MAX PLANCK INSTITUTE FOR CHEMISTRY
Max Planck Graduate Center mit der Johannes Gutenberg-Universität
JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Machine learning for the elucidation of multiphase processes and systems

Matteo Krüger
Born on 10.10.1993 in Bingen am Rhein, Germany

Dissertation
for the award of the academic degree of
'Doctor rerum naturalium '(Dr. rer. nat.) of the faculties:
08 - Physics, Mathematics and Computer Science
09 - Chemistry, Pharmaceutical Sciences, Geography and Geosciences
10 - Biology
University Medicine

Mainz, November 2025

Supervisor:

Second supervisor:

Day of the oral examination: 21 January 2026

I hereby declare that I wrote the dissertation submitted without any unauthorized external assistance and used only sources acknowledged in the work. All textual passages which are appropriated verbatim or paraphrased from published and unpublished texts as well as all information obtained from oral sources are duly indicated and listed in accordance with bibliographical rules. In carrying out this research, I complied with the rules of standard scientific practice as formulated in the statutes of Johannes Gutenberg-University Mainz to insure standard scientific practice.

Matteo Krüger
Mainz, 4 November 2025

Abstract

Atmospheric chemistry governs many processes in the context of air quality, climate, and human health. It includes, among others, multiphase interactions between gases and condensed phases, affecting the formation, transformation and removal of atmospheric constituents. Understanding multiphase chemistry is therefore essential for accurately describing atmospheric composition and its impacts. In this context, atmospheric aerosols play a particularly important role. They serve as cloud nuclei, transport reactive species, provide reaction surface and impact radiative forcing, the key driver of global warming. Among them, organic aerosols represent an abundant fraction, yet their formation and transformation remains poorly understood, largely contributing to uncertainty in climate and health modelling. The immense diversity of precursor compounds for organic aerosols, coupled with complex environmental conditions and multiphase chemistry, pose a major challenge for laboratory and modelling studies alike. At the same time, advances in measurement and monitoring technologies generate large volumes of atmospheric data. Data-centric methods like machine learning pose a powerful opportunity to use these data to forward the understanding of atmospheric processes and improve parameterizations of Earth system models. This study explores and advances machine learning applications in multiphase chemistry, including compound property prediction, model acceleration, uncertainty quantification, experiment design, and multiscale modelling. The focal points of this thesis can be summarized as follows:

1. Advancement of quantitative structure-activity relationship (QSAR) models with novel artificial neural network architectures. This includes the application of pattern-detecting convolutional neural networks on one-hot encoded simplified molecular input line entry system (SMILES) representations of molecular structures to estimate the reduction potentials of atmospherically-relevant quinones. Reduction potentials determine the quinones' reactivity, and therefore their ability to undergo redox-cycling, which leads to the catalytic production of reactive oxygen species (ROS). ROS are likely associated with adverse health effects of air pollution, as they react with biomolecules in the lungs of exposed humans. Trained convolutional neural network models permit the rapid screening of atmospherically-relevant quninones that pose an elevated risk of adverse health effects, without the need of expensive measurements. Another advancement of QSAR methods is the application of graph convolutional neural networks alongside graph representations of molecules to estimate their saturation vapor pressure. Vapor pressures determine the partitioning equilibrium of atmospheric compounds between the condensed and gas phase. Their accurate determination is of high relevance in various fields of atmospheric science, critically affecting the formation and growth of secondary organic aerosol, which account for a substantial mass fraction of tropospheric aerosols. A novel group contribution-assisted,

adaptive-depth graph convolutional neural network architecture outperforms existing methods for vapor pressure prediction, even when trained on relatively few experimental data.

2. Acceleration of kinetic multilayer process models for mass transport and chemical reactions in aerosols through machine learning surrogate models. While differential equation models are capable of accurately simulating the growth and aging of atmospheric aerosols, their computational expense often poses restrictions on their application in inverse or global atmospheric modelling. Artificial neural network and polynomial chaos expansion surrogate models can be trained on sampling data from differential equation models to reproduce their output. The trained surrogate models show high accuracy in comparison with their reference models, and offer immense acceleration by multiple orders of magnitude. They were successfully used to aid complex differential equation models in inverse modelling tasks, notably reducing the computational cost.

3. Development of the Numerical Compass, a computational method for automated uncertainty quantification of process models and experiment design. The Numerical Compass utilizes a fit ensemble, i.e., an ensemble of plausible model solutions consistent with experimental data, to identify combinations of laboratory parameters for experiments that lead to near-optimal model constraints. This enables researchers to use a more target-oriented approach when designing experiments by addressing the major sources of uncertainty in parametric models. The method was thoroughly tested in simulations of the ozonolysis of oleic acid.

4. Application of parametric equations and neural network models for the modelling of gas exchange rates of biological soil crusts to advance the calculation of the net primary production and long-term carbon balance of dryland ecosystems. Biological soil crusts, communities of photoautotrophic and heterotrophic organisms form a common ecological feature in dryland areas across the globe and are an important factor for carbon balance and cycling. NN and parametric equation models trained on laboratory data are capable of describing carbon dioxide (CO_2) gas exchange rates of various types of biological soil crusts based on the prevailing environmental conditions. The models are accurate, versatile and are applied to investigate the effect of ambient CO_2 concentrations on biological soil crusts in a follow-up study.

