM-Schichten $n$ (num_layers): Die Anzahl der LSTM-Schichten beeinflusst die Modellgröße/-komplexität über die Modelltiefe.
    - Untersuchter Bereich: 1 - 10
- Batch-Größe (batch_size): Die Batch-Größe gibt an, wie groß die Datensequenz ist, welche dem Modell während des Trainings zur Verfügung gestellt wird, bevor die Gewichtungen angepasst werden. Dies können ein kleiner Teil (bspw. 10 Sekunden realer Messzeit) oder alle verfügbaren Daten sein. Sie beeinflusst die Trainingsgeschwindigkeit (kleine Batch-Größe → viele Anpassungen der Gewichtungen) und die Fähigkeit zur Generalisierung (Über- bzw. Unteranpassung) des Modells in der Inferenz (Batch-Größe zu hoch → unter Umständen Überanpassung).
    - Untersuchter Bereich: 1 - 5000
- Dropout-Intensität (dropout): Dropout ist eine Methode, welche die Fähigkeit zur Regularisierung verbessern kann (siehe oben).
    - Untersuchter Bereich: 0.1 – 0.5
- Lernrate (learning rate): Die Lernrate ist ein Parameter, der die Anpassung der Gewichtungen im Trainingsvorgang durch das Gradientenverfahren beeinflusst.
    - Untersuchter Bereich: 0.00001 – 0.01 und adaptiv (Adam „Adaptive moment estimation“, siehe [113])
- Sequenzlänge (window size): Die Sequenzlänge gibt an, wie viele vergangene Zeitschritte für die Prädiktion des aktuellen Zeitschritts berücksichtigt werden. Dadurch

können langsam ablaufende zeitliche Phänomene besser beobachtet werden, was jedoch auch die Ansprechzeit des Modells auf schnelle Ereignisse reduzieren kann (Trägheit)
    - Untersuchter Bereich: 1 - 100

Für das Training wurden die in Kapitel 5 beschriebenen Daten verwendet, wobei ein Teil für die spätere Validierung bzw. das Testen zurückgehalten wurde. Der Trainingsdatensatz umfasst 154941 gemessene Zeitschritte, was circa 8.6 h RDE-ähnlicher Straßenfahrt entspricht. Die Daten wurden entsprechend den Ergebnissen von Kapitel 6.2.2 aufbereitet (normalisiert und zeitlich korrigiert). Während des Trainingsvorgangs werden die Eingangsdaten dem Modell zur Verfügung gestellt und über initiale Gewichtungen die ersten Prädiktionen erstellt. Diese werden mit den Messdaten verglichen und die Gewichtungen der Neuronen über Fehlerrückführung angepasst. Der Prozess ist iterativ und die Häufigkeit hängt von der Batch-Größe und der Anzahl der Trainingsepochen (Durchlauf des kompletten Trainingsdatensatzes) ab. Nach jeder Trainingsepoche werden die RMSE-Werte bezüglich des Trainings- und des Validierungsdatensatz abgespeichert, um den Trainingsfortschritt beurteilen zu können. In Abbildung 6-19 sind beispielhaft drei unterschiedliche Trainingsverläufe dargestellt, bestehend aus den Prädiktionsungenauigkeiten der drei Modelle bezogen auf die Trainingsdaten über 2000 Trainingsepochen. Zu Gunsten der besseren qualitativen Interpretierbarkeit der Ergebnisse sind in den nachfolgenden Diagrammen die Ordinatenachsen jeweils mit dem Ausdruck „Prädiktionsungenauigkeiten“ beschriftet, sodass ein abfallender Verlauf mit einer reduzierten Ungenauigkeit in der Modellvorhersage (und damit gleichbedeutend einer gesteigerten Prädiktionsgenauigkeit) einhergeht.

Für das dargestellte Training wurde der komplette Trainingsdatensatz verwendet. Die Modelle bestehen jeweils aus zwei LSTM-Schichten mit je sieben Neuronen, es wird eine Dropout Intensität von 0.1 angewendet und die Batch-Größe beträgt 50. Für die Prädiktion des aktuellen Wertes wurden die letzten 10 Zeitschritte als Sequenzlänge beachtet. Der Unterschied in den drei Modellen liegt in der Lernrate. Für den schwarzen Kurvenverlauf wurde eine zyklische wechselnde Lernrate zwischen 0.001 und 0.0001 gewählt, welche alle 50 Epochen geändert wurde. Dies ist ein häufig eingesetztes Verfahren, um den Einfluss der Lernrate auf den Trainingsprozess besser einschätzen zu können. Die graue Lernkurve basiert auf einem Modell mit konstanter Lernrate von 0.001, die gestrichelte auf einer Lernrate von 0.0001.

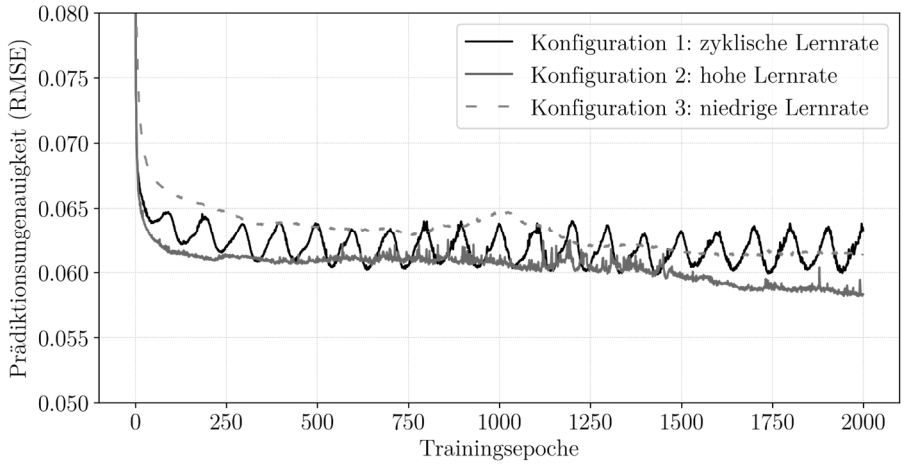


Abbildung 6-19: Verlauf des RMSE-Wertes über einen Trainingsvorgang von 2000 Epochen mit drei verschiedenen Lernraten

Anhand der in Abbildung 6-19 dargestellten Auswirkung der Lernrate der drei gewählten Konfigurationen wird der Einfluss eines einzelnen Hyperparameters verdeutlicht. Das Modell mit zyklisch wechselnder Lernrate (schwarz) zeigt einen sinusförmigen Verlauf. Dies resultiert aus der alle 50 Epochen veränderten Lernrate, was dazu führt, dass die Prädiktionsungenauigkeit über mehrere Epochen zunimmt und danach wieder abnimmt. Das lässt darauf schließen, dass die gewählten Lernraten in Kombination nicht geeignet sind, um sich dem Optimum gezielt zu nähern. Der sinusförmige Verlauf ist anfangs zwar noch erkennbar negativ, scheint ab Epoche 1250 dagegen im Mittel horizontal zu stagnieren. Die geringe Lernrate (gestrichelt) zeigt bei Trainingsbeginn die langsamste Verbesserung, da die Gewichtungen im Neuronalen Netzwerk entsprechend wenig pro Iteration angepasst werden. Dahingegen sind von Epoche zu Epoche keine großen Schwankungen zu erkennen, wie etwa bei der hohen Lernrate (grau). Die niedrige Lernrate liefert im gezeigten Beispiel die schlechteste Vorhersagegenauigkeit. Dies kann daran liegen, dass aufgrund der geringen Optimierungsgeschwindigkeit die Anzahl der Trainingsepochen nicht ausreicht, oder nur ein lokales und kein globales Minimum angenähert wird. Das Modell mit der hohen Lernrate erreicht mit einem RMSE von 0.058 über die 2000 Epochen den besten Wert (ca. 10 % besser als Konfiguration 3) und anhand des noch abfallenden Verlaufes ist bei längerem Training noch von weiteren Verbesserungen auszugehen. Das gewählte Beispiel verdeutlicht die Relevanz der Hyperparameter für die Vorhersagegenauigkeit, weshalb auch an dieser Stelle der Einsatz performanter Optimierungsmethoden sinnvoll ist. Weiterhin ist zu beachten, dass ein präzise trainiertes Modell keine alleinige Aussage auf die spätere Modellanwendung in der Inferenz mit unbekannten Eingangsdaten zulässt (Über- und Unteranpassung). Daher ist es hilfreich, bereits im Training und/oder beim Hyperparametertuning einen Validierungsdatensatz einzusetzen. Dieser wird nicht direkt genutzt, um die Gewichtungen anzupassen (bleibt also „unbekannt“), sondern dient vielmehr als Evaluationswerkzeug, um die Prädiktionsgenauigkeit gegenüber neuen Daten zu bewerten und damit

primär die Überanpassung des Modells auf die Trainingsdaten zu reduzieren. Abschließend kann das trainierte und validierte Modell mit zusätzlichen Daten (Testdaten) getestet werden. Dies wird in Kapitel 7.2 untersucht. Der dritte Datensatz (Testdaten) ist für eine faire Bewertung des Modells trotz der bereits verwendeten Validierungsdaten erforderlich. Letztere sind zwar im Training nicht direkt beteiligt und die Modellgewichtungen werden nicht auf die Validierungsdaten angepasst, jedoch kann die Festlegung der Hyperparameter auf eine gute Prädiktion der Validierungsdaten ebenfalls als eine Optimierung interpretiert werden und die Validierungsergebnisse können zu optimistisch ausfallen [114]. Dadurch ist die finale Beurteilung des Modells mit den Testdaten hilfreich.

Für die Wahl der Hyperparameter wurde eine Bayessche Optimierungsmethode namens „Tree-structured Parzen Estimator“ (TPE, siehe [115]) angewendet, welche in den vergangenen Jahren sehr gute Ergebnisse bei der Hyperparameteroptimierung komplexer Modellstrukturen erzielen konnte [116]). Hierzu wurde entsprechend den bereits eingeführten Hyperparametern und der gewählten Bereiche ein Suchraum aufgespannt, um die optimale Hyperparameterkonfiguration zu bestimmen. Für die Anpassung der Gewichtungen der einzelnen, vom Algorithmus gewählten Testkonfigurationen wurden die vollständigen Trainingsdaten genutzt. Als Bewertungskriterium für den Erfolg des TPE wurde hingegen der RMSE-Wert bezüglich des Validierungsdatensatzes berechnet. Die Hyperparameteroptimierung wurde zu Gunsten der Rechendauer über eine reduzierte Anzahl an Epochen (500) durchgeführt und die besten Konfigurationen anschließend über einen längeren Zeitraum (Mindestens 2000 Epochen, ohne Konvergenz entsprechend länger) trainiert. In Abbildung 6-20 ist das Ergebnis der Hyperparameteroptimierung dargestellt.

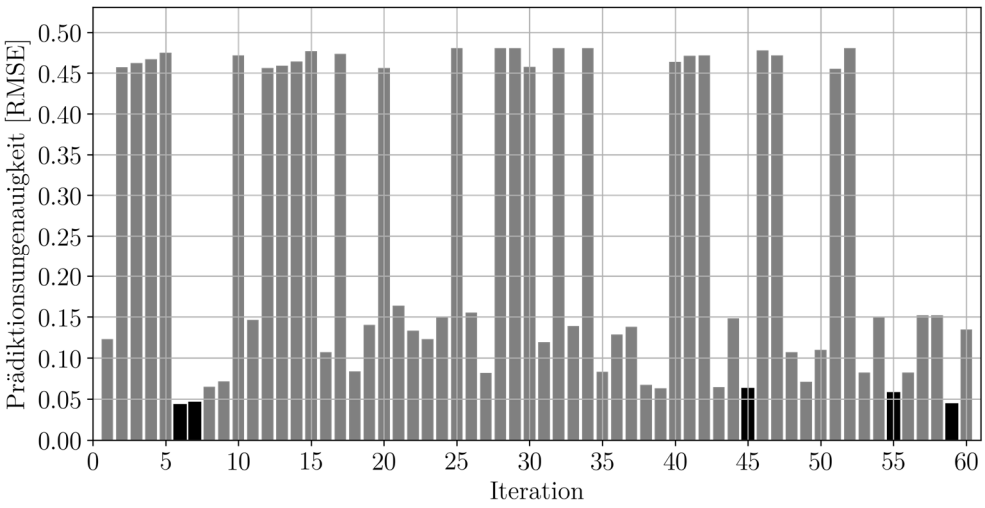


Abbildung 6-20: Prädiktionsungenauigkeit bezüglich des Validierungsdatensatzes bei der Hyperparameteroptimierung mit TPE. Die schwarz eingefärbten Balken sind die fünf besten Hyperparameterkonfigurationen

Dargestellt sind 60 Konfigurationen bzw. Iterationen („Trials“), welche der TPE-Algorithmus in Abhängigkeit der Randbedingungen (Hyperparameter und deren Bereiche) und des zu optimierenden Kriteriums (RMSE bezogen auf den Validierungsdatensatz) erstellt. Diese Konfigurationen werden trainiert und bewertet. Die Ergebnisse des Optimierungslaufes lassen sich in drei Bereiche zusammenfassen: Konfigurationen mit schlechtem Validierungsergebnis (RMSE 0.45 bis 0.50), Konfigurationen mit mittlerem Validierungsergebnis (RMSE 0.10 bis 0.15) und Konfigurationen mit gutem Validierungsergebnis (RMSE 0.04 bis 0.07). In die erste Kategorie fallen hauptsächlich Modelle, in welchen verschiedene Hyperparameter an den jeweiligen Randbereichen (Extremwerte) getestet wurden. Dazu gehören beispielsweise die Wahl von nur einem LSTM-Neuron pro Schicht, Dropout-Intensitäten von 0.5 oder sehr geringe Lernraten (0.00001). Die mittleren Ergebnisse bilden eine große Bandbreite möglicher Hyperparameterkonfigurationen ab und lassen sich nur bedingt klassifizieren. Die schwarz eingefärbten Balken, detailliert in Abbildung 6-20, repräsentieren die fünf besten Hyperparameterkonfigurationen und sind in Tabelle 6-1 dargestellt:

Tabelle 6-1: Hyperparameter der fünf besten Konfigurationen des datenbasierten Modells

| Nummer | Neuronen pro Schicht | Anzahl der Schichten | Batch-Größe | Dropout-Intensität | Lernrate | Sequenzlänge | RMSE |
|---|---|---|---|---|---|---|---|
| 6 | 20 | 8 | 50 | 0.279 | 0.000049 | 41 | 0.0429 |
| 7 | 20 | 2 | 50 | 0.204 | 0.000594 | 71 | 0.0457 |
| 45 | 2 | 10 | 50 | 0.129 | 0.000126 | 91 | 0.0626 |
| 55 | 2 | 9 | 50 | 0.250 | 0.000097 | 71 | 0.0575 |
| 59 | 20 | 8 | 50 | 0.223 | 0.000245 | 61 | 0.0439 |

Es zeigt sich, dass tendenziell zwei Netzaufbauten vielversprechende Ergebnisse liefern. Dazu gehören breite Netze (20 Neuronen pro Schicht) und tiefe Netze (mehr als neun LSTM-Schichten). Eine Ausreizung beider Parameter an die untersuchten Grenzen führt zu keiner weiteren Verbesserung, was an einer möglichen Überanpassung der Trainingsdaten und einer damit verbundenen schlechteren Generalisierungsfähigkeit liegen kann. Dies lässt sich beispielsweise an Aufbau Nummer acht erkennen, welche bei gleicher Breite eine weitere LSTM-Schicht hat, der RMSE-Wert bezogen auf die Validierungsdaten jedoch circa 50 % schlechter ist.

Im dargestellten Optimierungslauf wurde die beste Konfiguration bereits in der sechsten Iteration gefunden. Weiterhin ist zu erkennen, dass noch in verschiedene Richtungen exploriert wird und sich die guten Ergebnisse gegen Ende des Versuchs zwar verdichten, das anfängliche Ergebnis jedoch nicht weiter verbessert werden kann. Um dies zu bestätigen, wurde ein zusätzlicher Optimierungslauf gestartet und die Bandbreite der Hyperparameterwerte um die

besten Ergebnisse des Vorlaufes konzentriert. Daraus resultieren jedoch keine weiteren Verbesserungen, weshalb die in Tabelle 6-1 gezeigten Modelle als potenzielle Kandidaten für die weitere Modellierung verwendet werden. Da die aktuellen RMSE-Werte auf lediglich 500 Epochen basieren, ist die Durchführung von erweiterten Trainingsvorgängen notwendig, was in Abbildung 6-21 für die beste Konfiguration (Iteration Nummer sechs) dargestellt ist.