Zusammenfassung

Atmosphärenchemie bestimmt eine Vielzahl an Prozessen im Kontext von Luftqualität, Klima und Gesundheit. Sie umfasst unter anderem Multiphaseninteraktionen zwischen Gasen und kondensierten Phasen und erklärt die Bildung, Transformation und Entfernung atmosphärischer Bestandteile. Das Verständnis von Multiphasenchemie ist daher essenziell für die detaillierte Beschreibung der atmosphärischen Zusammensetzung und deren Einflüsse. In diesem Kontext kommt Aerosolen eine besondere Bedeutung zu. Sie dienen als Kondensationskeime in Wolken, transportieren reaktive Verbindungen, bieten Reaktionsoberflächen und beeinflussen den Strahlungsantrieb, eine entscheidende Komponente der globalen Erderwärmung. Organische Aerosole stellen einen bedeutenden Anteil aller Aerosole dar, doch ihre Bildung und Umwandlung sind bisher nur unzureichend verstanden, was maßgeblich zur Unsicherheit in Klimamodellen beiträgt. Die enorme Vielfalt an unterschiedlichen Verbindungen in Kombination mit komplexen Umweltbedingungen stellt eine große Herausforderung sowohl für Labor-, als auch für Modellierungsstudien dar. Gleichzeitig resultieren aus Fortschritten in Mess- und Monitoring-Technologien große Mengen atmosphärischer Daten. Datenzentrierte Methoden wie maschinelles Lernen sind eine vielversprechende Möglichkeit, solche Daten zu nutzen, um das Verständnis atmosphärischer Prozesse voranzubringen und Parametrisierungen von Erdsystemmodellen zu verbessern. Diese Arbeit befasst sich mit Anwendungen des maschinellen Lernens in der Multiphasenchemie, unter anderem zur Modellbeschleunigung, Vorhersage von Stoffeigenschaften, Unsicherheitsquantifizierung, experimentellen Versuchsplanung und Multiskalen-Modellierung. Die Schwerpunkte der Dissertation lassen sich wie folgt zusammenfassen:

1. Entwicklung von quantitativen Struktur-Wirkungs-Beziehungsmodellen (QSAR) durch den Einsatz neuartiger Architekturen künstlicher neuronaler Netze. Dazu gehört die Anwendung von mustererkennenden Convolutional Neural Networks, trainiert auf one-hot-codierte Simplified Molecular Input Line Entry System (SMILES)-Darstellungen molekularer Strukturen zur Abschätzung der Reduktionspotentiale atmosphärisch relevanter Chinone. Reduktionspotentiale bestimmen die Reaktivität von Chinonen und damit ihre Fähigkeit zum Redox-Cycling, das zur katalytischen Bildung reaktiver Sauerstoffspezies (ROS) führt. ROS stehen wahrscheinlich in Zusammenhang mit negativen Gesundheitseffekten durch Luftverschmutzung, da sie mit Biomolekülen in der Lunge exponierter Menschen reagieren. Trainierte Convolutional Neural Network-Modelle ermöglichen ein schnelles Screening von Verbindungen, die ein erhöhtes Risiko gesundheitsschädlicher Effekte bergen. Darüber hinaus erfolgt im Kontext von QSAR die Nutzung von Graph-Convolutional Neural Networks in Kombination mit Graph-Darstellungen von Molekülen zur Abschätzung ihres Sättigungsdampfdrucks. Dampfdrücke bestimmen das Verteilungsgleichgewicht atmosphärischer Verbindungen zwischen kondensierter und Gasphase. Ihre Bestimmung ist in ver-

schiedenen Bereichen der Atmosphärenwissenschaften von großer Relevanz, da sie die Bildung und das Wachstum sekundärer organischer Aerosole entscheidend beeinflussen, welche einen erheblichen Massenanteil troposphärischer Aerosole ausmachen. Eine neuartige Graph-Convolutional Neural Network Architektur mit Unterstützung von Gruppenbeitragsmethoden und adaptiver Tiefe übertrifft konventionelle Methoden zur Dampfdruckvorhersage, selbst wenn nur wenige experimentelle Daten zum Training zur Verfügung stehen.

2. Beschleunigung kinetischer, Multischicht-Prozessmodelle für Massentransport und chemische Reaktionen in Aerosolen durch Emulatoren des maschinellen Lernens. Während Differentialgleichungsmodelle in der Lage sind, das Wachstum und die Alterung atmosphärischer Aerosole präzise zu simulieren, ist ihr Einsatz in inverser Modellierung und in globalen Klimamodellen oft durch einen hohen Rechenaufwand beschränkt. Emulatoren basierend auf künstlichen neuronalen Netzen oder Polynomial Chaos Expansion, welche auf Simulationsdaten trainiert werden, zeigen im Vergleich zu den Referenzmodellen eine hohe Genauigkeit, und ermöglichen eine deutliche Beschleunigung um mehrere Größenordnungen. Sie wurden erfolgreich zur Unterstützung von rechenaufwändigen Referenzmodellen in inversen Modellierungsaufgaben angewendet.

3. Entwicklung des Numerical Compass, einer numerischen Methode zur automatisierten Quantifizierung der Unsicherheit von Prozessmodellen und zur Versuchsplanung. Der Numerical Compass nutzt ein Fit-Ensemble, d.h. ein Ensemble plausibler Modellösungen in Übereinstimmung mit experimentellen Daten, um Kombinationen von Umwelt- oder Versuchsparametern mit hohem Optimierungspotential zu identifizieren. Dies ermöglicht Forschenden einen zielgerichteteren Ansatz bei der Planung von Experimenten, indem die größten Unsicherheitsquellen in parametrischen Modellen adressiert werden. Die Methode wurde umfassend in Simulationen für die Ozonolyse von Ölsäure getestet.

4. Anwendung parametrischer Gleichungen und neuronaler Netzwerke zur Modellierung der Gasaustauschraten biologischer Bodenkrusten zur Verbesserung von Simulationen der Nettoprimärproduktion und des langfristigen Kohlenstoffhaushalts von Trockengebieten. Biologische Bodenkrusten, Gemeinschaften aus photoautotrophen und heterotrophen Organismen sind weltweit in Trockengebieten verbreitet und stellen eine wichtige Komponente im globalen Kohlenstoffkreislauf dar. Künstliche neuronale Netze und Modelle auf Basis parametrischer Gleichungen, die mit Labordaten trainiert wurden, sind in der Lage, die Kohlenstoffdioxid (CO_2)-Gasaustauschraten verschiedener Typen biologischer Bodenkrusten in Abhängigkeit von ihren Umweltbedingungen zu beschreiben. Die Modelle liefern genaue Ergebnisse, sind vielseitig einsetzbar, und werden in einer Folgestudie verwendet, um den Einfluss der Umgebungs-CO_2-Konzentration auf biologische Bodenkrusten zu untersuchen.

Acknowledgements

I would like to begin by expressing my gratitude to my parents - and -, and family for their unwavering support and encouragement throughout the course of my academic journey. I am especially grateful to my siblings - and -, and brother-in-law - for their experience and for providing balance during the more demanding periods of this work. I also extend my heartfelt thanks to - for her patience, understanding, and unconditional love; her confidence in my abilities and constant support were crucial to the completion of this work. I further wish to acknowledge my friends outside academia, most importantly -, -, - and -, who were consistently available to discuss my concerns and who offered welcome opportunities for exercise and shared recreational activities, providing a much-needed sense of balance.

I am deeply indebted to my mentor, -, for his invaluable guidance and insight. His intellectual rigor, constructive feedback, and consistent support have been fundamental to the development of this thesis, and he was always an exemplary scientist and a role model to me. My deep gratitude also extends to - for the opportunity to pursue my PhD at the Max-Planck Institute, and for all genuine care and support, not only in scientific matters. I also wish to thank the members of my dissertation committee, -, - and -, for their time, thoughtful comments, helpful suggestions, and admirable expertise.

I would also like to acknowledge my colleagues and collaborators, especially in the Max-Planck Institute for Chemistry, whose discussions, assistance, and camaraderie have contributed greatly to both the progress of this research and my professional development. The supportive environment within and beyond the group has made this endeavor productive, intellectually stimulating, and fun at the same time. Thank you, -, -, -, -, -, -, -, -, -, -, -, -, -, -, -, - and everyone else who at some point was with me on this journey.

I would further like to express my sincere appreciation to the administrative and technical staff, whose efficiency and dedication ensured that everything ran smoothly behind the scenes. In particular, I thank -, - and - for their help with scheduling, communication, and countless organizational matters. Their professionalism and kindness have greatly facilitated my work and created an environment in which I could focus fully on my research.

List of Abbreviations

biocrust	Biological Soil Crust
CNN	Convolutional Neural Network
CO_2	Carbon Dioxide
GCNN	Graph Convolutional Neural Network
GC^2NN	Group Contribution-Assisted Graph Convolutional Neural Network
KM-SUB	Kinetic Multi-layer Model of Aerosol Surface and Bulk Chemistry
ML	Machine Learning
NN	Artificial Neural Network
NC	Numerical Compass
PM	Particulate matter
QSAR	Quantitative Structure-Activity Relationship
ROS	Reactive Oxygen Species
RMSE	Root Mean Square Error
SMILES	Simplified Molecular Input Line Entry System
SOA	Secondary Organic Aerosol
VOC	Volatile Organic Compound

Contents

1. Introduction

1.1. Atmospheric Aerosols and Multiphase Processes

Atmospheric aerosols, suspensions of fine solid or liquid particles in the atmosphere, are of great relevance in the Earth system. They occur in various sizes and originate from a variety of natural sources, including sea spray (Vignati et al., 2010), mineral dust (Kok et al., 2023), volcanoes, wildfires or biological sources like pollen or bacteria (Lam et al., 2011). Anthropogenic sources of atmospheric aerosols include combustion and other industrial processes, transportation and agriculture (Mukherjee and Agrawal, 2017; Pozzer et al., 2017; Hopke et al., 2020). Depending on the atmospheric conditions, aerosols can undergo physical and chemical transformations that alter their size distribution, composition and atmospheric lifetime (Rudich et al., 2007; Berkemeier et al., 2014).

The impact of aerosols on the Earth system is manifold: climatically, they affect radiative forcing directly through scattering and absorption of solar and terrestrial radiation, and indirectly, by serving as cloud condensation nuclei or ice-nucleating particles (Haywood and Boucher, 2000). Such processes impact cloud properties, and therefore albedo and precipitation (Hoose et al., 2009; Tao et al., 2012). Fine particulate matter with a diameter less than 2.5 μm (PM2.5) can penetrate deeply into the human respiratory system and cause adverse health effects like asthma and cardiovascular diseases (Burnett et al., 2014; Cohen et al., 2017). Recent estimations globally attribute up to 7-9 million premature deaths to air pollution each year (Burnett et al., 2018; Lelieveld et al., 2020). Adverse health effects of air pollution are likely associated with the production of reactive oxygen species (ROS) in the respiratory system (Li et al., 2003; Shiraiwa et al., 2012b). Such ROS can be formed through redox-cycling of reactive compounds like transition metals or quinones, conjugated cyclic dione structures, which commonly occur in PM2.5 (Monks et al., 1992; Charrier et al., 2014). The ability of chemical compounds to undergo redox cycling is determined by their reduction potential (Roginsky et al., 1999).