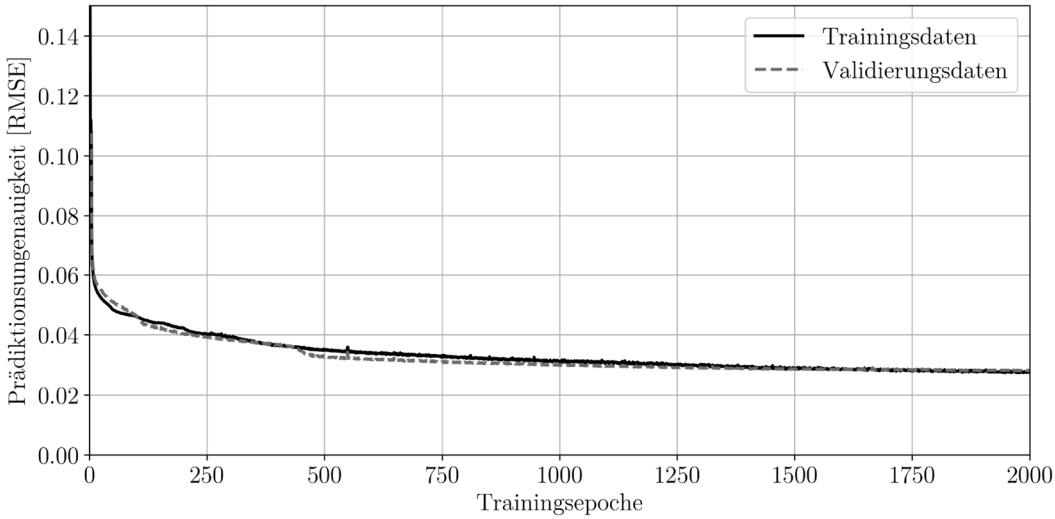


Abbildung 6-21: Prädiktionsungenauigkeit des datenbasierten Modells gegenüber den Trainings- und Validierungsdaten über 2000 Epochen

In der Abbildung ist die Prädiktionsungenauigkeit bezüglich der Trainingsdaten (schwarz) und der Validierungsdaten (grau, gestrichelt) über 2000 Epochen aufgetragen. Es ist generell zu erkennen, dass die Kurven bis auf die ersten circa zehn Epochen sehr gut übereinstimmen und das Modell beide Datensätze vergleichbar abbilden kann. Dies spricht für eine gute Generalisierung des Modells, was für die Prädiktion unbekannter Daten hilfreich ist. Qualitativ ist der Trainingsverlauf gleichmäßig (im Gegensatz zu Abbildung 6-19) und weist keine größeren Sprünge oder Oszillationen auf. Dies liegt neben dem Modellaufbau und den zugrundeliegenden Daten auch an der geringen Lernrate für das finale Training (0.0005). Erst in den letzten 250 Epochen nähern sich beide Verläufe der Horizontalen an. Bei Trainingsepoche 1990 ist die minimale Abweichung zu den Validierungsdaten erreicht mit einem RMSE von 0.02796. Dies ist nochmals circa 35 % besser als der Wert, der bei der Hyperparameteroptimierung erreicht wurde. Weitere Trainingsepochen zeigen keine Verbesserungen mehr und es ist eine Konvergenz zu erkennen.

## 6.3 Hybride Modellierungsansätze

Im nachfolgenden Kapitel werden Möglichkeiten erläutert, wie eine Verknüpfung zwischen physikalisch-phänomenologischen und datenbasierten Modellen geschaffen werden kann, um die Vorzüge beider Ansätze zu vereinen. Diese Integration zielt darauf ab, die Prädiktionsgenauigkeit zu verbessern. Das Kapitel beginnt mit der Definition von Voraussetzungen für die Entwicklung effizienter hybrider Modelle. Die Bedeutung einer einheitlichen Programmiersprache und fortschrittlicher Algorithmen zur Verarbeitung und Analyse dieser Daten sei hervorgehoben, da hiermit die Modelle sowohl robust als auch performant gestaltet, trainiert und angewendet werden können.

Anschließend werden zwei grundlegende Architekturen hybrider Modelle vorgestellt: die parallele und die serielle Architektur. Die parallele Architektur betrachtet die unabhängigen Prädiktionen beider Modelltypen und vereint diese zu einer Ausgabe, während die serielle Architektur das Potenzial eines Modells nutzt, um die Informationsdichte des anderen zu verbessern.

### 6.3.1 Voraussetzung für die Kombination der Modellansätze

Bevor die physikalisch-phänomenologischen und die datenbasierten Modellansätze zu einem oder mehreren hybriden Modellen kombiniert werden können, gilt es, die Voraussetzungen für diesen Prozess zu betrachten. Das physikalisch-phänomenologische Modell wurde mit der proprietären Software GT-Suite entwickelt, die nur über begrenzte Schnittstellen verfügt. Im Gegensatz dazu wurde das datenbasierte Modell unter Verwendung des Open Source ML-Frameworks PyTorch in Python erstellt. Aufgrund der unterschiedlichen Systemarchitekturen ist eine direkte Integration beider Modelle nicht möglich. Ebenso ist eine Co-Simulation aufgrund der hohen Rechenintensität und Ineffizienz von GT-Suite im Vergleich zu den schnellen ML-Frameworks, die in C-Code geschrieben sind und von der Nutzung zahlreicher GPUs profitieren, nicht praktikabel. Daher wurde eine Methode entwickelt, um das physikalisch-phänomenologische Modell auf das Leistungsniveau des datenbasierten Modells anzuheben, bei gleichzeitiger Überführung in das PyTorch ML-Framework. Eine schematische Darstellung dieser Methode ist in Abbildung 6-22 dargestellt.

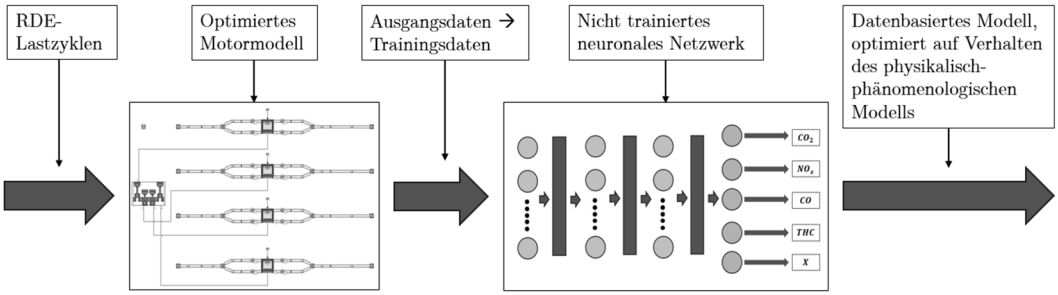


Abbildung 6-22: Transformation des physikalisch-phänomenologischen Modells in ein Neuronales Netzwerk

Die Grundlage bildet das in Kapitel 6.1.4 vorgestellte Motormodell, welches die physikalisch-phänomenologischen, optimierten Verbrennungs- und Emissionsmodelle umfasst. Dieses Motormodell kann für die Beschreibung des gewählten Vorgehens als virtueller Prüfstandsmotor betrachtet werden und wird in GT-Suite mit dem identischen, circa 10 Stunden dauernden RDE-Lastprofil (siehe Kapitel 5.3) wie der reale Prüfstandsmotor dynamisch betrieben. Die Ausgangsdaten beinhalten sowohl die bisher betrachteten Emissionsspezies als auch zahlreiche weitere Modellgrößen, stehen jedoch erst am Ende des Versuchslaufes (welcher auch mit der in Kapitel 6.1 beschriebenen performanten Hardware nicht in Echtzeit abläuft) zur Verfügung. Um diesen Vorgang zu beschleunigen und gleichzeitig beliebige Anpassungen an der Datenverarbeitung vornehmen zu können, werden die Ausgangsdaten als Trainings-, Validierungs- und Testdatensatz für ein weiteres Neuronales Netzwerk (fortan als „Imitationsmodell“ bezeichnet) genutzt. Dieses wird entsprechend trainiert, um das dynamische Verhalten des physikalisch-phänomenologischen Motormodells bestmöglich abzubilden. Dabei wird darauf geachtet, dass die gleichen drei Eingänge wie in Kapitel 6.2 genutzt werden. Als Prädiktionswerte werden erneut die vier Emissionsspezies verwendet. Der entscheidende Unterschied zum vorherigen datenbasierten Modell besteht darin, dass dieses Mal die vom Motormodell errechneten (bereits optimierten) und nicht die gemessenen Emissionswerte im Training genutzt werden. Letzteres würde unweigerlich zu einem zweiten, vergleichbaren datenbasierten Emissionsmodell führen. Im vorliegenden Fall soll hingegen gezielt das physikalisch-phänomenologische Verhalten nachgebildet werden, um neue Informationen in den hybriden Modellansätzen nutzen zu können. Wie in Abbildung 6-22 angedeutet, können neben den bereits erwähnten Prädiktionswerten auch noch weitere aus dem physikalisch-phänomenologischen Modell erlernt werden. Hierzu gehören beispielsweise die Verbrennungsspitzentemperatur oder der maximale Druckgradient im Brennraum. Diese Größen könnten die Informationsdichte in einer hybriden Modellstruktur weiter erhöhen und dadurch die Vorhersagegenauigkeit steigern, was in Kapitel 6.3.3 untersucht wird.

Das Vorgehen für die Modellerstellung und das Hyperparametertuning des datenbasierten Imitationsmodells entspricht dem des vorherigen Kapitels. In der folgenden Abbildung ist der schematische Aufbau des Modells dargestellt.

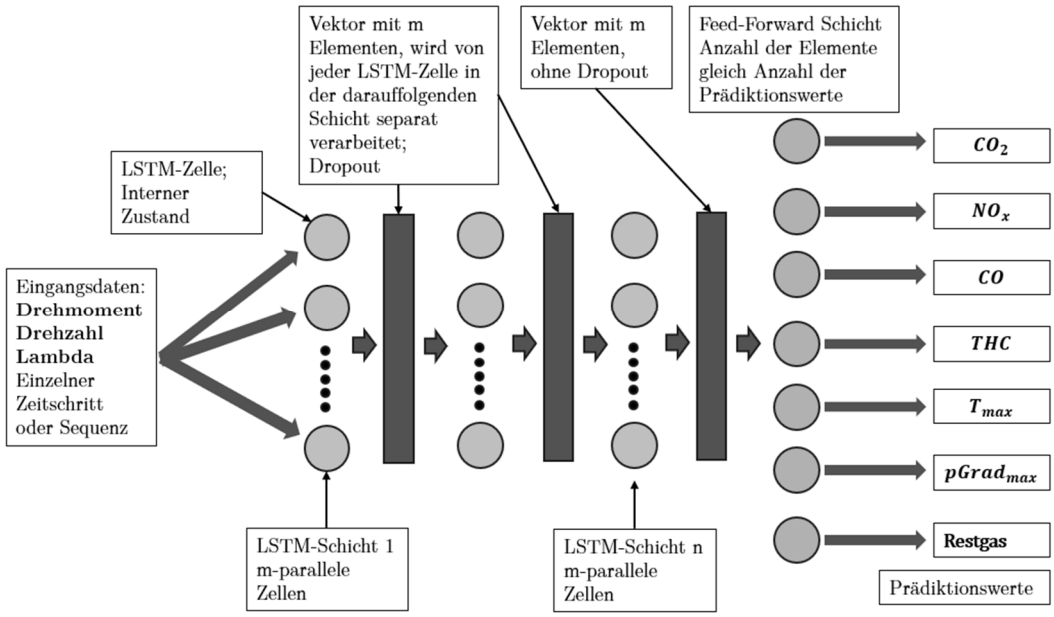


Abbildung 6-23: Schematischer Aufbau des datenbasierten Imitationsmodells

Die Architektur ist ebenfalls sehr ähnlich mit dem bereits vorgestellten Neuronalen Netzwerk. Dies liegt daran, dass die Modellierungsaufgabe vergleichbar ist und sich die gewählten Algorithmen äußerst gut für die Abbildung von zeitbasierten Regressionsaufgaben eignen. Erneut werden dem Modell drei Eingänge zur Verfügung gestellt, hiermit werden jedoch sieben Ausgangswerte prädiziert. Hinzugekommen sind die Spitzentemperatur während der Verbrennung ($T_{max}$), der maximale positive Druckgradient ($pGrad_{max}$) und die Masse des Restgases vor der Verbrennung. Dies sind im Hinblick auf die Emissionsbildung wichtige Parameter, welche jedoch aufwändig zu messen sind. Die Auswahl der zusätzlichen Prädiktionswerte ist in der konzeptionellen Phase der hybriden Modellierungsansätze noch nicht optimiert und könnte noch weiteres Verbesserungspotenzial bieten.

Für die Bestimmung der Hyperparameter wird die in Abschnitt 6.2.4 eingeführte Methodik verwendet. In Abbildung 6-24 ist die Hyperparameteroptimierung des Imitationsmodells dargestellt. Die 60 betrachteten Konfigurationen wurden über jeweils 500 Epochen trainiert; der beste Prädiktionswert bezüglich des Validierungsdatensatzes ist im Diagramm dargestellt.

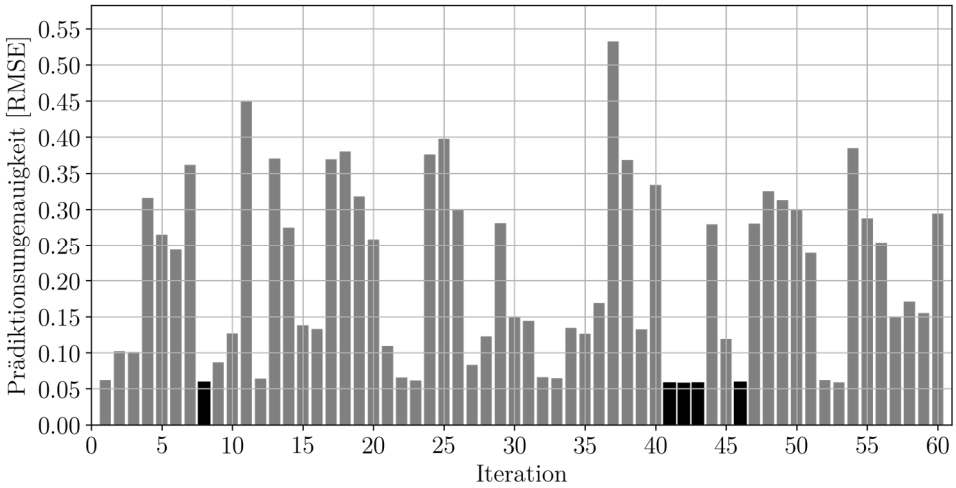


Abbildung 6-24: Prädiktionsungenauigkeit des Imitationsmodells bezüglich des Validierungsdatensatzes bei der Hyperparameteroptimierung mit TPE. Die schwarz eingefärbten Balken sind die fünf besten Hyperparameterkonfigurationen

Die fünf besten Konfigurationen sind in der oberen Abbildung in Schwarz hervorgehoben und in Tabelle 6-2 im Detail dargestellt. Es wird deutlich, dass primär breite Netzarchitekturen gute Ergebnisse erzielen, die Tiefe jedoch gering ausfallen kann. Mit 30 Neuronen pro Schicht wird der vorgegebene Bereich maximal ausgenutzt, eine Ausweitung führt jedoch zu keiner zusätzlichen Steigerung.

Tabelle 6-2: Hyperparameter der fünf besten Konfigurationen des Imitationsmodells

| Nummer | Neuronen pro Schicht | Anzahl der Schichten | Batch-Größe | Dropout-Intensität | Lernrate | Sequenzlänge | RMSE |
|---|---|---|---|---|---|---|---|
| 8 | 30 | 2 | 5050 | 0.161 | 0.001936 | 71 | 0.0589 |
| 41 | 30 | 3 | 1000 | 0.205 | 0.004901 | 71 | 0.0579 |
| 42 | 30 | 3 | 1000 | 0.214 | 0.005369 | 71 | 0.0573 |
| 43 | 30 | 3 | 1000 | 0.205 | 0.006404 | 61 | 0.0580 |
| 46 | 30 | 4 | 1000 | 0.205 | 0.007697 | 41 | 0.0586 |

Für die abschließende Festlegung der Modellparameter (Gewichtungen), wird Konfiguration 42 über insgesamt 2000 Trainingsepochen optimiert, was in Abbildung 6-25 dargestellt ist.

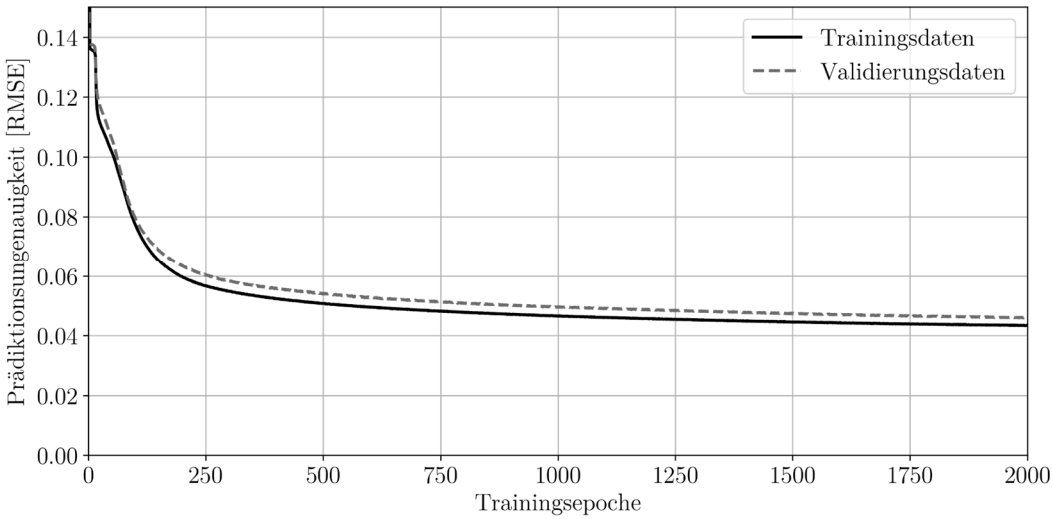


Abbildung 6-25: Prädiktionsungenauigkeit des Imitationsmodells gegenüber den Trainings- und Validierungsdaten über 2000 Epochen

Von Beginn an ist der Unterschied in den RMSE-Werten bezüglich der Trainings- und der Validierungsdaten gering. Dies deutet darauf hin, dass das Modell bereits nach wenigen Optimierungen gut generalisiert. Auch im weiteren Verlauf kommt es zu keiner Überanpassung hinsichtlich der Trainingsdaten, die Prädiktionsungenauigkeit der Validierungsdaten sinkt in vergleichbarem Maß. Ähnlich wie bei dem datenbasierten Modell aus Kapitel 6.2 ist eine Konvergenz ab Trainingsepoche 1750 erkennbar. Dies kann daran liegen, dass sowohl der Trainingsumfang als auch die Modellgröße vergleichbar sind und eine identische Lernrate benutzt wird. Die minimale Abweichung zu den Validierungsdaten wird bei Trainingsepoche 1995 mit einem RMSE von 0.0459 erreicht. Dies ist circa 20 % besser als der Wert, der bei der Hyperparameteroptimierung erreicht wurde. Die Abweichung zu den Trainingsdaten beträgt bei der gleichen Epoche 0.04331 und ist damit nur 5 % besser im Vergleich mit den Validierungsdaten. Dies unterstreicht die Leistungsfähigkeit des Modells bei der Verarbeitung neuer Daten.