A major difficulty in aerosol sciences is chemical complexity. Atmospheric aerosols contain a wide range of compounds including inorganic salts such as sulfates, nitrates and ammonium (Xu and Penner, 2012). In addition, they may incorporate a vast array of organic compounds. Secondary organic aerosols (SOA) are particularly abundant and therefore have major impacts on the climate (Kanakidou et al., 2005; Jimenez et al., 2009; Shrivastava et al., 2017). Volatile organic compounds (VOCs) are emitted from both natural and anthropogenic sources and undergo aging processes in the atmosphere, e.g., oxidation or photochemistry (George et al., 2015; Mellouki et al., 2015). Depending

on the volatility of organic compounds, they partition into the particle phase (Wilson et al., 2021). Due to the immense diversity of organic precursors and wide spectrum of oxidation pathways, atmospheric aerosols, especially SOA, pose a major uncertainty in atmospheric modelling studies (Intergovernmental Panel On Climate Change (IPCC), 2023).

Multiphase chemistry on atmospheric aerosols - coupled mass transport and chemical reactions across gas, liquid and solid phases - can be described through kinetic multilayer models (Shiraiwa et al., 2010). Physical mass transfer (e.g., gas-particle partitioning and diffusion) and chemical transformations (e.g., particle-phase reactions and oligomerization) modify aerosol size, composition, and their ability to act as cloud condensation nuclei (Tolocka et al., 2004; Shiraiwa et al., 2012a; Li et al., 2019). The processes are governed by physical properties like volatility, viscosity, diffusivity, or reaction rates of involved compounds (Worsnop et al., 2002; Roldin et al., 2014; Shiraiwa et al., 2014).

Mechanistic process models like the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) are based on differential equations which explicitly couple transport and reaction kinetics within particles (Pöschl et al., 2007; Kolb et al., 2010). These models are capable of simulating even complex multiphase chemistry processes, for instance the production of ROS in the respiratory tract through interaction of various pollutants and biomolecules (Lelieveld et al., 2021; Dovrou et al., 2023; Mishra et al., 2023). Alongside experimental data, multiphase chemical kinetics models can be used to constrain or deduct chemical and physical parameters, a process referred to as inverse modelling (Berkemeier et al., 2017).

While such models capture the fundamental physics and chemistry and offer a deep understanding of multiphase systems, their computational cost grows rapidly with their complexity, i.e., the number of involved compounds and interactions, often making them impractical for inclusion in large-scale models (Semeniuk and Dastoor, 2020). This mismatch of deep mechanistic description with large-scale applicability leads to the wide-spread use of parametrizations, which attempt to capture the net effect of complex systems using simplified descriptors (Christensen and Zanna, 2022).

1.2. Gas Exchange of Biological Soil Crusts

While aerosols, by modifying Earth's cloud and surface albedo, impact radiative forcing and thus global warming, the main contributor to climate change are increased greenhouse gas concentrations due to anthropogenic emissions (Al-Ghussain, 2019). In this context, land-atmosphere gas exchange processes play a crucial role, as they regulate atmospheric composition by mediating the exchange of carbon dioxide, methane, and water vapor. The biosphere functions as a key regulator in this system, acting as both source and sink of greenhouse gases through photosynthesis, respiration or decomposition (Steiner, 2020).

Biological soil crusts (biocrusts), communities of photoautotrophic cyanobacteria, algae, lichens, and bryophytes, are a key feature of dryland ecosystems across the globe, and contribute significantly to this exchange (Housman et al., 2006; Pointing and Belnap, 2012). Therefore, accurate estimations of carbon dioxide (CO_2) gas exchange rates of biocrusts as a response to environmental parameters such as light, temperature, water content, and ambient CO_2 concentration are crucial to assess their long-term net primary production and impact on global warming (Lange, 2002).

1.3. Machine Learning and Data Science in Atmospheric Chemistry

Advances in instrumentation rapidly expand the data available in atmospheric science. New and larger data sets arise from more efficient laboratory experiments, monitoring networks, remote sensing, and computational simulations (DeCarlo et al., 2006; Wang et al., 2017; Giles et al., 2019; Tabor et al., 2019; Besel et al., 2023; Vance et al., 2024). However, the sheer volume and complexity of these data may overwhelm traditional modelling strategies (Benedetti et al., 2018; Sandström et al., 2023). Consequently, complementary approaches that leverage new data sets and scale efficiently to the needs of atmospheric models are of increased interest in the field of atmospheric science (Gianquintieri et al., 2024).

In this context, data-centric methods and machine learning (ML) applications offer promising new opportunities. ML describes a variety of algorithms to analyze and draw inferences from patterns in data through statistical methods, without following explicit instructions. ML applications and related data-centric approaches provide the means to detect patterns, reduce dimensionality and potentially develop predictive models (Ruske et al., 2018; Nair et al., 2021; Yorks et al., 2021; Tang et al., 2024). Among machine learning algorithms, artificial neural networks (NN) are a group of algorithms inspired by biological brains, where complex signal processing and response patterns are based on comparably simple interactions of large numbers of interconnected nodes, or neurons (Kröse and van der Smagt, 1996). NN are commonly organized in layers, where an individual neuron obtains signals from neurons in the previous layer and maps them to a single new signal that is passed to neurons of the following layer.

One area in atmospheric science where ML applications have demonstrated great potential is in quantitative structure-activity relationship (QSAR). QSAR models attempt to establish a link between the molecular structure and certain physico-chemical properties of chemical species (Pankow, 1987; Donahue et al., 2009; Compernolle et al., 2011; Li et al., 2016). In atmospheric chemistry, QSAR models are particularly useful for the prediction of vapor pressures, Henry's law constants or redox potential, as these properties are crucial for the understanding of partitioning, reactivity or toxicity (Bilde and Pandis, 2001; Sander, 2015). Trained on data sets of experimental measurements or quantum-mechanical calculations, successful QSAR models allow for accurate predictions and rapid screening of large quantities of atmospheric compounds, thus

addressing vast chemical diversity of organic molecules in the atmosphere (Lumiaro et al., 2021; Galeazzo and Shiraiwa, 2022; Armeli et al., 2023).

Beyond molecular property prediction, ML can be applied in the context of aerosol and air quality modelling to predict concentrations or sources of pollutants like PM2.5 or ozone from monitoring network, meteorological, or satellite data, often outperforming traditional statistical models (Masmoudi et al., 2020; Muthukumar et al., 2022). Furthermore, unsupervised ML like clustering methods can be used to pre-process measurement data, e.g., through dimensionality reduction or automated source apportionment analysis (Christopoulos et al., 2018; Kumar et al., 2022). Other emerging applications are the use of ML to develop surrogate models for computationally costly process simulations (Weber et al., 2020; Yang et al., 2024).

In summary, these examples illustrate how ML has the potential to reshape the landscape of atmospheric chemistry and aerosol science. It opens new avenues to handle high-dimensional data, accelerate, interpret or improve complex models, and uncover relationships in data that are difficult to capture with traditional approaches.

1.4. Research Objectives

While traditional methods and modelling approaches continue advancing our understanding in the field of atmospheric chemistry, its complexity and multi-scale character pose a challenge. On the other hand, ML and other data-centric methods revolutionize many aspects of science and everyday life. Especially with regards to the rapid growth and progressing automatization of measurement techniques, there is an increased demand in the use of novel methods and their integration into traditional approaches in the context of atmospheric chemistry and aerosol science. Thus, the research goals of this PhD project were to explore ML techniques to aid, advance or substitute multiphase chemical kinetics models, and to bridge the gaps between different fields and scales in atmospheric chemistry.

The specific objectives and activities of this PhD work can be summarized as follows:

1. Explore novel techniques and neural network architectures like convolutional neural networks (CNN) or graph convolutional neural networks (GCNN) in the context of QSAR modelling to enhance and accelerate the prediction of relevant physicochemical properties like reduction potential or saturation vapor pressure of compounds in the vast chemical space of atmospherically relevant organic compounds. Benchmark the new models against common traditional methods and provide a rigorous framework for future benchmarking of QSAR models.

2. Develop ML-based surrogate models for computationally expensive multiphase chemical kinetics models to accelerate aerosol simulations at high accuracy. Explore their potential in inverse modelling tasks as well as in bridging the gaps between various scales of atmospheric modelling approaches.

3. Develop a new framework for uncertainty analysis using ensemble methods in order to guide future experiments based on a process model and recent experimental

data. Apply the framework to a model system and perform simulations to prove that the suggested experiments are near-optimal to constrain kinetic parameter uncertainty. Provide the framework as an efficient, easy-to-use open source program.

4. Use ML and high performance computing to solve computational constraints and other difficulties, especially in the context of multiscale modelling and big data analysis, e.g., in the context of ecosystem net primary production modelling.

2. Results

2.1. Overview

The main findings and methods of this PhD project are described in three first and one second author papers published in peer-reviewed scientific journals. One additional first author manuscript is currently under peer-review and one more is in preparation. An overview of the studies and the broader picture, i.e., which research fields and aspects they address, is given in Fig. 2.1. Each project falls into at least one of three main fields of machine learning application in atmospheric chemistry:

1. Inverse modelling and optimization to infer or constrain physical parameters from mechanistic models and laboratory data.
2. Property prediction to efficiently screen the vast space of organic compounds with regards to certain physicochemical properties.
3. Multiscale modelling, to bridge the gaps between small-scale and large-scale systems and models.

Chapter 2.2 addresses advances in quantitative structure-activity relationship (QSAR) modelling for the prediction of reduction potentials of quinones, common atmospheric compounds that, depending on their reactivity, may undergo redox-cycling and produce harmful reactive oxygen speices (ROS) in the respiratory tract. Pattern-detecting convolutional neural networks (CNN) are presented as suitable models for redox potential estimation. This work is expanded on in chapter 2.3, where adaptive-depth group contribution assisted graph convolutional neural networks (GC^2NN) are utilized to accurately determine the saturation vapor pressure of atmospheric compounds, an important parameter that affects a compound's partitioning equilibrium between the condensed and gas phase. Chapter 2.4 addresses the utilization of machine learning (ML) methods to generate fast surrogate models for complex kinetic multilayer models, facilitating their applicability in multiscale and inverse modelling tasks where large quantities of model evaluations are required. In chapter 2.5, a novel method for uncertainty quantification and experiment design is presented, the Numerical Compass (NC). This chapter also links to chapter 2.4, as fast surrogate models are successfully applied to significantly reduce the computational effort required for the evaluation of the NC method. Chapter 2.6 demonstrates how artificial neural networks (NN) and parameteric equations can be used to address multiscale modelling problems, i.e., linking the gas exchange of biological soil crusts (biocrusts) in laboratory experiments to the net primary production of dryland ecosystems.