Das trainierte GT-Suite Imitationsmodell ermöglicht es, neben den Emissionswerten auch die bereits genannten Verbrennungsparameter mit hoher Genauigkeit zu prädizieren und erfasst damit indirekt einige der physikalisch-phänomenologischen Zusammenhänge, auf welcher das ursprüngliche GT-Suite Modell basiert. In der neu geschaffenen, auf Neuronalen Netzwerken basierenden Form kann es jedoch auch mit einer deutlich reduzierten Anzahl an Eingangsparametern (drei gegenüber ursprünglich zwölf) und zusätzlich effizienter in Echtzeit auf leistungsarmer Hardware eingesetzt werden. Aufgrund dieser Vorteile bildet das Imitationsmodell eine wichtige Grundlage für die nachfolgend beschriebenen hybriden Modelle.

### 6.3.2 Parallele Architektur

Die untersuchte parallele Architektur basiert darauf, die Vorhersagen beider vorgestellter Modelle zu nutzen und sie anschließend zu einer neuen Prädiktion zu kombinieren. Die Art der Kombination kann dabei grundsätzlich von einer einfachen Mittelwertbildung über feste Gewichtungen bis hin zu einer flexiblen Verarbeitung der individuellen Modellausgänge reichen. In der vorliegenden Arbeit wird die letztgenannte Möglichkeit untersucht. Die Annahme ist, dass die Modelle in Abhängigkeit des Lastprofils (Eingangsprofil) individuelle Stärken aufweisen. Das physikalisch-phänomenologische Modell (und das davon abgeleitete Imitationsmodell) ist auf der Grundlage stationärer Messungen aufgebaut, die Stützstellen im Motorkennfeld darstellen. In diesen Betriebspunkten und bei geringer Dynamik könnte das physikalisch-phänomenologische Modell die realen Emissionen besser darstellen als das auf transienten RDE-Profilen trainierte datenbasierte Modell. Der (quasi-)stationäre Betrieb hat auch im realen Straßenverkehr einen Anteil – beispielsweise bei gleichbleibender Geschwindigkeit auf ebener Strecke – weshalb die Verknüpfung zweier spezialisierter Modelle Vorteile bringen könnte.

Die angesprochene Abhängigkeit der individuellen Prädiktionsungenauigkeit vom Lastprofil und damit das Entscheidungskriterium, wie die Modellausgänge kombiniert bzw. verarbeitet werden sollen, ist jedoch vorab nicht hinreichend bekannt. Aus diesem Grund wird eine hybride Architektur untersucht, welche neben den jeweiligen Emissionsprädiktionen auch das Lastprofil berücksichtigt. Dadurch soll in Abhängigkeit der Betriebsweise erlernt werden, wie die Prädiktionen des datenbasierten und physikalisch-phänomenologischen Emissionsmodells zu verarbeiten sind. Dieser Aufbau ist schematisch in Abbildung 6-26 dargestellt.

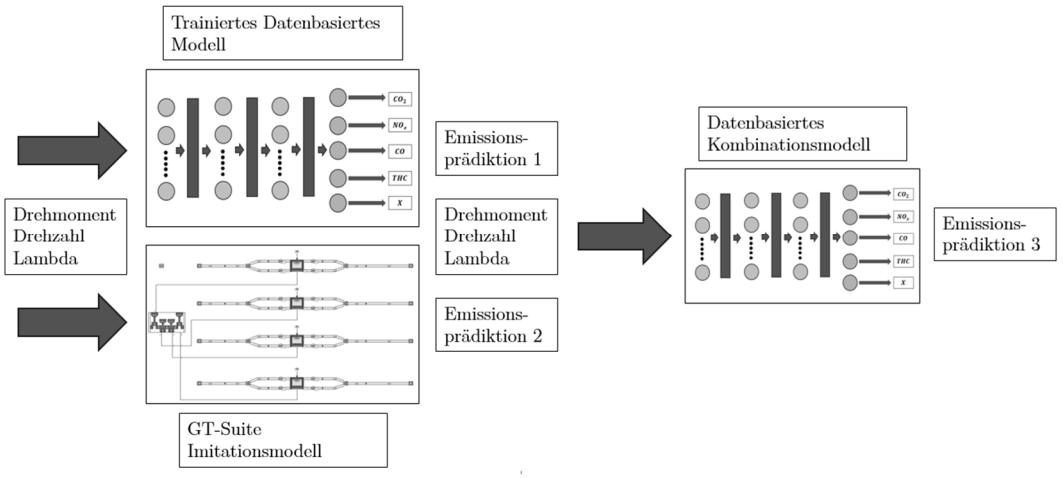


Abbildung 6-26: Schematischer Aufbau des parallelen Hybridmodells

Die Abbildung zeigt den Informationsfluss durch das Modell von den Eingängen (links) zur finalen Emissionsprädiktion (rechts). Das datenbasierte Modell und das physikalisch-phänomenologische Modell, repräsentiert durch das Imitationsmodell, sind entsprechend den vorhergehenden Kapiteln aufgebaut. Beide verarbeiten jeweils die Parameter Drehmoment, Drehzahl und Lambda als Eingangswerte und berechnen daraus individuelle Emissionsprädiktionen (Emissionsprädiktion 1 und Emissionsprädiktion 2). Diese Informationen werden gemeinsam mit den Eingangswerten einem weiteren datenbasierten Modell (Kombinationsmodell) zur Verfügung gestellt. Dieses besitzt elf Eingänge (acht Emissionsprädiktionen, Drehmoment, Drehzahl und Lambda) und errechnet eine finale Emissionsprädiktion (Emissionsprädiktion 3). Das Kombinationsmodell wird anhand von Trainingsdaten optimiert, um die Verarbeitung der Eingänge und damit auch die Gewichtungen der parallel angeordneten vorherigen Modelle in Abhängigkeit der Motorbetriebsweise zu erlernen. Da das Kombinationsmodell als zweite Stufe dieses Hybridaufbaus fungiert, kann der Aufbau schematisch auch als parallel-serielle Anordnung beschrieben werden.

Das Kombinationsmodell soll erneut in der Lage sein, Lastprofile zu erkennen, um die Dynamik der Betriebsweise analysieren zu können. Aus diesem Grund ist die Architektur des Neuronalen Netzes bis auf die Anzahl der Eingangswerte vergleichbar mit dem datenbasierten Modell aus Kapitel 6.2.4. Die ersten $n$ Schichten bestehen aus LSTM-Neuronen, gefolgt von einer Feed-Forward Schicht. Für die Bestimmung der Hyperparameter wird erneut der Tree-structured Parzen Estimator zur automatisierten Optimierung eingesetzt, der Optimierungsvorgang ist in Abbildung 6-27 dargestellt.

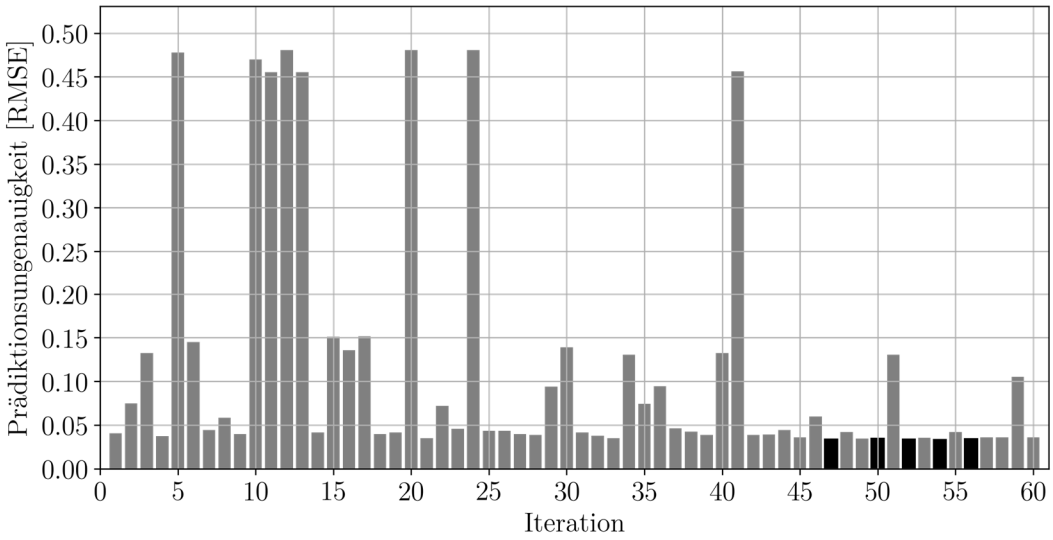


Abbildung 6-27: Prädiktionsungenauigkeit des parallelen Hybridmodells bezüglich des Validierungsdatensatzes bei der Hyperparameteroptimierung mit TPE. Die schwarz eingefärbten Balken sind die fünf besten Hyperparameterkonfigurationen

In der obigen Abbildung sind 60 verschiedene Hyperparameterkonfigurationen und die dazugehörige Prädiktionsungenauigkeit bezüglich des Validierungsdatensatzes über 500 Trainingsepochen dargestellt. Die besten Konfigurationen sind als schwarze Balken dargestellt und werden ab Iteration 47 erreicht. Die geringste Prädiktionsungenauigkeit (geringster RMSE-Wert) liegt bei Iteration 54 mit einem RMSE-Wert von 0.0332 vor. Dieser ist circa 23 % geringer als der Bestwert bei der Hyperparameteroptimierung des ausschließlich datenbasierten Modells aus Kapitel 6.2.4. Die Hyperparameter der fünf besten Konfigurationen sind in Tabelle 6-3 dargestellt.

Tabelle 6-3: Hyperparameter der fünf besten Konfigurationen des parallelen Hybridmodells

| Nummer | Neuronen pro Schicht | Anzahl der Schichten | Batch-Größe | Dropout-Intensität | Lernrate | Sequenzlänge | RMSE |
|---|---|---|---|---|---|---|---|
| 47 | 30 | 2 | 50 | 0.111 | 0.001588 | 91 | 0.0336 |
| 50 | 30 | 3 | 50 | 0.110 | 0.001565 | 91 | 0.0342 |
| 52 | 30 | 3 | 50 | 0.101 | 0.001536 | 101 | 0.0336 |
| 54 | 30 | 3 | 50 | 0.100 | 0.000915 | 81 | 0.0332 |
| 56 | 30 | 3 | 50 | 0.100 | 0.002760 | 101 | 0.0340 |

Aus der Tabelle lässt sich eine klare Tendenz für die Architektur des Kombinationsmodells (als zweite Stufe des parallelen Hybridmodells) erkennen. Die besten Ergebnisse werden ausschließlich bei breiten Netzen (30 Neuronen pro Schicht) erreicht. Dies kann daraus resultieren, dass bedingt durch die elf Eingangsparameter viele Informationen pro Zeitschritt parallel verarbeitet werden müssen. Die Netztiefe ist jedoch gering und liegt zwischen zwei und drei LSTM-Schichten. Tiefere Netze bieten dahingegen keinen weiteren Vorteil. An dieser Stelle muss jedoch beachtet werden, dass das Kombinationsmodell durch die Emissionsprädiktionen des datenbasierten (acht Schichten) und des Imitationsmodells (drei Schichten) bereits aufbereitete Daten erhält und dadurch eventuell durch einen flachen Netzaufbau von einer besseren Generalisierungsfähigkeit profitieren kann. Die beste Konfiguration wird vor der abschließenden Modellanalyse in Kapitel 7.3 über mindestens 2000 Epochen trainiert, was in Abbildung 6-28 dargestellt ist.

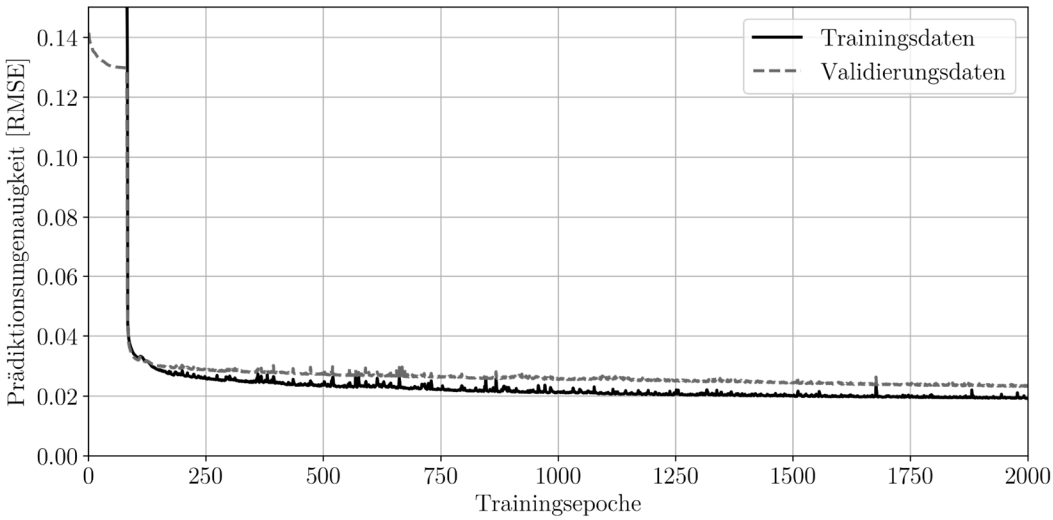


Abbildung 6-28: Prädiktionsungenauigkeit des parallelen Hybridmodells gegenüber den Trainings- und Validierungsdaten über 2000 Epochen

Zu Beginn des Trainings ist ein großer Unterschied in der Prädiktionsungenauigkeit zwischen den Trainings- und Validierungsdaten erkennbar. Dies ist nach circa 100 Epochen mit einer erheblichen Verbesserung der Modellgewichtungen behoben. In dem darauffolgenden Verlauf sind die Unterschiede gering und das Modell kann die vorliegenden Daten generalisiert darstellen. Bis zu den letzten circa 10 % des Trainings ist ein leichtes Gefälle in der schwarzen und der grauen Kurve erkennbar, danach wird ein Plateau erreicht. Der geringste RMSE-Wert bezogen auf die Validierungsdaten beträgt 0.02315 und wird bei Epoche 1897 erreicht. Die dazugehörige Prädiktionsungenauigkeit der Trainingsdaten liegt bei 0.01918. Bei der Abbildung der Validierungsdaten zeigt das parallele Hybridmodell über den gleichen Trainingsumfang eine 17 % bessere Leistung als das datenbasierte Modell aus Kapitel 6.2.

### 6.3.3 Serielle Architektur

Für die serielle Architektur wird im Gegensatz zum parallelen Aufbau nur das physikalisch-phänomenologische Imitationsmodell als Basis verwendet, siehe Abbildung 6-29.

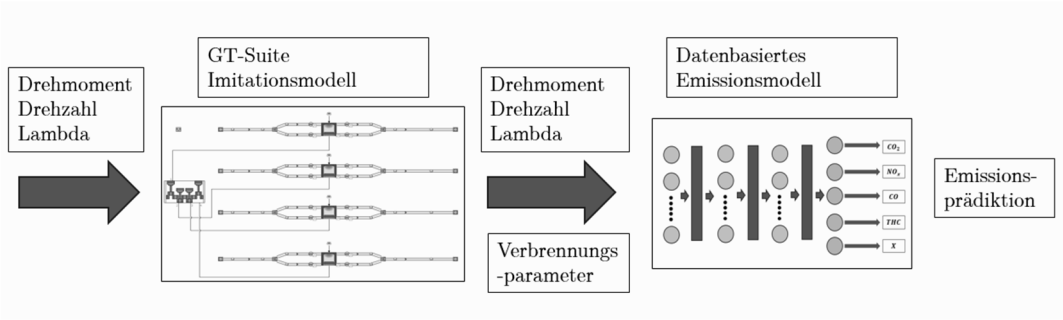


Abbildung 6-29: Schematischer Aufbau des seriellen Hybridmodells

Weiterhin wird auf eine Emissionsprädiktion im ersten Modellschritt verzichtet, obwohl das Imitationsmodell mit den Eingangswerten Drehmoment, Drehzahl und Lambda über die notwendigen Informationen verfügt. Stattdessen werden die Parameter „Verbrennungsspitzentemperatur“, „maximaler Druckgradient“ und „Restgasgehalt“ berechnet. Zusammen mit dem Drehmoment, der Drehzahl und Lambda bilden sie die sechs Eingangsparameter für ein nachgeschaltetes datenbasiertes Emissionsmodell. Die Idee dieses Ansatzes ist es, eine höher verdichtete und zudem breitere Datengrundlage zu bilden, auf welcher die finale Emissionsprädiktion erfolgen kann. Im Gegensatz zum parallelen Ansatz stehen durch die Verbrennungsparameter neue Entscheidungskriterien zur Verfügung. Aufgrund der Anwendung der ursprünglich stationär optimierten Verbrennungsmodelle in einer dynamischen Umgebung und dem zusätzlich vorhandenen Informationsverlust bei der Umwandlung in ein datenbasiertes Modell (Imitationsmodell), ist bezüglich der prädizierten Verbrennungsparameter jedoch mit einer Unschärfe zu rechnen.

Die unter Umständen erheblichen Abweichungen der Verbrennungsparameter von den realen Werten werden bei diesem Ansatz im Vorfeld nicht als Ausschlusskriterium beurteilt. Begründet wird dies unter anderem dadurch, dass eine präzise Abbildung der physikalischen Größen keine Voraussetzung für die Prädiktionsgenauigkeit des nachfolgenden datenbasierten Modells darstellt. Das Neuronale Netz hat keine Informationen bezüglich der Herkunft der Eingangsparameter, deren ursprünglicher Relevanz oder eventueller physikalischer Querbeeinflussungen. Stattdessen wird die Verarbeitung von Grund auf ohne jegliche Vorbelastung erlernt, um die Ausgangswerte bestmöglich zu prädizieren. Daher ist die Motivation für die serielle Architektur, dass die prädizierten Verbrennungsparameter trotz einer realen Ungenauigkeit wichtige Muster und Informationen (etwa durch eine tendenziell richtige Darstellung der Parameterverläufe) bereitstellen.

In Abbildung 6-30 ist der Verlauf der Hyperparameteroptimierung dargestellt.

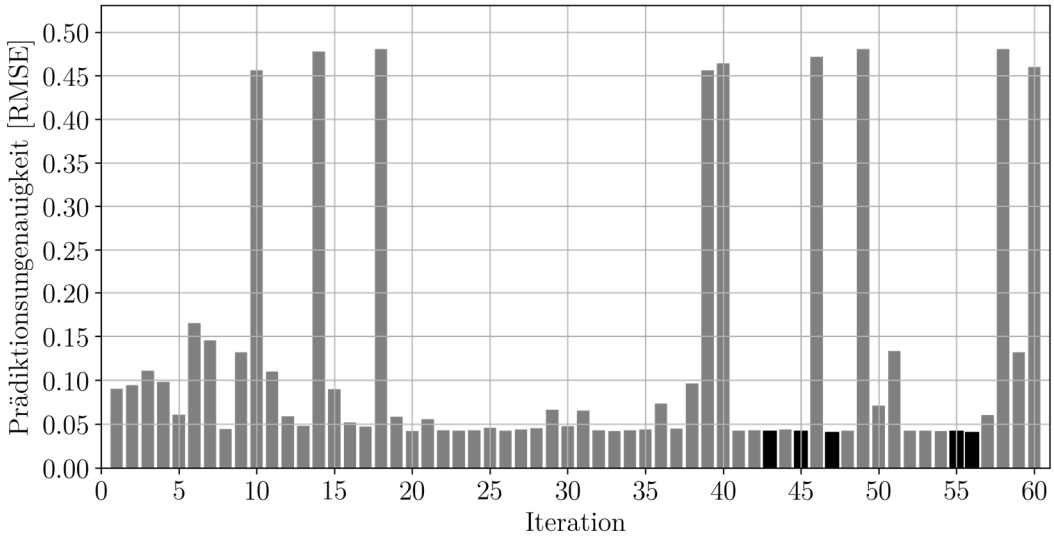


Abbildung 6-30: Prädiktionsungenauigkeit des seriellen Hybridmodells bezüglich des Validierungsdatensatzes bei der Hyperparameteroptimierung mit TPE. Die schwarz eingefärbten Balken sind die fünf besten Hyperparameterkonfigurationen

Die besten Prädiktionsergebnisse erreichen die Konfigurationen zwischen Iteration 43 und 56, wobei Konfiguration 56 mit einem RMSE-Wert von 0.0400 die höchste Genauigkeit aufweist. Dies ist circa 7 % besser als das rein datenbasierte Modell, jedoch 17 % schlechter als das Ergebnis der parallelen Architektur. Bei dieser Bewertung müssen jedoch die jeweilig frühe Phase der Modellbildung und die verringerte Anzahl der Trainingsepochen berücksichtigt werden. Erst wenn alle Modelle abschließend trainiert sind, kann ein Vergleich mit den Testdatensatz eine finale Einschätzung ermöglichen. Die Hyperparameter der fünf besten Konfigurationen sind in Tabelle 6-4 abgebildet.

Tabelle 6-4: Hyperparameter der fünf besten Konfigurationen des seriellen Hybridmodells

| Nummer | Neuronen pro Schicht | Anzahl der Schichten | Batch-Größe | Dropout-Intensität | Lernrate | Sequenzlänge | RMSE |
|---|---|---|---|---|---|---|---|
| 43 | 30 | 3 | 50 | 0.254 | 0.000956 | 41 | 0.0413 |
| 45 | 30 | 3 | 50 | 0.233 | 0.000281 | 31 | 0.0416 |
| 47 | 30 | 2 | 50 | 0.234 | 0.000253 | 21 | 0.0401 |
| 55 | 30 | 3 | 50 | 0.228 | 0.000545 | 21 | 0.0415 |
| 56 | 30 | 2 | 50 | 0.228 | 0.000552 | 21 | 0.0400 |

Die Hyperparameterkombinationen mit einer niedrigen Prädiktionsungenauigkeit sind bei dem neuen Modell der seriellen Architektur vergleichbar mit denen des Kombinationsmodells aus Kapitel 6.2.4. Es werden Netzaufbauten bevorzugt, welche breit (30 Neuronen pro Schicht), jedoch flach (zwei bis drei Schichten) sind. Auch hier könnte dies an den verarbeiteten, erweiterten Informationen liegen, da das Modell sechs Eingänge besitzt. Die Konfiguration mit der Nummer 56 wird über einen erweiterten Umfang von mindestens 2000 Epochen trainiert. Der Verlauf des Trainings ist in nachfolgender Abbildung dargestellt.

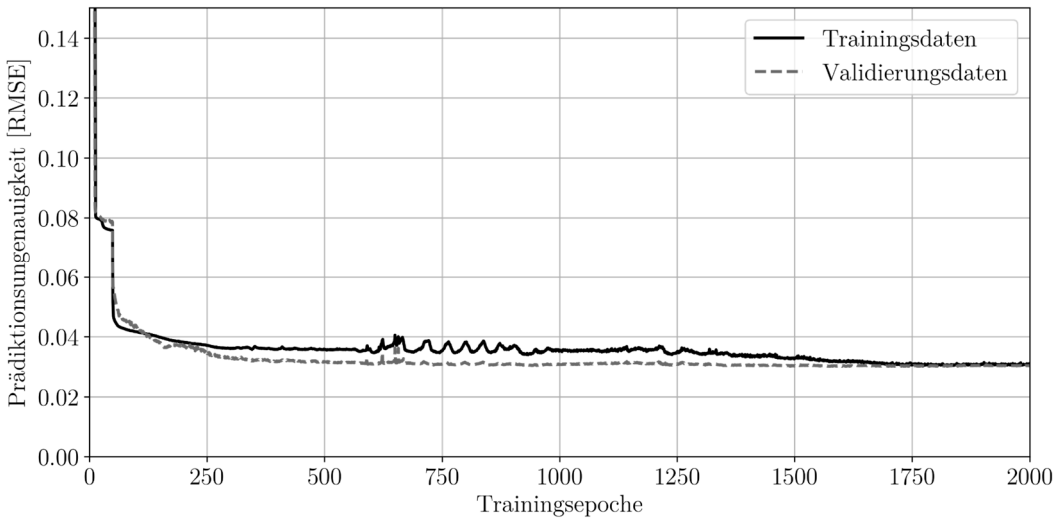


Abbildung 6-31: Prädiktionsungenauigkeit des seriellen Hybridmodells gegenüber den Trainings- und Validierungsdaten über 2000 Epochen

Die dargestellten Trainingsverläufe sind von Beginn an auf einem ähnlichen Niveau. Verglichen mit den Trainingsverläufen der bisher gezeigten Modelle sind in der obigen Abbildung einige Störungen erkennbar, etwa zwischen Epoche 600 und 1000. Die auftretende Oszillation, welche sich besonders in der Prädiktionsungenauigkeit der Trainingsdaten zeigt, lässt auf eine zu hohe Lernrate in diesem Bereich schließen, wodurch die Anpassung der Gewichtungen auch zu schlechteren Ergebnissen führen können. Dies wird im weiteren Trainingsverlauf durch den Adam-Algorithmus angeglichen, beispielsweise durch eine Rückkehr zur niedrig gewählten Basis-Lernrate. Ab Epoche 1500 sind die Verläufe homogen und nach einem kurzen Gefälle ist eine Konvergenz erkennbar. Bei Trainingsepoche 1575 wird die niedrigste Prädiktionsungenauigkeit der Validierungsdaten mit einem RMSE-Wert von 0.0301 erreicht, was circa 30 % schlechter als das parallele Hybridmodell und 8 % schlechter als das datenbasierte Modell ist.

# 7 Analyse der Modellqualität

In diesem Kapitel wird die Qualität der in den vorherigen Abschnitten entwickelten Modelle abschließend analysiert und verglichen. Um eine unvoreingenommene Bewertung zu gewährleisten, wird dabei ein Testdatensatz verwendet, welcher weder Teil des Trainings noch der Validierung war.

Der Testdatensatz repräsentiert eine auf der Straße gemessene RDE-Fahrt mit einer Dauer von etwa 2000 Sekunden. Diese Fahrt umfasst verschiedene Fahrsituationen (innerorts, Landstraße, Autobahn), wodurch ein alltagsnahes Profil abgebildet wird. Die Drehzahl , das Drehmoment und die gemessenen Stickstoffoxide sind in Abbildung 7-1 dargestellt.

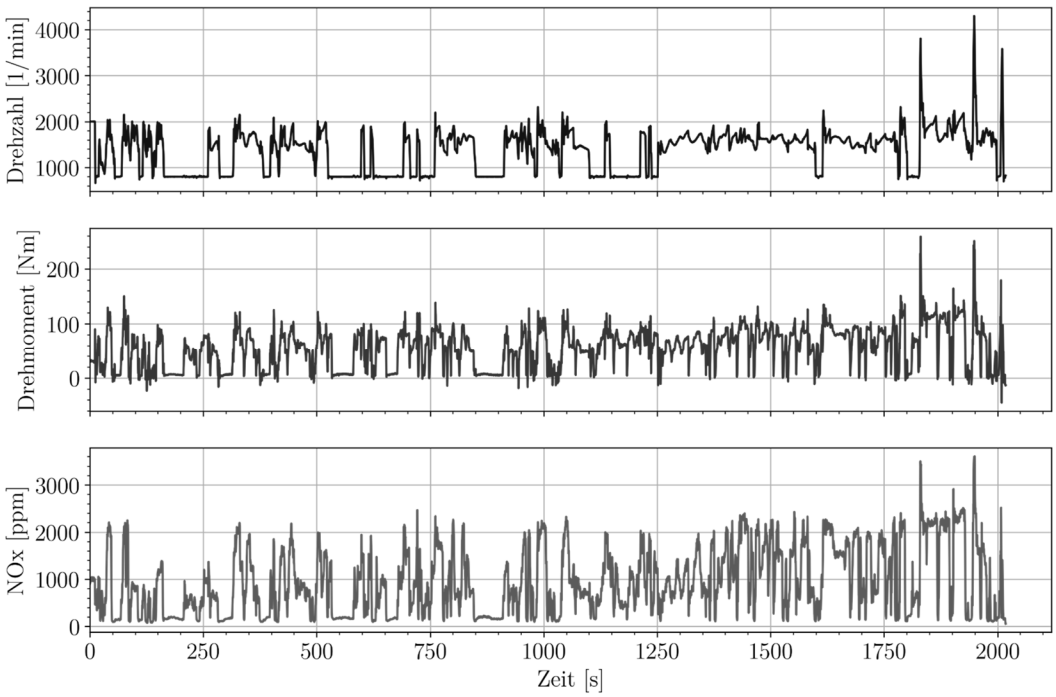


Abbildung 7-1: Darstellung des Testprofils für die Bewertung der Modelle

Anhand der Verläufe des Drehzahl- und Drehmomentprofils, aber auch dem Stickstoffoxidverlauf als exemplarische Emissionsspezies, kann die Dynamik des Testdatensatzes beurteilt werden. In Zusammenhang mit der Dauer von über 30 Minuten wird eine hohe Variationsbreite an Betriebssituationen gewährleistet, was wiederum eine umfangreiche Bewertung der Modelle erlaubt.

Analyse der Modellqualität

In der Analyse werden sowohl das physikalisch-phänomenologische Imitationsmodell als auch das datenbasierte Modell sowie die beiden hybriden Modellierungsansätze untersucht. Ziel ist es, die Modelle hinsichtlich ihrer Genauigkeit und Robustheit zu bewerten und ihre Eignung für praktische Anwendungen zu ermitteln. In den nachfolgenden Unterkapiteln werden die Emissionsprädiktionen der genannten Modelle jeweils über das dargestellte Testprofil mit den Messwerten des Prüfstands verglichen. Dabei werden alle vier Emissionsspezies betrachtet. Neben den Verläufen als qualitatives Beurteilungskriterium werden die einzelnen RMSE-Werte über den circa 2000 Sekunden langen Testlauf berechnet, um eine quantitative Einschätzung der Prädiktionsungenauigkeit und einen direkten Vergleich zu ermöglichen.

Am Ende des Kapitels werden die beiden besten Modellarchitekturen direkt verglichen und die jeweiligen Vor- und Nachteile in Bezug auf die Fahrsituationen diskutiert. Daraus können Verbesserungspotenziale für weiterführende Arbeiten abgeleitet werden.

## 7.1 Physikalisch-phänomenologisches Imitationsmodell

Das physikalisch-phänomenologische Imitationsmodell zeigt die schlechteste Prädiktionsgenauigkeit über alle vier Emissionsspezies im Vergleich zu den anderen drei Modellen. Die Wurzel des mittleren quadratischen Fehlers beträgt im Schnitt 0.1455 und das Modell kann die dynamischen Verläufe nicht präzise abbilden, wie aus Abbildung 7-2 deutlich wird.

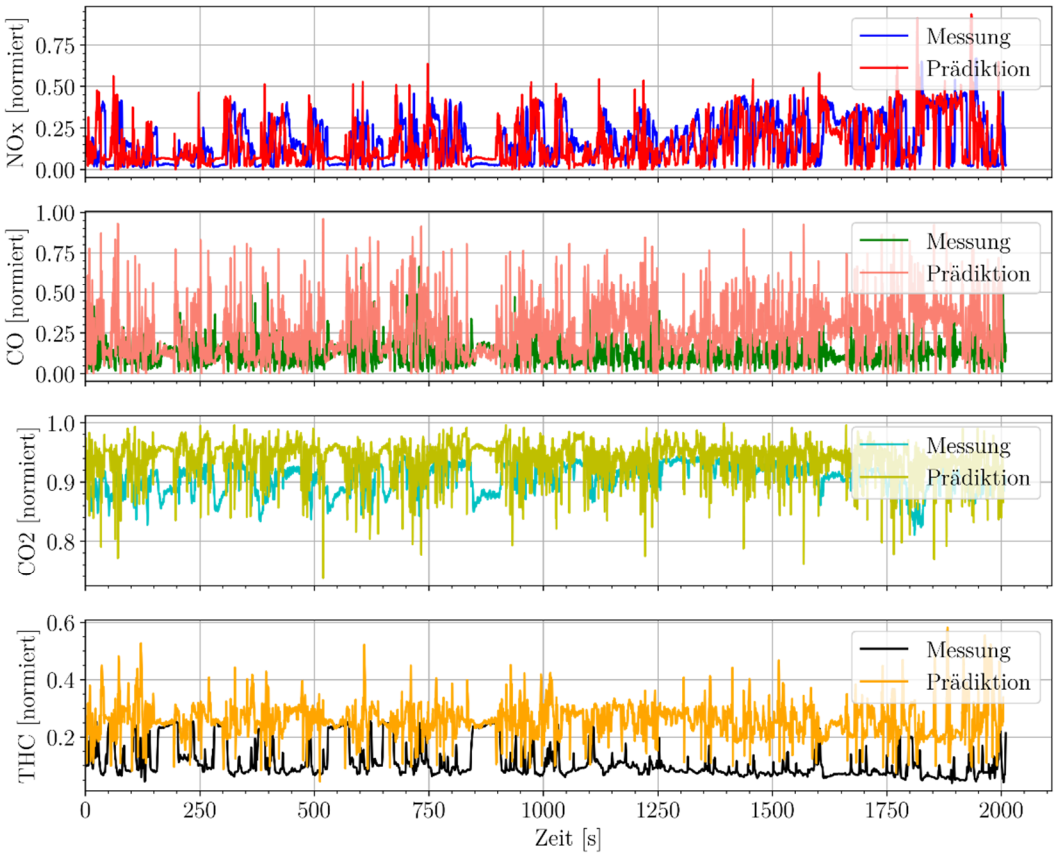


Abbildung 7-2: Vergleich zwischen Messung und Prädiktionswerten des physikalisch-phänomenologischen Imitationsmodells über den Testzyklus

Die mangelhafte Prädiktionsgenauigkeit des physikalisch-phänomenologischen Imitationsmodells lässt sich durch den Modellaufbau erklären. Als Grundlage dient das GT-Suite Modell, dessen Verbrennungs- und Emissionsmodelle auf stationäre Betriebspunkte optimiert sind. Da das Imitationsmodell das GT-Suite Modell präzise abbilden soll, ist es auf dessen Verhalten in stationären und dynamischen Betriebsarten trainiert. Jedoch bildet das GT-Suite Modell die transiente Emissionsentstehung nicht optimal ab, was daher auch für das Imitationsmodell gilt. Werden die Emissionsspezies untereinander verglichen, gibt es jedoch Unterschiede in der Prädiktionsgenauigkeit. Die Stickstoffoxide können qualitativ am besten abgebildet werden.

Dies kann unter anderem durch die in 6.2.3 nachgewiesene, hohe (0.68) lineare Korrelation zwischen dem Drehmoment und den Stickstoffoxiden erklärt werden. Dadurch lässt sich der Verlauf der Stickstoffoxide im Vergleich zu den anderen Emissionsspezies einfach prädizieren.

Kohlenstoffmonoxid, Kohlenstoffdioxid und unverbrannte Kohlenwasserstoffe werden dahingegen im Schnitt zu hoch errechnet. Weiterhin weisen alle drei Verläufe eine höhere Variabilität auf, was sich in den Messungen nicht widerspiegelt. Es gibt jedoch auch vereinzelte Phasen, in welchen das Modell eine höhere Genauigkeit aufweist, wie beispielsweise in den Leerlaufphasen (Sekunde 150-200 und 850-900). In diesen Phasen stimmen die Prädiktionswerte von Kohlenstoffmonoxid, Stickstoffoxiden und unverbrannten Kohlenwasserstoffen gut mit den Messwerten überein.

Als Konsequenz lässt sich ableiten, dass sich das physikalisch-phänomenologische Imitationsmodell zwar nur sehr bedingt für die Emissionsprädiktion eignet, jedoch in bestimmten Betriebssituationen hilfreiche Informationen liefern kann. Diese müssen gezielt bestimmt werden können, um eine Grundlage für die Erstellung leistungsfähiger Hybridmodelle zu bilden.