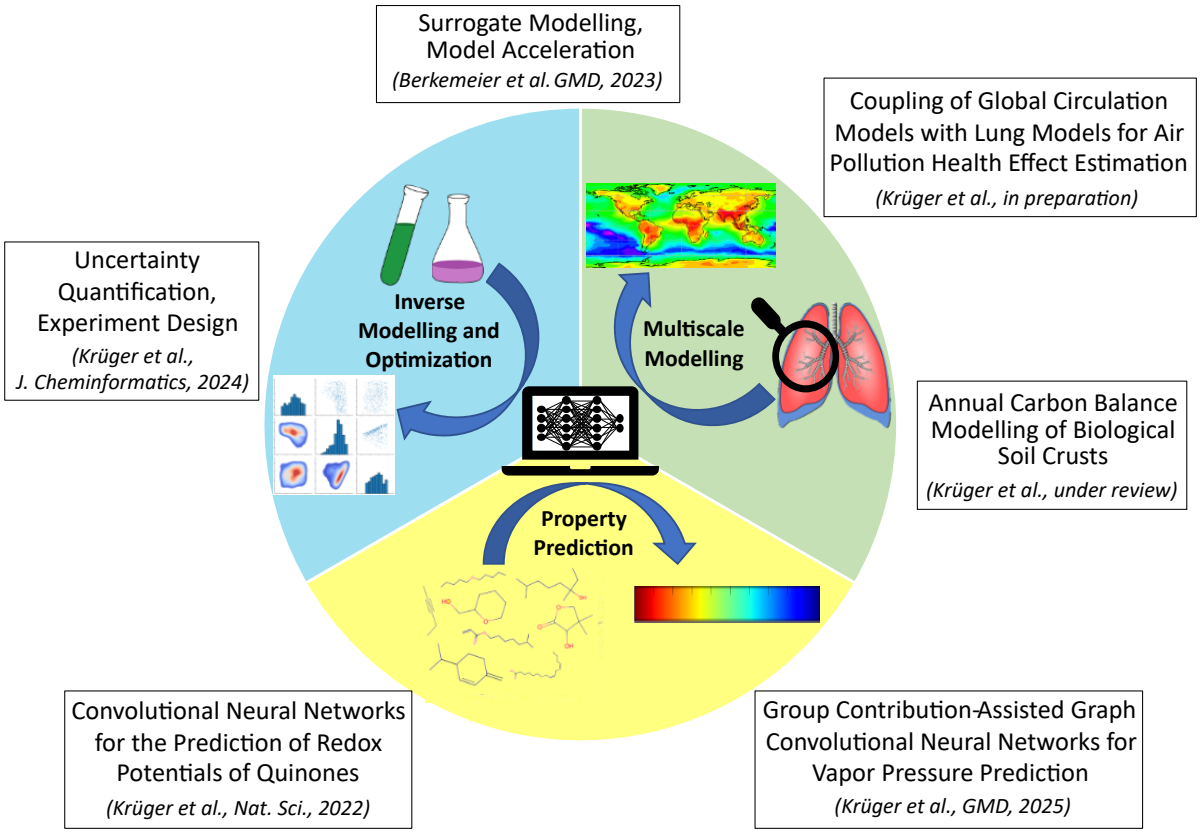

Figure 2.1.: *Overview of the research fields covered by the PhD project, and associated publications and manuscripts for peer-reviewed journals. Shown are only first- and second-author papers.*

2.2. Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects

This chapter presents a research article published in the new highlight journal from Wiley, Natural Sciences. I am the first author and the main contributor to this paper. This project originated in my master's thesis, where I investigated CNN models for reduction potential prediction. I subsequently expanded the work and prepared the paper during my PhD studies. I compiled and pre-processed the training data, explored ML architectures suitable for the task and wrote all code for the initialization, training and evaluation of the CNN model. I also prepared and supervised the model training using parallelization and high performance computing. Finally, I prepared all figures, wrote the manuscript draft together with Thomas Berkemeier and revised the manuscript during peer-review, assisted by my co-authors. More detailed information on the author contributions are provided at the end of the manuscript.

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B.A.M., Lammel, G., Schmidt, B., Pöschl, U., Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, *Nat. Sci.*, doi: 10.1002/ntls.20220016, (2022).

Convolutional neural networks (CNN) were used as quantitative structure-activity relationship (QSAR) models to relate the one-electron reduction potentials of quinones to their molecular structure. For CNN training and testing, a data set of more than 100,000 quinones with associated reduction potential values derived from density functional theory calculations was encoded in simplified molecular input line entry system (SMILES). The best performing CNN model strongly outperformed linear regression models fitted on common molecular descriptors. Augmentation methods were newly adapted or applied to support CNN training with smaller data sets, improving the root mean square error (RMSE) by up to approximately 37% for a data set of only 321 molecules. Using the newly developed model, a subset of atmospherically relevant quinones was identified, that are likely to have a high oxidative potential and may play a role in aerosol health effects. The supplement to this work can be found in appendix B1.

Received: 12 April 2022 | Revised: 30 July 2022 | Accepted: 2 August 2022

DOI: 10.1002/ntls.20220016

Natural Sciences

RESEARCH ARTICLE

Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects

Matteo Krüger[1] | **Jake Wilson**[1] | **Marco Wietzoreck**[1] | **Benjamin A. Musa Bandowe**[1] | **Gerhard Lammel**[1] | **Bertil Schmidt**[2] | **Ulrich Pöschl**[1] | **Thomas Berkemeier**[1]

[1]Multiphase Chemistry Department, Max Planck Institute for Chemistry, Mainz, Germany

[2]Institute of Computer Science, Johannes Gutenberg University, Mainz, Germany

Correspondence
Thomas Berkemeier, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany.
Email: t.berkemeier@mpic.de