## 7.2 Datenbasiertes Modell

Das datenbasierte Modelle kann die gemessenen Emissionsverläufe über den Testzyklus (siehe Abbildung 7-3) sehr gut darstellen, der durchschnittliche RMSE-Wert beträgt 0.0401, aufgeteilt auf die einzelnen Komponenten ergibt sich folgende Genauigkeit (RMSE):

- Stickstoffoxide: 0.0496
- Kohlenstoffmonoxid: 0.0586
- Kohlenstoffdioxid: 0.0154
- Unverbrannte Kohlenwasserstoffe: 0.0172

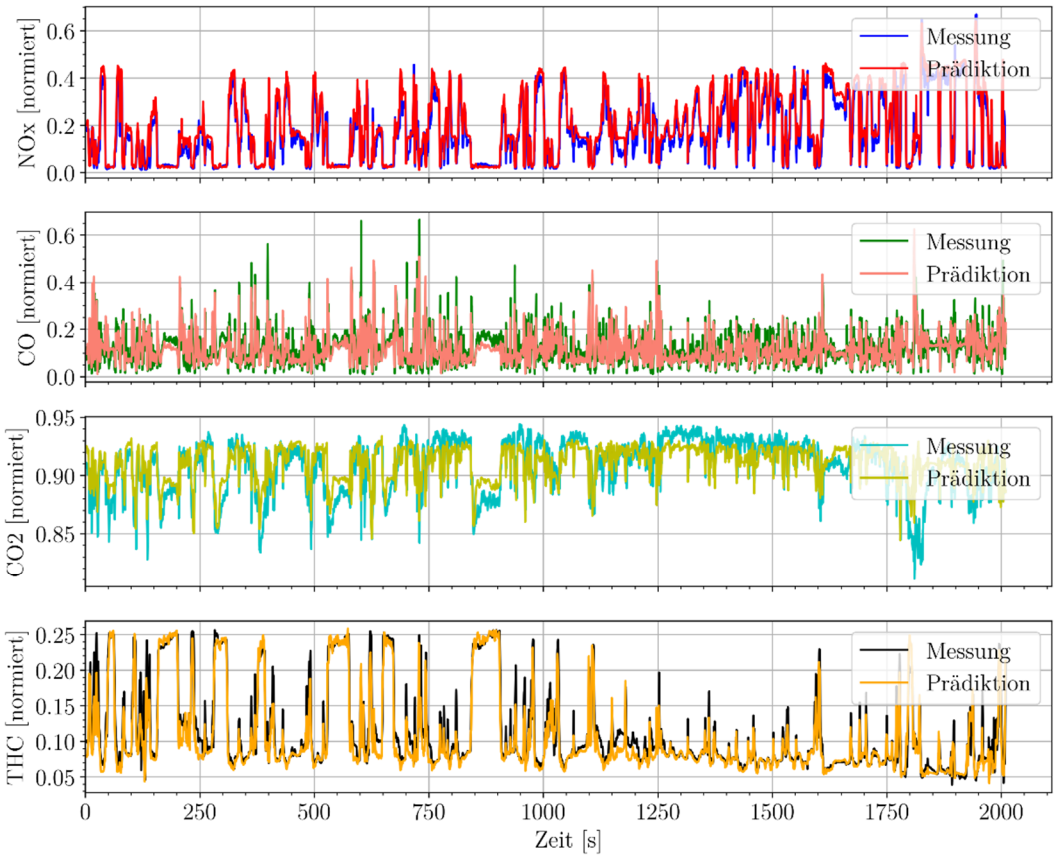


Abbildung 7-3: Vergleich zwischen Messung und Prädiktionswerten des datenbasierten Modells über den Testzyklus

Das Testergebnis bestätigt einerseits die Leistungsfähigkeit datenbasierter Architekturen aus den Vorarbeiten (siehe [60]), andererseits aber auch die guten Ergebnisse während dem Training und der Validierung des Modells in Kap. 6.2.4. Der RMSE-Wert bezüglich des Testdatensatzes liegt wie zu erwarten über dem Validierungsergebnis (0.0279), dennoch ist festzuhalten, dass das Modell eine hohe Generalisierungseigenschaft aufweist.

Allgemein betrachtet werden bei allen Emissionsspezies die Niveaus treffend prädiziert. Größere Abweichungen sind vor allem in dynamischen Änderungen erkennbar, was sich beispielsweise im obersten und untersten Diagramm der obigen Abbildung erkennen lässt. Bei den Stickstoffoxiden werden hauptsächlich Verläufe mit negativer Steigung schlechter prädiziert; das Modell senkt das Emissionslevel zu langsam. Bei nachfolgenden stationären Phasen stimmt es wieder mit der Messung überein. Die unverbrannten Kohlenwasserstoffe werden dahingegen besonders bei Spitzen vom Modell eher unterschätzt.

Die Prädiktionsgenauigkeit von Kohlenstoffmonoxid und Kohlenstoffdioxid ist schwer auf einzelne Effekte oder Betriebsszenarien zurückzuführen. Beide simulierten Verläufe unter- und überschätzen die realen Messwerte in vergleichbarer Häufigkeit. Die Prädiktion von Kohlenstoffdioxid könnte anhand der oberen Abbildung qualitativ schlechter eingeschätzt werden, dabei muss jedoch der beschränkte Wertebereich (nach der Normierung) dieser Emissionsspezies im regulären Betrieb berücksichtigt werden. Größere Abweichungen treten in erster Linie bei nicht stöchiometrischem Betrieb auf. Dies kann beispielsweise bei Volllast und hohen Drehzahlen der Fall sein, ist jedoch nicht Bestandteil des Testzyklus.

Das datenbasierte Modell kann die Emissionsverläufe auch bei hochdynamischen, völlig unbekannten Betriebsszenarien sehr genau prädizieren und könnte sich dadurch auch beispielsweise für die Auslegung der Betriebsstrategie von Hybridfahrzeugen oder Regelungsfunktionen im Steuergerät (Abgasnachbehandlung) eignen. Somit erfüllt das Modell bereits ein zentrales Ziel der vorliegenden Arbeit (Kapitel 4).

### 7.3 Paralleles Hybridmodell

Das parallele Hybridmodell zeigt die höchste Prädiktionsgenauigkeit der erstellten Emissionsmodelle; die dazugehörigen Verläufe können Abbildung 7-4 entnommen werden. Der durchschnittliche RMSE-Wert liegt für den Testdatensatz bei 0.0295 und dadurch nochmals unter dem Ergebnis des datenbasierten Modells (0.0401). Damit bestätigen sich die Ergebnisse aus Kapitel 6.3.2, in welchem das parallele Hybridmodell auch bei den Trainings- und Validierungsdaten jeweils die höchste Genauigkeit aufweisen konnte.

Die RMSE-Werte der einzelnen Emissionskomponenten betragen:

- Stickstoffoxide: 0.0366
- Kohlenstoffmonoxid: 0.0410
- Kohlenstoffdioxid: 0.0154
- Unverbrannte Kohlenwasserstoffe: 0.0148

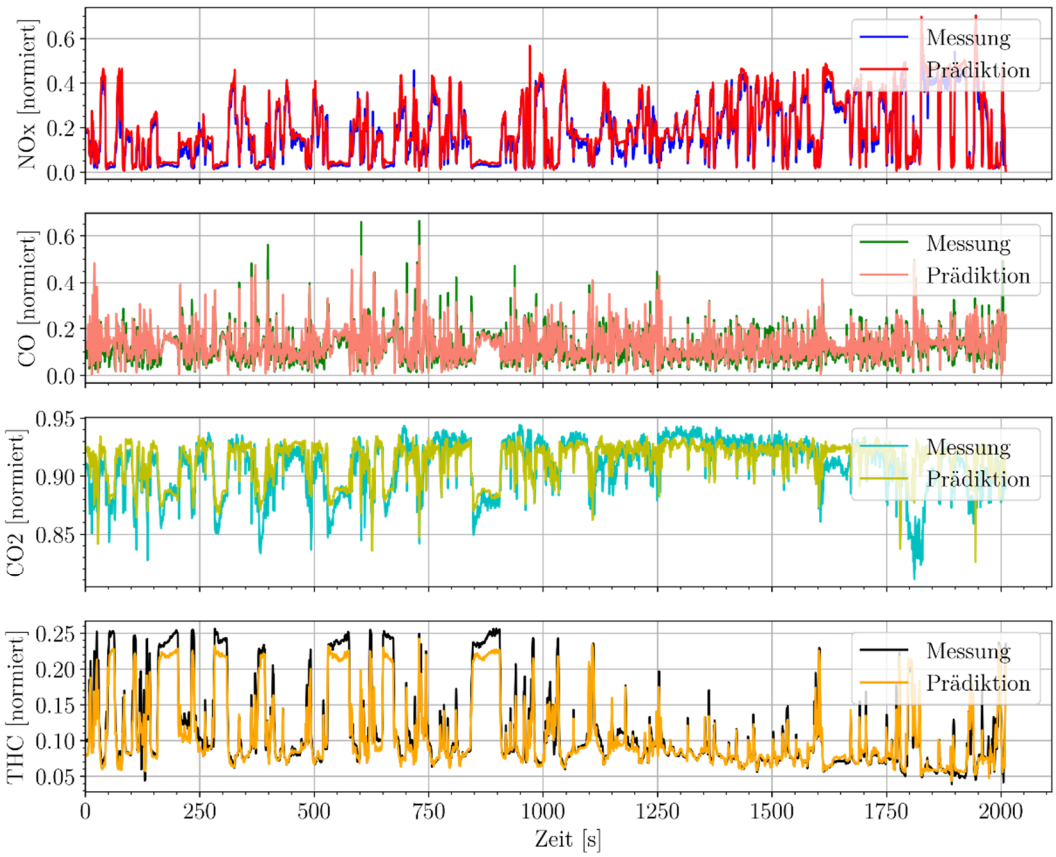


Abbildung 7-4: Vergleich zwischen Messung und Prädiktionswerten des parallelen Hybridmodells über den Testzyklus

Analyse der Modellqualität

Allgemein sind die Verläufe zwischen Prädiktion und Messung in der obigen Abbildung vergleichbar mit den in 7.2 analysierten Ergebnissen. Daher gilt auch für das parallele Hybridmodell, dass die prädizierten Emissionsniveaus die reale Messung sehr präzise widerspiegeln, jedoch bei dynamischen Phasen und Extremwerten Abweichungen vorhanden sind.

Besonders bei der Prädiktion der Stickstoffoxide und der Kohlenstoffmonoxide zeigt das parallele Hybridmodell die größten Verbesserungen, was sich einerseits in den RMSE-Werten (-27 % und -30 % bezogen auf die bisherigen Bestwerte des datenbasierten Modells), andererseits aber auch qualitativ im Verlauf erkennen lässt. In beiden Fällen werden vor allem die dynamischen Anteile besser dargestellt.

Somit lässt sich auch für das parallele Hybridmodell eine sehr hohe Generalisierungsfähigkeit und die Eignung als effektives Werkzeug für Optimierungs- und Regelungsfunktionen feststellen. Weiterhin demonstriert es die Fähigkeit hybrider Modelle, durch die Steigerung der vorhandenen Informationsdichte Vorteile gegenüber rein datenbasierten Modellen zu erreichen. Dies ist besonders bei einer beschränkten Anzahl verfügbarer Trainingsdaten wertvoll und kann dadurch den Entwicklungsaufwand reduzieren.

Der Vergleich zwischen dem parallelen Hybridmodell und dem datenbasierten Modell erfolgt in Kapitel 7.5. An dieser Stelle soll auch diskutiert werden, inwiefern das Hybridmodell von den suboptimalen Prädiktionen des Imitationsmodells profitieren kann.

## 7.4 Serielles Hybridmodell

Das serielle Hybridmodell bestätigt in der Analyse der Testdaten die Ergebnisse aus dem Training und der Validierung. Es liegt mit einem durchschnittlichen RMSE-Wert von 0.0547 hinter dem datenbasierten Modell und dem parallelen Hybridmodell. Auffällig ist dabei die deutliche Verschlechterung von den Validierungs- zu den Testdaten. Wird die Prädiktionsgenauigkeit des seriellen Hybridmodells mit dem datenbasierten Modell verglichen, so liegt die Differenz bei den Validierungsdaten nur bei circa 8 % zu Gunsten des datenbasierten Modells, bei den Testdaten sind es bereits 27 %. Dies spricht für eine geringe Generalisierungsfähigkeit des seriellen Hybridmodells mit dem aktuellen Trainingsstand. Die RMSE-Werte der einzelnen Emissionskomponenten betragen:

- Stickstoffoxide: 0.0712
- Kohlenstoffmonoxid: 0.0774
- Kohlenstoffdioxid: 0.0159
- Unverbrannte Kohlenwasserstoffe: 0.0547

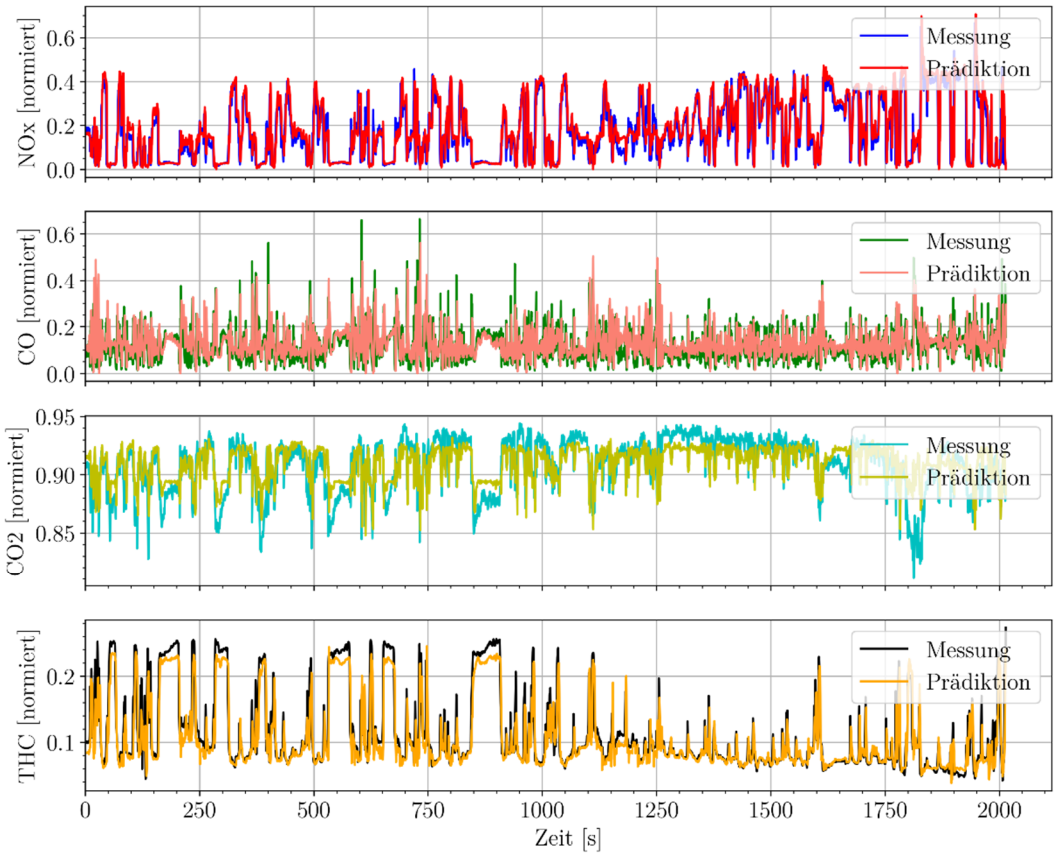


Abbildung 7-5: Vergleich zwischen Messung und Prädiktionswerten des seriellen Hybridmodells über den Testzyklus

Werden die Ergebnisse auf die einzelnen Emissionskomponenten reduziert betrachtet, können die Verläufe im Mittel gut angenähert werden, sind jedoch ausnahmslos schlechter als bei dem datenbasierten Modell und dem parallelen Hybridmodell. Eine mögliche Erklärung liegt in dem Netzaufbau der zweiten Stufe des seriellen Hybridmodells. Dieser ist mit lediglich zwei Schichten sehr flach und kann zudem nicht auf die Informationen des datenbasierten Modells zugreifen wie das parallele Hybridmodell. Dieser Aufbau funktionierte für die Trainings- und Validierungsdaten zwar noch gut – bei der Hyperparameteroptimierung zeigte das serielle Hybridmodell eine höhere Genauigkeit als das datenbasierte Modell –, bei den Testdaten kann diese Leistung jedoch nicht bestätigt werden.

Als mögliche Verbesserungen für das serielle Hybridmodell können zwei Ansätze verfolgt werden. Einerseits könnte die zweite Modellstufe „vergrößert“ werden (mehr Neuronen pro Schicht und/oder mehr Schichten). In Kombination mit weiteren Trainingsdaten könnte das serielle Hybridmodell verlässlicher erlernen, in welchem Maße es auf die prädizierten Verbrennungsparameter vertraut. Da die zweite Modellstufe ebenfalls die Eingangsparameter des datenbasierten Modells (Drehmoment, Drehzahl und Lambda) integriert, sollte das serielle Hybridmodell mit ausreichender Datengrundlage unabhängig von der Genauigkeit des vorgeschalteten Imitationsmodells mindestens auf die Leistung des datenbasierten Modells kommen.

Da dies im theoretischen Extremfall die Grundzüge eines Hybridmodells verfehlt (wenn die Ausgangswerte des Imitationsmodells ignoriert werden), kann andererseits auch an der Prädiktionsgenauigkeit des Imitationsmodells gearbeitet werden. Hierzu muss das grundlegende GT-Suite Modell verbessert werden, was ebenfalls mehr Daten und eine umfangreiche Optimierung mit transienten Anteilen beinhaltet. Dieser Prozess ist zeitaufwändig, wodurch unter gleichen Bedingungen (aktuelle Datenbasis) das serielle Hybridmodell keine Vorteile gegenüber der datenbasierten Emissionsmodellierung aufweisen kann. Das Potenzial, welches in der Informationssteigerung durch physikalische Daten begründet ist, ist zwar vorhanden, muss jedoch mit einer für den Anwendungsfall (dynamischer Betrieb) besseren physikalisch-phänomenologischen Basis demonstriert werden.