Funding information
Max-Planck-Gesellschaft

Abstract

Quinones are chemical compounds commonly found in air particulate matter (PM). Their redox activity can generate reactive oxygen species (ROS) and contribute to the oxidative potential (OP) of PM leading to adverse health effects of aerosols. The quinones' OP and ability to form ROS are linked to their reduction potential (RP, measured in volts), a metric for the tendency to lose electrons in redox reactions. Here, we use convolutional neural networks (CNN) as quantitative structure-activity relationship (QSAR) models to relate the one-electron RP of quinones to their molecular structure. For CNN training and testing, a data set of more than 100,000 quinones with associated RP values derived from density functional theory calculations was encoded in simplified molecular input line entry system (SMILES). The best performing CNN model achieved a root mean square error (RMSE) of 0.115 V for an independent test data set and outperformed linear regression models fitted on common molecular descriptors (≥ 0.140 V RMSE). Augmentation methods were newly adapted or applied to support CNN training with smaller data sets, improving RMSE by up to approximately 37% for a data set of 321 molecules. Adjusted for solvent effects, the CNN-derived RP predictions showed good agreement with experimental data. Using the newly developed method, we identified a subset of atmospherically relevant quinones that are likely to have a high OP and play a role in aerosol health effects, which remains to be further elucidated by experimental studies. We suggest to use the presented machine learning approach in further investigations of atmospheric aerosol chemistry and health effects as well as other studies that require a target-oriented screening of the properties and effects of large classes of substances.

KEYWORDS
quinone, redox chemistry, environmental chemistry, redox potential, oxidative potential, convolutional neural network, machine learning, QSAR

Key Points

1. Convolutional neural networks can be used to estimate unknown physical and chemical properties of chemical substances and outperform additive group contribution methods.
2. Augmentation methods aid in the prevalent problem of data availability.
3. Quinone species detected in the environment are screened for potential relevance in atmospheric chemistry and public health and presented in this study.

INTRODUCTION

Particulate matter (PM) is the aggregate of nano- to micrometer-sized solid or liquid particles suspended in air. Origin and composition of PM are diverse over time and space.[1,2] Short-term increases of PM exposure to humans have shown to increase mortality and morbidity in adults and children, mostly associated with heart conditions,[3] asthma,[4] and chronic obstructive pulmonary disease.[5] Long-term exposure to PM is linked to an increased risk of lung cancer,[6] asthma,[7] and cardiovascular diseases.[8] Among several mechanisms explaining the adverse health effects of PM, reactive oxygen species (ROS) production and the resulting oxidative stress are considered main contributors.[9–11] In addition to inflammation processes in exposed tissue as a response to oxidative stress, a continuous imbalance between ROS formation and antioxidant activity in a cell can lead to damage of lipids, proteins, RNA, or DNA.[12,13]

The contribution of individual components of PM to the formation of ROS is not fully understood and difficult to disentangle, as a large variety of chemical compounds take part in these reactions. Typical reactive compounds in PM are ROS such as H_2O_2, environmentally-persistent free radicals, as well as redox-active components such as transition metals and quinones.[14–17] These redox-active compounds often participate in catalytic redox cycles, which lead to formation and interconversion of ROS.[18–21] The oxidative potential (OP) of redox-active substances has been proposed as a descriptor for this ability. OP is typically quantified with cellular assays or acellular antioxidant assays like the dithiothreitol (DTT) or ascorbic acid assay.[22–26]

Quinones are conjugated cyclic dione structures, which are ubiquitously present in the human environment, including ambient PM.[27] Their electrophilic properties have been associated with cellular damage through modifications of proteins and DNA and their redox properties can make them efficient redox cyclers.[27,28] Estimates for the contribution of quinones to the OP of PM vary considerably across previous studies and depend on PM composition. Atmospheric concentrations up to 2 ng/m^3 have been reported for quinones, and they can account for up to 20% of DTT consumption by atmospheric aerosol samples.[29–31]

Experimental studies find the OP of quinones to differ widely based on their chemical structure.[30] Roginsky et al.[32] showed a correlation between the one-electron reduction potential (RP), an indicator for the tendency to participate as electron acceptor or donor in redox reactions, and the effective kinetic rate constant of oxygen consumption, k_{eff}, of quinones in ascorbic acid (AscH) solution. These solutions form significant amounts of semiquinone (SQ^-) through reduction of the quinone (Q) and oxidation of AscH to the ascorbyl radical (Asc, Equation (1)). The semiquinone is able to consume molecular oxygen, recycling the quinone and forming superoxide (O_2^-, Equation (2)). This ability of quinone/semiquinone pairs to catalytically form ROS and deplete antioxidants results in high OP in acellular assays.[26] Thus, following up on Wietzoreck et al.,[26] we propose that an estimation of the OP of a quinone could be derived from its RP.

$$Q + AscH \rightarrow SQ^- + Asc + H^+, \quad (1)$$

$$SQ^- + O_2 \rightarrow Q + O_2^-. \quad (2)$$

In addition to the potential relevance of the RP of quinones for their OP, it may provide information regarding several other aspects and applications of quinone chemistry. A quinone's lifetime under specific conditions, and therefore environmental persistence, is strongly dependent on its redox reactivity. Quinones also play a role as intermediate products in metabolic processes related to drugs, and their RP can be an indicator for toxicity.[33] Recently, there has been increased interest in quinones due to their potential as active materials in aqueous redox flow batteries. Such batteries are based on inexpensive organic materials that fit certain requirements regarding their physicochemical properties, including their RP.[34,35]

Common methods for obtaining the RP include experimental approaches such as cyclic voltammetry[36] or computational approaches such as quantum-mechanical calculations based on density functional theory,[37] hence requiring time-consuming laboratory experiments or large computational effort. With preexisting data, however, data-centric approaches for prediction of the RP are feasible. Existing approaches often utilize group contribution methods, where chemical or structural properties, for example the presence and position of functional groups or substituents, are mapped to thermodynamic properties of the associated molecule using algebraic equations.[38–41]