## 7.5 Vergleich datenbasiertes und paralleles Hybridmodell

Das datenbasierte und das parallele Hybridmodell weisen bezüglich der Validierungs- und Testdaten die höchsten Genauigkeiten auf und werden in diesem Kapitel im Detail gegenübergestellt. Ziel dieser Untersuchung ist es, die Bereiche zu identifizieren, in denen das Hybridmodell im Vergleich zum datenbasierten Modell Vorteile zeigt, um den Einfluss der zusätzlich bereitgestellten Informationen durch das Imitationsmodell nachvollziehen zu können. Zur Erreichung dieses Ziels werden Differenzverläufe zwischen den realen Messwerten und den vorhergesagten Werten für die vier Ausgangsparameter erstellt. Dabei wird besonders darauf geachtet, ob Auffälligkeiten im Zusammenhang mit den Verläufen der Eingangswerte bestehen. Diese Analyse soll nicht nur die Stärken des Hybridmodells herausarbeiten, sondern auch Verbesserungsmöglichkeiten für Folgearbeiten aufzeigen.

In Abbildung 7-6 ist der Vergleich zwischen dem datenbasierten Modell (blaue Verläufe) und dem parallelen Hybridmodell (rote Verläufe) dargestellt. Im Gegensatz zu den vorrangehenden Abbildungen sind die Farben über alle Emissionsspezies konsistent, da der Fokus dieser Abbildung auf dem Vergleich der Modelle liegt. Die Emissions-Abweichungen und das Lastprofil beziehen sich jeweils auf die normierten Werte. Für die Differenzen wird der prädizierte Wert von dem realen Messwert abgezogen, sodass beispielsweise eine positive Abweichung auf eine zu geringe Modellprädiktion schließen lässt. Das fünfte Diagramm stellt das Lastprofil des Testdatensatzes mit dem Drehmoment in Gelb und der Drehzahl in Schwarz dar.

Die RMSE-Werte bezogen auf die vier Ausgangsparameter sind nachfolgend nochmals zusammengestellt, um die Abbildung besser bewerten zu können:

- Stickstoffoxide → datenbasiertes Modell: 0.0496 > Hybridmodell: 0.0366
- Kohlenstoffmonoxid → datenbasiertes Modell: 0.0586 > Hybridmodell: 0.0410
- Kohlenstoffdioxid → datenbasiertes Modell 0.0154 = Hybridmodell: 0.0154
- Unverbrannte Kohlenwasserstoffe → datenbasiertes Modell: 0.0172 > Hybridmodell: 0.0148

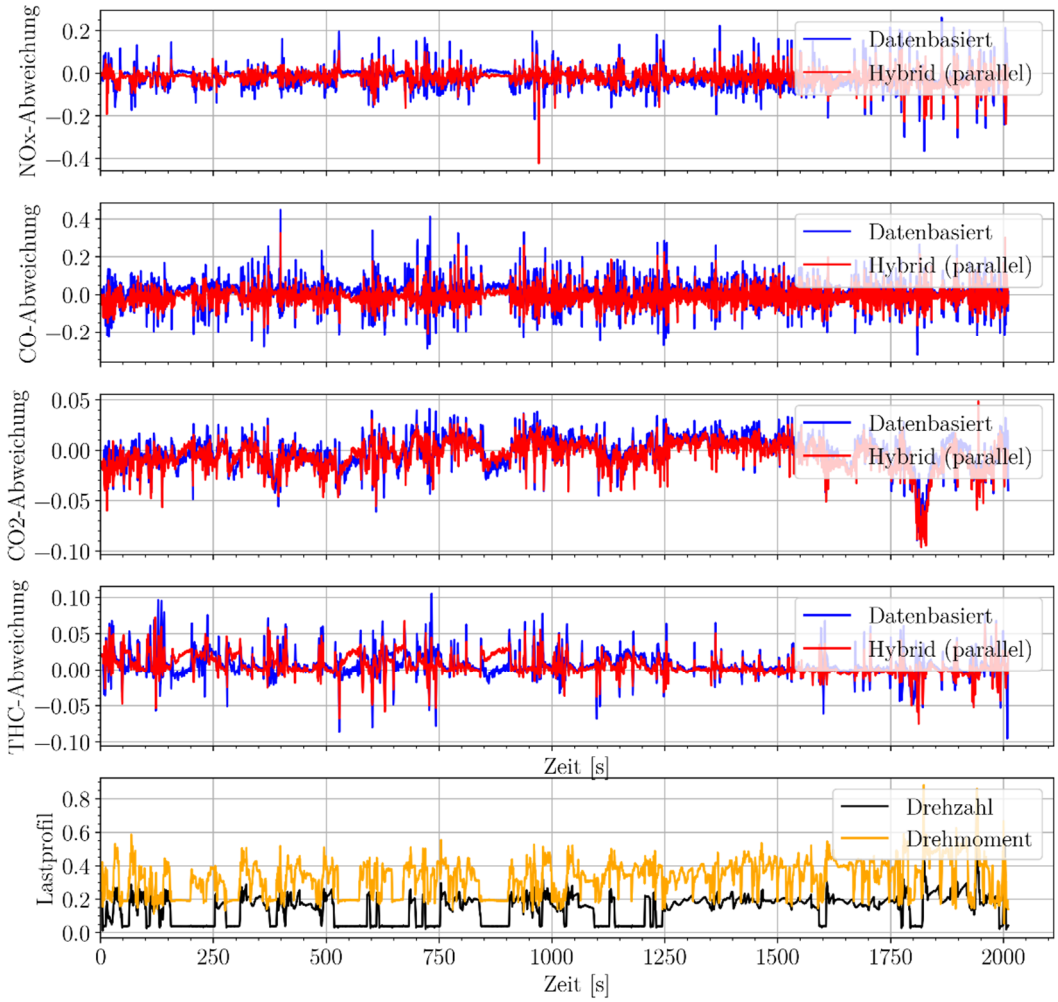


Abbildung 7-6: Vergleich zwischen datenbasiertem Modell und parallelem Hybridmodell

Die Vorteile des parallelen Hybridmodells werden besonders bei den Stickstoffoxiden und Kohlenstoffmonoxid deutlich. Dies zeigt sich in den circa 30 % besseren RMSE-Werten und lässt sich auch in der obigen Abbildung erkennen. In den ersten zwei Diagrammen ist der Verlauf des Hybridmodells enger um die Nulllinie und damit näher an den realen Messwerten. Die maximalen Abweichungen sind geringer als bei dem datenbasierten Modell, welches einige Ausreißer in Form von Spitzen aufweist. Dies ist bei dynamischen Drehmomentänderungen vermehrt der Fall, bei denen das datenbasierte Modell die beiden Emissionsspezies eher über- oder unterschätzt als das Hybridmodell. In den unteren beiden Emissionsdiagrammen ist diese Tendenz nur geringfügig erkennbar und auch in den durchschnittlichen RMSE-Werten liegen beide Modelle auf einem vergleichbaren Niveau. Eine Auffälligkeit besteht im Bereich der unverbrannten Kohlenwasserstoffe. In Phasen der Nulllast (entspricht dem normierten Wert 0.2, da es auch negative Drehmomente gibt, welche folglich näher an der „normierten Null“

liegen), beispielsweise von 150-200 Sekunden, ist die Prädiktionsqualität des Hybridmodells zu gering und verglichen mit dem datenbasierten Modell schlechter. Die Einbeziehung des Imitationsmodells führt zu Nachteilen, obwohl dieses (Vergleich Abbildung 7-2) die unverbrannten Kohlenwasserstoffe in dem betreffenden Zeitbereich sehr gut prädiziert. Das Imitationsmodell berechnet die unverbrannten Kohlenwasserstoffe im Mittel jedoch zu hoch, weshalb es sein kann, dass das Kombinationsmodell diesen Bereich nicht als Ausnahme erkennt, die kombinierte Kohlenwasserstoffprädiktion reduziert und daher auch die finale Prädiktion zu gering ist. Auch an dieser Stelle wäre eine Erweiterung der Trainingsdaten sinnvoll, um bessere Ergebnisse erzielen zu können.

Eine mögliche Ursache für die selektive Verbesserung des parallelen Hybridmodells bei zwei von vier Emissionsspezies lässt sich bei der Analyse des Imitationsmodells als Teil des Hybridmodells ableiten. Wird die Prädiktionsgenauigkeit des Imitationsmodells hinsichtlich der einzelnen Emissionsspezies (siehe Abbildung 7-2) verglichen, so fällt auf, dass Stickstoffoxide und Kohlenstoffmonoxid qualitativ abgebildet werden können und das Niveau im Mittel dem Messverlauf entspricht. Kohlenstoffdioxid und unverbrannte Kohlenwasserstoffe werden durch das Imitationsmodell hingegen schlechter abgebildet, weder das mittlere Niveau noch der qualitative Verlauf können den Messungen folgen. Daher scheint es, dass das Hybridmodell bei den ersten beiden Emissionsspezies relevante Informationen durch das Einbeziehen der Prädiktionen des Imitationsmodells erhält und dadurch die Messdaten genauer abbilden kann. Erneut ist zu beobachten, dass durch das Training des Hybridmodells nicht zwingend physikalisch korrekte Informationen durch das Imitationsmodell notwendig sind, um die Modellgenauigkeit zu verbessern. Dies ist beispielsweise bei der Berechnung von Stickstoffoxiden und Kohlenstoffmonoxid in transienten Betriebsphasen der Fall – das Hybridmodell ist besser als das rein datenbasierte Modell, obwohl das Imitationsmodell schlechte, jedoch offenbar in brauchbaren Mustern vorliegende Prädiktionen liefert. Bezüglich $CO_2$ und $THC$ kann das Imitationsmodell zu keiner relevanten Verbesserung beitragen. Es wird vermutet, dass das Kombinationsmodell des Hybridmodells in diesem Fall vermehrt auf die Prädiktionswerte des datenbasierten Modells vertraut und nur in wenigen Ausnahmen das Imitationsmodell berücksichtigt. Dies unterstreicht die Wichtigkeit, das Lastprofil bei der Architektur der Hybridmodelle als Eingangsparameter zu integrieren. Dadurch können in Abhängigkeit der Fahrsituation die Emissionsprädiktionsergebnisse der individuellen Modelle korrekt interpretiert werden.

Der Vergleich zwischen der Leistungsfähigkeit des datenbasierten und des parallelen Hybridmodells zeigt erneut, dass die Erklärbarkeit der Prädiktionen von Deep Learning Algorithmen eine Herausforderung darstellt. Dies kann einerseits negativ interpretiert werden, da im Voraus oft nicht eindeutig ist, welche Maßnahmen zu Verbesserung der Modellqualität führen können; andererseits demonstriert der vorliegende Anwendungsfall eindrucksvoll, wie Methoden des Maschinellen Lernens hilfreiche Informationen aus vermeintlich unpräzisen Signalen gewinnen können und dadurch neue Möglichkeiten in der simulativen Abbildung eröffnen.

# 8 Fazit und Ausblick

Das vorrangige Ziel der vorliegenden Arbeit (siehe Kapitel 4) bestand in der Entwicklung innovativer Methoden zur Modellierung von Emissionen im hochtransienten Motorbetrieb. Diesbezüglich wurden mehrere Emissionsmodelle ausgehend von einer physikalisch-phänomenologischen und einer datenbasierten Beschreibung konzipiert, optimiert und analysiert, wobei besonders das datenbasierte Emissionsmodell und das parallele Hybridmodell sehr gute Prädiktionsgenauigkeiten aufweisen.

Daraus lässt sich schlussfolgern, dass beide Modelle als effiziente Werkzeuge für Optimierungs- und Regelungsfunktionen in der modernen Fahrzeugtechnik eingesetzt werden können. Durch die Festlegung der Eingangsparameter auf (auch im dynamischen Betrieb) einfach zu prädizierende Parameter können die Emissionsmodelle ebenfalls für die Optimierung der Betriebsstrategie von Hybridfahrzeugen eingesetzt werden.

Das Potenzial in der Kombination von physikalisch-phänomenologischen und datenbasierten Ansätzen wurde durch die Konzeption von neuartigen Hybridmodellen auf zwei unterschiedliche Arten demonstriert. Dabei bietet die Nutzung des aus dem physikalisch-phänomenologischen Modell abgeleiteten Imitationsmodells den Vorteil, dass die kombinierten Modelle auf einer gemeinsamen Programmiersprache basieren und eine hohe Effizienz in der Inferenz aufweisen. Dadurch können alle betrachteten Modellierungsansätze auch auf leistungsarmer Hardware in Echtzeit eingesetzt werden, was für die direkte Anwendung auf Fahrzeugsteuergeräten unabdingbar ist.

Die vorgestellten Emissionsmodelle unterscheiden sich sehr stark in ihren Prädiktionsgenauigkeiten. Das Imitationsmodell ist dabei am fehleranfälligsten, was nicht primär dem Modellaufbau, sondern dem grundlegenden GT-Suite Modell geschuldet ist. Das GT-Suite Modell entstand in einem mehrstufigen Prozess (Druckverlaufsanalyse, Erstellung des Verbrennungsmodell, Optimierung der Emissionsmodelle), welcher in jedem Schritt auf stationäre Betriebspunkte zurückgreifen musste und zudem auch durch teilweise fehlende Motorparameter (CAD-Geometrien, Wandwärmeverluste, etc.) zusätzliche Ungenauigkeiten beinhaltet. Dadurch kommt es im dynamischen Einsatz zu größeren Abweichungen zwischen Prädiktion und Messung. Für eine bessere physikalisch-phänomenologische Grundlage ist ein erheblicher Mehraufwand notwendig, was die Ressourcen des dieser Arbeit zugrundeliegenden Projektes deutlich überstiegen hätte. Mögliche Verbesserungen könnten beispielsweise durch eigene Strömungsmessungen und -simulationen, die Einbindung nicht in der Software enthaltener und leistungsfähigerer Emissionsmodelle und einen dynamischen Abgleich mit den Messwerten erzielt werden. Die Genauigkeit des vorhandenen Modells ist in dieser Arbeit jedoch keine

Einschränkung, da in dieser frühen, konzeptionellen Entwicklungsphase der neuen Hybridansätze bereits Vorteile demonstriert werden konnten.

Wie auch in den Vorarbeiten bestätigt, erreichen Modellierungsansätze, welche ausschließlich datenbasiert sind, bei der hochdynamischen Prädiktion von Emissionen sehr gute Ergebnisse. Der größte Vorteil bei der Verwendung dieser Methoden liegt in der Zeitersparnis und der Reduktion von anwendungsspezifischem Fachwissen. Dies lässt sich durch die hier vorgestellte Anwendung erörtern, da für das datenbasierte Modell kein physikalisch-phänomenologisches Grundlagenmodell notwendig war. Letzteres erforderte einen hohen Anteil an manueller Arbeit und verbrennungsmotorspezifischer Expertise, um beispielsweise nicht vorhandene Motorkennwerte sinnvoll abschätzen zu können oder die Brennverläufe zu beurteilen. Weiterhin profitieren datenbasierte Modelle von der Vielzahl an unterschiedlichen Architekturen. Im Rahmen dieser Arbeit wurden hauptsächlich Neuronale Netze mit LSTM-Zellen verwendet, je nach Anwendungsfall kann jedoch auch auf andere Arten von Neuronen zurückgegriffen werden, deren Anzahl und Vielfalt auch zukünftig weiter steigen wird.

Hybridmodelle stellen einen äußerst interessanten Ansatz dar, um datenbasierte Modelle weiter zu verbessern. Mit der Steigerung der Informationsdichte können bessere Prädiktionsgenauigkeiten besonders bei einer limitierten Datenmenge erzielt werden. Das parallele Hybridmodell erzielt insgesamt das beste Ergebnis, was besonders beachtlich im Hinblick auf das bereits hohe Leistungsniveau des datenbasierten Modells ist. Es zeichnet sich durch ein kleines (im Sinne einer geringen Tiefe) Kombinationsmodell aus, welches ohne vorher definierte Regeln erlernt, wie es die teilweise stark von den Messungen abweichenden bereitgestellten Informationen verarbeitet, um adäquatere Vorhersagen treffen zu können. Das serielle Hybridmodell kann dahingegen noch nicht überzeugen und erreicht ein schlechteres Ergebnis als das datenbasierte Modell. Es kann aus den physikalischen Parametern des Imitationsmodells keine ausreichenden Muster ableiten, um hochwertige Prädiktionen bereitzustellen.

Modellübergreifend lässt sich beobachten, dass eine Erweiterung der zugrundeliegenden Datenbasis vorteilhaft für ein besseres Training und letztendlich für höhere Generalisierungsfähigkeiten wäre. Dieser Schluss kann daraus gezogen werden, dass bei allen Modellen die Trainings- und Validierungsgenauigkeiten deutlich passender als die Prädiktionsergebnisse bezüglich des Testdatensatzes sind. Das kann daraus folgen, dass die Modelle während des Trainings noch nicht alle im Testzyklus auftretenden Trajektorien der Eingangsparameter erfassen konnten. Die Erweiterung der Trainingsdaten könnte auch größere Modelle zur Folge haben. Im Kontext dieser Arbeit wird die Datengrundlage jedoch nicht als Limitierung bewertet, da die methodische Ausarbeitung innovativer Modelle als primäres Ziel festgelegt wurde.

Dennoch kann die Erweiterung des Datenumfangs als erste Überleitung zum Ausblick auf mögliche Folgearbeiten genutzt werden. Auf Grundlage der hier vorgestellten Inhalte wäre es höchst interessant, die Potenziale der vorgestellten Modellansätze durch ein erweitertes Training und eine mögliche Modellvergrößerung auszureizen.

Neben der Erweiterung der Datengrundlage kann auch ein gegensätzlicher Ansatz verfolgt werden, welcher besonders im praktischen, unternehmerischen Kontext durch eine Zeit- und damit auch Kostenersparnis Vorteile bieten kann. Auf der Grundlage der in Kapitel 5.3 vorgestellten Methode der Systemanalyse könnte gezielt untersucht werden, welche Genauigkeiten Modelle erreichen können, wenn sie lediglich auf diese speziellen Untersuchungen zurückgreifen. Weiterhin sind auch andere Lastzyklen denkbar. Ein mögliches Ziel wäre, einen Betriebszyklus für den vorliegenden Versuchsträger zu finden, welcher mit möglichst geringer Messdauer eine vorher definierte Prädiktionsgenauigkeit für reale Straßenzyklen erreichen kann. Zusätzlich kann nachfolgend untersucht werden, ob sich dieser Betriebszyklus auf andere Verbrennungsmotoren (mit entsprechender Skalierung) übertragen lässt. Dies wäre für die zukünftige Entwicklung und die damit verbundene Modellierung für Regelungs- und Optimierungsfunktionen eine interessante Option.

Ein zusätzlicher Forschungsbedarf kann bei der Analyse der Modelleingangsparameter abgeleitet werden. In der vorliegenden Arbeit wird bewusst auf wenige und einfach zu prädizierende Parameter zurückgegriffen, um auch dem Anwendungsgebiet der Emissionsmodelle bei der Hybridstrategieoptimierung gerecht zu werden. Neben den betrachteten Werten können auch Steuergeräteparameter verwendet werden, welche bereits in einem Modell präzise abgebildet oder tabellarisch definiert sind (z. B. Einspritzzeitpunkt, Zündzeitpunkt, etc.). Davon unabhängig sind auch andere Eingangsparameter denkbar, besonders, wenn diese durch ein anderes Anwendungsgebiet des Modells nicht vorhersagbar sein müssen. Wenn die Emissionsmodelle beispielsweise für eine unmittelbare Emissionsabschätzung verwendet werden würden, könnte auch auf unterschiedliche Sensorwerte (Ladedruck, Abgastemperatur, etc.) zurückgegriffen werden, was den Informationsgehalt direkt erhöhen würde.

Durch den konsequenten Einsatz von Künstlicher Intelligenz kann der gesamte Arbeitsablauf zur Erstellung und Optimierung von Emissionsmodellen und hierauf basierenden Betriebsstrategien effizienter gestaltet werden. Das vorgestellte datenbasierte Modell ist bereits ein erster Schritt in diese Richtung. Es wäre zusätzlich dazu aber auch denkbar, dass die Datenerfassung, die Modellauswahl, die Optimierung und die Validierung des Emissionsmodells automatisiert werden und dadurch ein einsatzfähiges Modell resultiert, ohne dass verbrennungsmotorisches oder auf Maschinelles Lernen ausgerichtetes Fachwissen erforderlich wären. Hierzu müssten Methoden untersucht werden, welche automatisiert Messdaten erfassen, Modellansätze vorschlagen, trainieren und testen und anschließend basierend auf den Abweichungen spezifische neue Messzyklen erstellen. Dieser iterative Prozess könnte bis zu einer definierten Testgenauigkeit fortgeführt werden.

Zusammengefasst bietet die Emissionsmodellierung und generell auch die Abbildung physikalischer Vorgänge durch den Einsatz von datenbasierten und hybriden Methoden enorme Potenziale. Diese wurden im Rahmen der vorliegenden Arbeit an einer hochkomplexen Problemstellung untersucht und demonstriert. Durch die steigenden Rechenleistungen und der damit

einhergehenden Preisreduktion wird auch die Effizienz in der simulativen Abbildung steigen, was einen vermehrten Einsatz der gezeigten Methoden begünstigen kann.

# Literaturverzeichnis


[1] UNFCCC, „Report of the Conference of the Parties on its twenty-first session, held in Paris from 30 November to 13 December 2015. Addendum. Part two: Action taken by the Conference of the Parties at its twenty-first session. | UNFCCC“. Zugegriffen: 13. Juni 2024. [Online]. Verfügbar unter: https://unfccc.int/documents/9097

[2] *Regulation (EU) 2021/1119 of the European Parliament and of the Council of 30 June 2021 establishing the framework for achieving climate neutrality and amending Regulations (EC) No 401/2009 and (EU) 2018/1999 ('European Climate Law')*, Bd. 243. 2021. Zugegriffen: 13. Juni 2024. [Online]. Verfügbar unter: http://data.europa.eu/eli/reg/2021/1119/oj/eng

[3] P. Hofmann, *Hybridfahrzeuge: Grundlagen, Komponenten, Fahrzeugbeispiele*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2023. doi: 10.1007/978-3-662-66894-8.

[4] DIE EUROPÄISCHE KOMMISSION, *VERORDNUNG (EU) 2016/427 DER KOMMISSION vom 10. März 2016 zur Änderung der Verordnung (EG) Nr. 692/2008 hinsichtlich der Emissionen von leichten Personenkraftwagen und Nutzfahrzeugen (Euro 6)*.

[5] H. D. Baehr, *Thermodynamik*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1989. doi: 10.1007/978-3-662-10527-6.

[6] R. Pischinger, Hrsg., *Thermodynamik der Verbrennungskraftmaschine*. in Die Verbrennungskraftmaschine : Neue Folge / hrsg. v. Hans List u. Anton Pischinger, no. 5. 1989.

[7] J. B. Heywood, *Internal combustion engine fundamentals*, Second revised edition. in McGraw-Hill series in mechanical engineering. New York Chicago San Francisco [und 9 andere]: McGraw-Hill Education, 2018.

[8] G. P. Merker und R. Teichmann, Hrsg., *Grundlagen Verbrennungsmotoren: Funktionsweise, Simulation, Messtechnik*. Wiesbaden: Springer Fachmedien Wiesbaden, 2014. doi: 10.1007/978-3-658-03195-4.

[9] K. Reif, D. Lejsek, A. Kufferath, und A. Kulzer, „Grundlagen des Ottomotors“, in *Abgastechnik für Verbrennungsmotoren*, K. Reif, Hrsg., Wiesbaden: Springer Fachmedien Wiesbaden, 2015, S. 19–58. doi: 10.1007/978-3-658-09522-2_2.

[10] R. R. Maly, „Die Zukunft der Funkenzündung“, *MTZ Motortech Z*, Bd. 59, Nr. 7–8, S. XIX–XXIV, Juli 1998, doi: 10.1007/BF03226465.

[11] R. Van Basshuysen und F. Schäfer, Hrsg., *Handbuch Verbrennungsmotor: Grundlagen, Komponenten, Systeme, Perspektiven*. Wiesbaden: Springer Fachmedien Wiesbaden, 2015. doi: 10.1007/978-3-658-04678-1.

[12] R. H. Thring, „Homogeneous-Charge Compression-Ignition (HCCI) Engines“, gehalten auf der 1989 SAE International Fall Fuels and Lubricants Meeting and Exhibition, Sep. 1989, S. 892068. doi: 10.4271/892068.

[13] G. P. Merker und R. Teichmann, Hrsg., *Grundlagen Verbrennungsmotoren: Funktionsweise und alternative Antriebssysteme Verbrennung, Messtechnik und Simulation*. Wiesbaden: Springer Fachmedien Wiesbaden, 2019. doi: 10.1007/978-3-658-23557-4.

[14] A. Wimmer und J. Glaser, *Indizieren am Verbrennungsmotor - Anwenderbuch*, 1. Auflage. Graz: AVL List GmbH, 2002.

[15] A. Witt, „Analyse der thermodynamischen Verluste eines Ottomotors unter den Randbedingungen variabler Steuerzeiten“, Dissertation, TU Graz, 1999.

[16] M. Langwiesner, *Konzepte für bestpunktoptimierte Verbrennungsmotoren innerhalb von Hybridantriebssträngen*. in Wissenschaftliche Reihe Fahrzeugtechnik Universität Stuttgart. Wiesbaden: Springer Fachmedien Wiesbaden, 2018. doi: 10.1007/978-3-658-22893-4.

[17] R. Pischinger, M. Klell, und T. Sams, „Analyse und Simulation des Systems Brennraum“, in *Thermodynamik der Verbrennungskraftmaschine*, Vienna: Springer Vienna, 2009, S. 157–301. doi: 10.1007/978-3-211-99277-7_4.

[18] T. Maurer, *Einführung in die Realprozessrechnung von Verbrennungsmotoren: Modellbildung und Berechnungsprogramm*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020. doi: 10.1007/978-3-662-59262-5.

[19] N. Peters, „Laminar diffusion flamelet models in non-premixed turbulent combustion“, *Progress in Energy and Combustion Science*, Bd. 10, Nr. 3, S. 319–339, Jan. 1984, doi: 10.1016/0360-1285(84)90114-X.

[20] C. Bossung, M. Grill, M. Bargende, und O. Dingel, „A quasi-dimensional charge motion and turbulence model for engine process calculations“, in *15. Internationales Stuttgarter Symposium*, M. Bargende, H.-C. Reuss, und J. Wiedemann, Hrsg., in Proceedings. , Wiesbaden: Springer Fachmedien Wiesbaden, 2015, S. 1001–1019. doi: 10.1007/978-3-658-08844-6_68.

[21] M. Metghalchi und J. C. Keck, „Burning velocities of mixtures of air with methanol, isooctane, and indolene at high pressure and temperature“, *Combustion and Flame*, Bd. 48, S. 191–210, Jan. 1982, doi: 10.1016/0010-2180(82)90127-4.

[22] J. Wallesten, *Modeling of flame propagation in spark ignition engines*. in Doktorsavhandlingar vid Chalmers Tekniska Högskola, no. N.S., 2049. Göteborg: Chalmers Univ. of Technology, 2003.

[23] G. Damköhler, „Der Einfluss der Turbulenz auf die Flammengeschwindigkeit in Gasgemischen“, *Zeitschrift für Elektrochemie und angewandte physikalische Chemie*, Bd. 46, Nr. 11, S. 601–626, Nov. 1940, doi: 10.1002/bbpc.19400461102.

[24] N. Peters, *Turbulent Combustion*, 1. Aufl. Cambridge University Press, 2000. doi: 10.1017/CBO9780511612701.

[25] N. C. Blizard und J. C. Keck, „Experimental and Theoretical Investigation of Turbulent Burning Model for Internal Combustion Engines“, gehalten auf der 1974 Automotive Engineering Congress and Exposition, Feb. 1974, S. 740191. doi: 10.4271/740191.

[26] R. Tabaczynski, F. Trinker, und B. Shannon, „Further refinement and validation of a turbulent flame propagation model for spark-ignition engines“, *Combustion and Flame*, Bd. 39, Nr. 2, S. 111–121, Okt. 1980, doi: 10.1016/0010-2180(80)90011-5.

[27] J. C. Livengood und P. C. Wu, „Correlation of autoignition phenomena in internal combustion engines and rapid compression machines“, *Symposium (International) on Combustion*, Bd. 5, Nr. 1, S. 347–356, Jan. 1955, doi: 10.1016/S0082-0784(55)80047-1.

[28] C. Elmqvist, F. Lindström, H.-E. Ångström, B. Grandin, und G. Kalghatgi, „Optimizing Engine Concepts by Using a Simple Model for Knock Prediction“, gehalten auf der SAE Powertrain & Fluid Systems Conference & Exhibition, Okt. 2003, S. 2003-01–3123. doi: 10.4271/2003-01-3123.

[29] M. Hess, „Klopfmodellierung bei Ottomotoren“, 2023, doi: 10.18419/OPUS-13817.

[30] T. Koch, K. Schänzlin, und K. Boulouchos, „Characterization and Phenomenological Modeling of Mixture Formation and Combustion in a Direct Injection Spark Ignition Engine“, gehalten auf der SAE 2002 World Congress & Exhibition, März 2002, S. 2002-01–1138. doi: 10.4271/2002-01-1138.

[31] B. L. Salvi und K. A. Subramanian, „Experimental investigation and phenomenological model development of flame kernel growth rate in a gasoline fuelled spark ignition engine“, *Applied Energy*, Bd. 139, S. 93–103, Feb. 2015, doi: 10.1016/j.apenergy.2014.11.012.

[32] J. B. Seldowitsch, „The Oxidation of Nitrogen in Combustion and Explosions“, *Acta Physicochimica U.S.S.R.*, Bd. 21, S. 577–628, 1946.

[33] G. A. Lavoie, J. B. Heywood, und J. C. Keck, „Experimental and Theoretical Study of Nitric Oxide Formation in Internal Combustion Engines“, *Combustion Science and Technology*, Bd. 1, Nr. 4, S. 313–326, Feb. 1970, doi: 10.1080/00102206908952211.

[34] D. L. Baulch *u. a.*, „Evaluated Kinetic Data for Combustion Modelling“, *Journal of Physical and Chemical Reference Data*, Bd. 21, Nr. 3, S. 411, Mai 1992, doi: 10.1063/1.555908.

[35] D. L. Baulch *u. a.*, „Evaluated Kinetic Data for Combustion Modeling: Supplement II“, *Journal of Physical and Chemical Reference Data*, Bd. 34, Nr. 3, S. 757–1397, Sep. 2005, doi: 10.1063/1.1748524.

[36] C. P. Fenimore, „Formation of nitric oxide in premixed hydrocarbon flames“, *Symposium (International) on Combustion*, Bd. 13, Nr. 1, S. 373–380, Jan. 1971, doi: 10.1016/S0082-0784(71)80040-1.

[37] J. A. Miller und C. T. Bowman, „Mechanism and modeling of nitrogen chemistry in combustion“, *Progress in Energy and Combustion Science*, Bd. 15, Nr. 4, S. 287–338, Jan. 1989, doi: 10.1016/0360-1285(89)90017-8.

[38] L. V. Moskaleva, W. S. Xia, und M. C. Lin, „The CH+N2 reaction over the ground electronic doublet potential energy surface: a detailed transition state search“, *Chemical Physics Letters*, Bd. 331, Nr. 2–4, S. 269–277, Dez. 2000, doi: 10.1016/S0009-2614(00)01160-X.

[39] J. Sutton, B. Williams, und J. Fleming, „Laser-induced fluorescence measurements of NCN in low-pressure CH4/O2/N2 flames and its role in prompt NO formation“, *Combustion and Flame*, Bd. 153, Nr. 3, S. 465–478, Mai 2008, doi: 10.1016/j.combustflame.2007.09.008.

[40] J. A. Harrington und R. C. Shishu, „A Single-Cylinder Engine Study of the Effects of Fuel Type, Fuel Stoichiometry, and Hydrogen-to-Carbon Ratio on CO, NO, and HC Exhaust Emissions“, gehalten auf der National Automobile Engineering Meeting, Feb. 1973, S. 730476. doi: 10.4271/730476.

[41] C. T. Bowman, „Kinetics of pollutant formation and destruction in combustion“, *Progress in Energy and Combustion Science*, Bd. 1, Nr. 1, S. 33–45, Jan. 1975, doi: 10.1016/0360-1285(75)90005-2.

[42] H. K. Newhall, „Kinetics of engine-generated nitrogen oxides and carbon monoxide“, *Symposium (International) on Combustion*, Bd. 12, Nr. 1, S. 603–613, Jan. 1969, doi: 10.1016/S0082-0784(69)80441-8.

[43] W. K. Cheng, D. Hamrin, J. B. Heywood, S. Hochgreb, K. Min, und M. Norris, „An Overview of Hydrocarbon Emissions Mechanisms in Spark-Ignition Engines“, gehalten auf der International Fuels & Lubricants Meeting & Exposition, Okt. 1993, S. 932708. doi: 10.4271/932708.

[44] G. A. Lavoie und P. N. Blumberg, „A Fundamental Model for Predicting Emissions and Fuel Consumption for the Conventional Spark- Ignition Engine“, gehalten auf der Fall Technical Meeting, Hartford, Connecticut, 1977.

[45] G. A. Lavoie, „Correlations of Combustion Data for S. I. Engine Calculations - Laminar Flame Speed, Quench Distance and Global Reaction Rates“, gehalten auf der 1978 Automotive Engineering Congress and Exposition, Feb. 1978, S. 780229. doi: 10.4271/780229.

[46] K. Yoshimura *u. a.*, „Predicting Unburned Hydrocarbons in the Thermal Boundary Layer Close to the Combustion-chamber Wall in a Gasoline Engine Using a 1-D Model“, *Int.J Automot. Technol.*, Bd. 23, Nr. 1, S. 233–242, Feb. 2022, doi: 10.1007/s12239-022-0020-3.

[47] B. Boust, J. Sotton, S. A. Labuda, und M. Bellenoue, „A thermal formulation for single-wall quenching of transient laminar flames“, *Combustion and Flame*, Bd. 149, Nr. 3, S. 286–294, Mai 2007, doi: 10.1016/j.combustflame.2006.12.019.

[48] W. Ertel, *Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung*. Wiesbaden: Springer Fachmedien Wiesbaden, 2013. doi: 10.1007/978-3-8348-2157-7.

[49] J. McCarthy, Minsky, N. Rochester, und C. E. Shannon, „A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE“. 31. August 1955. Zugegriffen: 23. Februar 2024. [Online]. Verfügbar unter: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

[50] K. P. Murphy, *Machine learning: a probabilistic perspective*, 4. print. (fixed many typos). in Adaptive computation and machine learning series. Cambridge, Mass.: MIT Press, 2013.

[51] E. Alpaydin, *Introduction to machine learning*, Third edition. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2014.

[52] B. Boehmke und B. M. Greenwell, *Hands-on machine learning with R*. in Chapman & Hall/CRC the R series. Boca Raton: CRC Press, 2019.

[53] S. Bhattacharyya, V. Snasel, A. Ella Hassanien, S. Saha, und B. K. Tripathy, Hrsg., *Deep Learning: Research and Applications*. De Gruyter, 2020. doi: 10.1515/9783110670905.

[54] E. Alpaydin, *Maschinelles Lernen*. De Gruyter, 2022. doi: 10.1515/9783110740196.

[55] J. Frochte, *Maschinelles Lernen: Grundlagen und Algorithmen in Python*, 2., Aktualisierte Auflage. München: Hanser, 2019.

[56] R. Xu und D. C. Wunsch, *Clustering*, 1. Aufl. Wiley, 2008. doi: 10.1002/9780470382776.

[57] E. Bilgin, *Mastering reinforcement learning with Python: build next-generation, self-learning models using reinforcement learning techniques and best practices*. Birmingham Mumbai: Packt, 2020.

[58] R. S. Sutton und A. Barto, *Reinforcement learning: an introduction*, Second edition. in Adaptive computation and machine learning. Cambridge, Massachusetts London, England: The MIT Press, 2020.

[59] Y. LeCun, Y. Bengio, und G. Hinton, „Deep learning“, *Nature*, Bd. 521, Nr. 7553, S. 436–444, Mai 2015, doi: 10.1038/nature14539.

[60] G. Sundaram, T. Gehra, J. Ulmen, M. Heubaum, D. Görges, und M. Guenthner, „Modeling of Transient Gasoline Engine Emissions using Data-Driven Modeling Techniques“, gehalten auf der WCX SAE World Congress Experience, Detroit, Michigan, United States, Apr. 2023, S. 2023-01–0374. doi: 10.4271/2023-01-0374.

[61] W. S. McCulloch und W. Pitts, „A logical calculus of the ideas immanent in nervous activity“, *Bulletin of Mathematical Biophysics*, Bd. 5, Nr. 4, S. 115–133, Dez. 1943, doi: 10.1007/BF02478259.

[62] F. Rosenblatt, „The perceptron: A probabilistic model for information storage and organization in the brain.“, *Psychological Review*, Bd. 65, Nr. 6, S. 386–408, 1958, doi: 10.1037/h0042519.

[63] D. Sonnet, *Neuronale Netze kompakt: Vom Perceptron zum Deep Learning*. in IT kompakt. Wiesbaden: Springer Fachmedien Wiesbaden, 2022. doi: 10.1007/978-3-658-29081-8.

[64] A. V. Joshi, *Machine Learning and Artificial Intelligence*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-26622-6.

[65] M. Minsky und S. A. Papert, *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, 2017. doi: 10.7551/mitpress/11301.001.0001.

[66] R. Kruse, C. Borgelt, C. Braune, F. Klawonn, C. Moewes, und M. Steinbrecher, „Mehrschichtige Perzeptren“, in *Computational Intelligence*, Wiesbaden: Springer Fachmedien Wiesbaden, 2015, S. 43–79. doi: 10.1007/978-3-658-10904-2_5.

[67] G. Görz, U. Schmid, und T. Braun, Hrsg., *Handbuch der Künstlichen Intelligenz*. De Gruyter, 2020. doi: 10.1515/9783110659948.

[68] S. Hochreiter und J. Schmidhuber, „Long Short-Term Memory“, *Neural Computation*, Bd. 9, Nr. 8, S. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[69] S. Hochreiter, „Untersuchungen zu dynamischen neuronalen Netzen“, Institut für Informatik, TUM, München, Diplomarbeit, 1991.

[70] T. Hastie, R. Tibshirani, und J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.

[71] R. E. Schapire, „The strength of weak learnability“, *Mach Learn*, Bd. 5, Nr. 2, S. 197–227, Juni 1990, doi: 10.1007/BF00116037.

[72] Y. Freund und R. E. Schapire, „A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting“, *Journal of Computer and System Sciences*, Bd. 55, Nr. 1, S. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[73] L. G. Valiant, „A theory of the learnable“, *Commun. ACM*, Bd. 27, Nr. 11, S. 1134–1142, Nov. 1984, doi: 10.1145/1968.1972.

[74] D. H. Wolpert, „Stacked generalization“, *Neural Networks*, Bd. 5, Nr. 2, S. 241–259, Jan. 1992, doi: 10.1016/S0893-6080(05)80023-1.

[75] C. Ericson, B. Westerberg, M. Andersson, und R. Egnell, „Modelling Diesel Engine Combustion and NOx Formation for Model Based Control and Simulation of Engine and Exhaust Aftertreatment Systems“, gehalten auf der SAE 2006 World Congress & Exhibition, Apr. 2006, S. 2006-01–0687. doi: 10.4271/2006-01-0687.

[76] G. Sundaram, T. Gehra, M. Heubaum, J. Ulmen, D. Görges, und M. Günthner, „Learning-Based Frameworks for Minimizing Pollutant Emissions in Hybrid Electric Vehicles for Dynamic Driving Conditions“, in *2023 IEEE Vehicle Power and Propulsion Conference (VPPC)*, Milan, Italy: IEEE, Okt. 2023, S. 1–6. doi: 10.1109/VPPC60535.2023.10403280.

[77] S. Esposito, L. Diekhoff, und S. Pischinger, „Prediction of gaseous pollutant emissions from a spark-ignition direct-injection engine with gas-exchange simulation“, *International Journal of Engine Research*, Bd. 22, Nr. 12, S. 3533–3547, Dez. 2021, doi: 10.1177/14680874211005053.

[78] L. Cai, A. Ramalingam, H. Minwegen, K. Alexander Heufer, und H. Pitsch, „Impact of exhaust gas recirculation on ignition delay times of gasoline fuel: An experimental and modeling study“, *Proceedings of the Combustion Institute*, Bd. 37, Nr. 1, S. 639–647, 2019, doi: 10.1016/j.proci.2018.05.032.

[79] C. Janssen, „Möglichkeiten zur Prädiktion von unverbrannten Kohlenwasserstoffen in einem direkteinspritzenden Ottomotor“, Dissertation, Universität Rostock, 2010.

[80] S. Frommater, „Phenomenological modelling of particulate emissions in direct injection spark ignition engines for driving cycle simulations“, Dissertation, TU Darmstadt, 2018.

[81] A. Bajwa, G. Zou, F. Zhong, X. Fang, F. Leach, und M. Davy, „Development of a semi-empirical physical model for transient $NO_x$ emissions prediction from a high-

speed diesel engine“, *International Journal of Engine Research*, S. 14680874241255165, Juni 2024, doi: 10.1177/14680874241255165.

[82] P. K. Wong, H. C. Wong, C. M. Vong, Z. Xie, und S. Huang, „Model predictive engine air-ratio control using online sequential extreme learning machine“, *Neural Comput & Applic*, Bd. 27, Nr. 1, S. 79–92, Jan. 2016, doi: 10.1007/s00521-014-1555-7.

[83] S. Lee, H. Choi, und K. Min, „Reduction of engine emissions via a real-time engine combustion control with an egr rate estimation model“, *Int.J Automot. Technol.*, Bd. 18, Nr. 4, S. 571–578, Aug. 2017, doi: 10.1007/s12239-017-0057-x.

[84] R. Lutchen, A. Krätschmer, und H. C. Reuss, „AI-based classification of CAN measurements for network and ECU identification“, *Automot. Engine Technol.*, Bd. 7, Nr. 3–4, S. 317–330, Dez. 2022, doi: 10.1007/s41104-022-00116-6.

[85] N. Papaioannou, X. Fang, F. Leach, A. Lewis, S. Akehurst, und J. Turner, „A Random Forest Algorithmic Approach to Predicting Particulate Emissions from a Highly Boosted GDI Engine“, gehalten auf der 15th International Conference on Engines & Vehicles, Sep. 2021, S. 2021-24–0076. doi: 10.4271/2021-24-0076.

[86] S. Shin, Y. Lee, J. Park, M. Kim, S. Lee, und K. Min, „Predicting transient diesel engine NOx emissions using time-series data preprocessing with deep-learning models“, *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, Bd. 235, Nr. 12, S. 3170–3184, Okt. 2021, doi: 10.1177/09544070211005570.

[87] Q. Huang, J. Liu, C. Ulishney, und C. E. Dumitrescu, „On the use of artificial neural networks to model the performance and emissions of a heavy-duty natural gas spark ignition engine“, *International Journal of Engine Research*, Bd. 23, Nr. 11, S. 1879–1898, Nov. 2022, doi: 10.1177/14680874211034409.

[88] X. (Leo) Fang, N. Papaioannou, F. Leach, und M. H. Davy, „On the application of artificial neural networks for the prediction of $NO_x$ emissions from a high-speed direct injection diesel engine“, *International Journal of Engine Research*, Bd. 22, Nr. 6, S. 1808–1824, Juni 2021, doi: 10.1177/1468087420929768.

[89] M. Nazoktabar, S. A. Jazayeri, M. Parsa, D. D. Ganji, und K. Arshtabar, „Controlling the optimal combustion phasing in an HCCI engine based on load demand and minimum emissions“, *Energy*, Bd. 182, S. 82–92, Sep. 2019, doi: 10.1016/j.energy.2019.06.012.

[90] Z. Zhao *u. a.*, „Physics Informed Neural Network-based High-frequency Modeling of Induction Motors“, *Chin. J. Electr. Eng.*, Bd. 8, Nr. 4, S. 30–38, Dez. 2022, doi: 10.23919/CJEE.2022.000036.

[91] S. Cuomo, V. S. di Cola, F. Giampaolo, G. Rozza, M. Raissi, und F. Piccialli, „Scientific Machine Learning through Physics-Informed Neural Networks: Where we are and What's next“, 2022, doi: 10.48550/ARXIV.2201.05624.

[92] M. Lang, P. Bloch, T. Koch, T. Eggert, und R. Schifferdecker, „Application of a combined physical and data-based model for improved numerical simulation of a medium-

duty diesel engine“, *Automot. Engine Technol.*, Bd. 5, Nr. 1–2, S. 1–20, Juni 2020, doi: 10.1007/s41104-019-00054-w.

[93] F. Steinparzer, C. Schwarz, T. Brüner, und W. Mattes, „Die neuen BMW 3- und 4-Zylinder Ottomotoren mit TwinPower Turbo Technologie“, in *Vortragsunterlagen des Wiener Motorensymposiums 2014*, Wien, 2014.

[94] Gamma Technologies, „Engine Performance Application Manual“, Gamma Technologies, Westmont, USA, 2022.

[95] Y. Zhang *u. a.*, „Characteristics of Transient NOx Emissions of HEV under Real Road Driving“, gehalten auf der WCX SAE World Congress Experience, Apr. 2020, S. 2020-01–0380. doi: 10.4271/2020-01-0380.

[96] V. Vellandi, A. Krishnasamy, und A. Ramesh, „Transient Emission Characteristics of a Light Duty Commercial Vehicle Powered by a Low Compression Ratio Diesel Engine“, gehalten auf der SAE Powertrains, Fuels & Lubricants Digital Summit, Sep. 2021, S. 2021-01–1181. doi: 10.4271/2021-01-1181.

[97] P. Annus, R. Land, M. Min, und J. Ojar, „Simple Signals for System Identification“, in *Fourier Transform - Signal Processing*, S. Salih, Hrsg., InTech, 2012. doi: 10.5772/35697.

[98] W. Khalil und E. Dombre, *Modeling, Identification and Control of Robots*. Elsevier, 2002. doi: 10.1016/B978-1-903996-66-9.X5000-3.

[99] G. Woschni, „A Universally Applicable Equation for the Instantaneous Heat Transfer Coefficient in the Internal Combustion Engine“, gehalten auf der National Fuels and Lubricants, Powerplants, Transportation Meetings, Feb. 1967, S. 670931. doi: 10.4271/670931.

[100] K. Deb und H. Jain, „An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints“, *IEEE Trans. Evol. Computat.*, Bd. 18, Nr. 4, S. 577–601, Aug. 2014, doi: 10.1109/TEVC.2013.2281535.

[101] R. Dudgeon, „Exploring and Improving the GT-SUITE Genetic Algorithm“, Whitepaper, Dez. 2020. [Online]. Verfügbar unter: https://www.gtisoft.com/wp-content/uploads/2020/12/Exploring-and-Improving-the-GT-SUITE-Genetic-Algorithm.pdf

[102] G. I. Taylor, „Statistical Theory of Turbulence“, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, Bd. 151, Nr. 873, S. 421–444, 1935.

[103] *GT-Suite v2022*. Gamma Technologies LLC.

[104] A. Paszke *u. a.*, „PyTorch: An Imperative Style, High-Performance Deep Learning Library“, 2019, doi: 10.48550/ARXIV.1912.01703.

[105] J. Ghorpade, „GPGPU Processing in CUDA Architecture“, *ACIJ*, Bd. 3, Nr. 1, S. 105–120, Jan. 2012, doi: 10.5121/acij.2012.3109.

[106] Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, „Resources of the High Performance Computer ‚Elwetritsch‘“, Apr. 2024. [Online]. Verfügbar unter: https://hpc.rz.rptu.de/elwetritsch/hardware.shtml

[107] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, und L. Shao, „Normalization Techniques in Training DNNs: Methodology, Analysis and Application“, 27. September 2020, *arXiv*: arXiv:2009.12836. Zugegriffen: 17. April 2024. [Online]. Verfügbar unter: http://arxiv.org/abs/2009.12836

[108] D. Singh und B. Singh, „Investigating the impact of data normalization on classification performance“, *Applied Soft Computing*, Bd. 97, S. 105524, Dez. 2020, doi: 10.1016/j.asoc.2019.105524.

[109] N. Henze, „Kovarianz und Korrelation“, in *Stochastik für Einsteiger*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2023, S. 215–230. doi: 10.1007/978-3-662-67729-2_21.

[110] F. Hutter, L. Kotthoff, und J. Vanschoren, Hrsg., *Automated Machine Learning: Methods, Systems, Challenges*. in The Springer Series on Challenges in Machine Learning. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-05318-5.

[111] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, und R. Salakhutdinov, „Dropout: A Simple Way to Prevent Neural Networks from Overfitting“, *Journal of Machine Learning Research*, Bd. 15, Nr. 56, S. 1929–1958, 2014.

[112] B. Bischl *u. a.*, „Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges“, *WIREs Data Min & Knowl*, Bd. 13, Nr. 2, S. e1484, März 2023, doi: 10.1002/widm.1484.

[113] D. P. Kingma und J. Ba, „Adam: A Method for Stochastic Optimization“, 2014, doi: 10.48550/ARXIV.1412.6980.

[114] J. A. Westerhuis *u. a.*, „Assessment of PLSDA cross validation“, *Metabolomics*, Bd. 4, Nr. 1, S. 81–89, März 2008, doi: 10.1007/s11306-007-0099-6.

[115] J. Bergstra, R. Bardenet, Y. Bengio, und B. Kégl, „Algorithms for hyper-parameter optimization“, in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, in NIPS'11. Red Hook, NY, USA: Curran Associates Inc., Dez. 2011, S. 2546–2554.

[116] S. Watanabe, „Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance“, 2023, doi: 10.48550/ARXIV.2304.11127.

# Lebenslauf

| | |
|---|---|
| Name | Tobias Gehra |
| Geburtsort | Sindelfingen |
| Staatsangehörigkeit | deutsch |
| <u>Bildungsweg</u> | |
| Seit Mai 2019 | **RPTU Kaiserslautern-Landau**, Kaiserslautern<br>*Wissenschaftlicher Mitarbeiter, Lehrstuhl für Antriebe in der Fahrzeugtechnik* |
| April 2016 – Feb. 2019 | **Technische Universität Kaiserslautern,** Kaiserslautern<br>*Studium Fahrzeugtechnik, Abschluss Master of Science Fahrzeugtechnik* |
| Okt. 2011 – Feb. 2016 | **Hochschule Kaiserslautern,** Kaiserslautern<br>*Studium Maschinenbau*, A*bschluss Bachelor of Engineering* |
| Okt. 2010 – Juli 2011 | **Technische Universität Kaiserslautern**, Kaiserslautern<br>*Studium Wirtschaftsingenieurwesen, Schwerpunkt Maschinenbau* |
| Sep. 2001 – März 2010 | **Gymnasium Ramstein-Miesenbach**, Ramstein-Miesenbach<br>*Abschluss Abitur* |
| Aug. 1997 – Juni 2001 | **Grundschule**, Ramstein-Miesenbach |

# Antriebe in der Fahrzeugtechnik

herausgegeben von Prof. Dr.-Ing. Michael Günthner
2941-4326

| | | | |
|---|---|---|---|
| 1 | David Woike | Die gezielte Steuerung des Ladungswechsels als Werkzeug in der Brennverfahrensentwicklung | |
| | | ISBN 978-3-8325-5678-5 | 45.00 € |
| 2 | Florian Müller | Innermotorische Effizienzsteigerung luftansaugender und gemischansaugender Dual-Fuel Brennverfahren | |
| | | ISBN 978-3-8325-5757-7 | 80.00 € |
| 3 | Tobias Mink | Analyse der Zusammenhänge von Öl- und Partikelemissionen am Ottomotor mit Direkteinspritzung | |
| | | ISBN 978-3-8325-5796-6 | 45.50 € |
| 4 | Tobias Gehra | KI-unterstützte hybride Modellierung von Emissionen im hochtransienten Motorbetrieb | |
| | | ISBN 978-3-8325-5940-3 | 44.50 € |

In der Entwicklung emissionsarmer Antriebstechnologien bieten batterieelektrische Fahrzeuge durch den hohen Gesamtwirkungsgrad und die lokal emissionsfreie Betriebsweise einen vielversprechenden Lösungsansatz. Gleichzeitig stehen ihnen Herausforderungen wie die notwendige Ladeinfrastruktur und in der Regel geringere Reichweiten gegenüber. Hybridfahrzeuge können diese Nachteile kompensieren, indem sie die Vorteile elektrischer und konventioneller Antriebe vereinen.

Um das Potenzial von Hybridantrieben auszuschöpfen, sind Betriebsstrategien für die optimale Lastverteilung notwendig. Diese lassen sich effizient in virtuellen Umgebungen entwickeln, was unter anderem eine präzise Abbildung des Fahrzeugs erfordert. Vor diesem Hintergrund setzt sich die vorliegende Arbeit mit der Modellierung von Emissionen im hochtransienten Motorbetrieb auseinander. Neben den physikalisch-phänomenologischen Ansätzen kommen auch Methoden des Maschinellen Lernens (ML) zum Einsatz, die besonders bei dynamischen Fahrzyklen überzeugen.

Durch die Kombination beider Modellansätze (Hybridmodelle) wird eine noch genauere Emissionsvorhersage ermöglicht. Die Ergebnisse zeigen, dass das parallele Hybridmodell zu einer deutlichen Präzisionssteigerung (ca. 25 % gegenüber dem reinen ML-Modell) führt und durch eine effiziente Architektur für Echtzeit-Anwendungen prädestiniert ist. Dies gilt nicht nur für die vorliegende Problemstellung, sondern könnte sich auch auf andere Forschungsbereiche ausweiten lassen.

**Logos Verlag Berlin**

**ISBN 978-3-8325-5940-3** **ISSN 2941-4326**