In this work, convolutional neural network (CNN) models are applied as data-centric approach to predict the RP of compounds based on their molecular structure. Over the last years, the field of machine learning, especially deep learning, has undergone massive developments in conjunction with new applications, addressing various problems in science and daily life.[42] Artificial neural networks are inspired by the underlying functionality of biological neuronal structures and represent mathematical models that link well-defined input

to output vectors and can be trained on data with the goal of optimizing prediction errors.[43] A CNN is a specific deep neural network type that uses convolution operations, in which so-called kernels are applied on the data. Element-wise mathematical operations of such kernels with sections of the input data allow the detection of patterns by a trained CNN model.[44]

To use molecular structures as input for CNN models, the molecules need to be represented as arrays that allow element-wise operations. Simplified molecular input line entry system (SMILES) is a molecule encoding system for chemical informatics applications, which applies graph theory to transform molecular structures into textual strings.[45,46] The textual representation can then be transformed to sparse binary matrices using one-hot encoding, where each possible character is assigned a reference level that can be indicated present (1) or absent (0) at a given string position.[47] This results in a sparse, two-dimensional binary matrix unambiguously encoding a single molecule.

Similar quantitative structure-activity relationship (QSAR) approaches with machine learning algorithms have been proposed in various studies. Hirohara et al.[48] applied CNN models for the detection of chemical motifs, aiming for fast screening of lead components for new drugs. For this purpose, they encoded molecular structures in SMILES, followed by one-hot encoding. Sanchez-Lengeling et al.[49] presented objective-reinforced generative adversarial networks (ORGAN), a method that allows generation of nonrepetitive, sensible molecular structures with a bias of the generative distribution toward specified attributes. Similarly, translation into SMILES followed by one-hot encoding is applied for structural encoding.

Successful applications of machine learning usually require large amounts of training data. In some cases, data augmentation methods can be used to artificially enhance the training data by adding data points that represent modified original data. An effective modification represents a new data point that improves the ability of the machine learning model to extract correlations and principles from the original data, despite not adding any new information. Data augmentation of images by cropping, rotating, and flipping has shown to improve the accuracy of deep learning classification models in various studies.[50,51] Text data augmentation, however, is not very common and mainly based on semantics, which cannot be directly related to the processing of molecule structures based on their SMILES representation.[52,53] When deviating from the stricter rule set of canonical SMILES, multiple variants of a SMILES representation of a molecule are possible, for instance, by allowing different starting positions of the enumeration within the molecule. This method has been called RandomSMILES and was successfully applied to increase training data sets and boost model performance.[54,55]

The objective of this study is the prediction of the standard one-electron RP of quinones based on molecular structure. Different CNN models are trained on subsets of computationally derived RP values of quinones by Tabor et al.[56] and Kristensen et al.[57], who applied density functional theory to identify quinone–hydroquinone pairs with RP values that would allow application in redox flow batteries. The CNN models are evaluated against multiple linear regression models using molecular descriptors that are common in cheminformatics: Molecular ACCess System (MACCS),[58] topological fingerprints developed for RDKit,[59] and Morgan fingerprints[60] that all can be obtained from the structural information provided in SMILES representations without the computationally costly application of quantum mechanical calculations. To test transferability to other data sets with dissimilar chemical profiles, a CNN model trained only on the Tabor et al.[56] data is applied to predict the RP of quinone species from the Kristensen et al.[57] data. Furthermore, we use a CNN model to predict the RP of quinone species reported in the literature and provide predictions for a compilation of quinones that have been found in atmospheric samples, act as metabolites in living systems, or have been associated with lipid peroxidation. These predictions are used to estimate OP and to compile a list of quinones that may be of relevance to public health or atmospheric chemistry based on their high RP. Additionally, data augmentation and feature map representation, common techniques in the application of CNN models on images, are adapted to the application on molecule representations.

METHODS

Data and preprocessing

The workflow of data preprocessing, data augmentation, model training, and model evaluation applied in this study is shown in Figure 1. Data used in this work were obtained from literature and originates from two different studies. Tabor et al.[56] systematically generated quinone molecular structures based on a set of backbones (benzene, naphthalene, anthracene, phenanthrene, and tetracene) and possible transformations through Michael addition and subsequent tautomerizations. They employed Gaussian-type orbital calculations on either hybrid or GGA functionals to calculate reaction energies for reduction and gem–diol reactions and calibrated the RP with PM7 COSMO and B3LYP methods.[56] They kindly provided access to the full data set of 146,857 quinone–hydroquinone pairs and associated RP and solvation free energy. Filtering for quinones and removing duplicates results in a set of 103,040 samples (referred to as "Tabor_all" data in the following). A smaller subset of 69,598 quinones is obtained from this data by removing all quinones containing sulfate functional groups (referred to as "Tabor_nosulf" data), which may be of less relevance in an atmospheric context. Kristensen et al.[57] investigated quinones of various biological sources and reported RP and solvation free energy of each compound, using the Perdew-Burke-Ernzerhof (PBE) functional and the 6-31G** basis set. We used an automated script to derive canonical SMILES for 321 compounds in this data set (referred to as "Kristensen" data in the following). Further descriptions of the data sets and applied methods are provided in Supporting Information Note 1. Note that the Tabor and Kristensen data sets exhibit different chemical profiles. While average molecular size is similar, distribution of functional groups varies strongly as evident from Table 1.

From the larger Tabor_all and Tabor_nosulf data sets, 5000 molecules are excluded randomly as test data and fully withheld

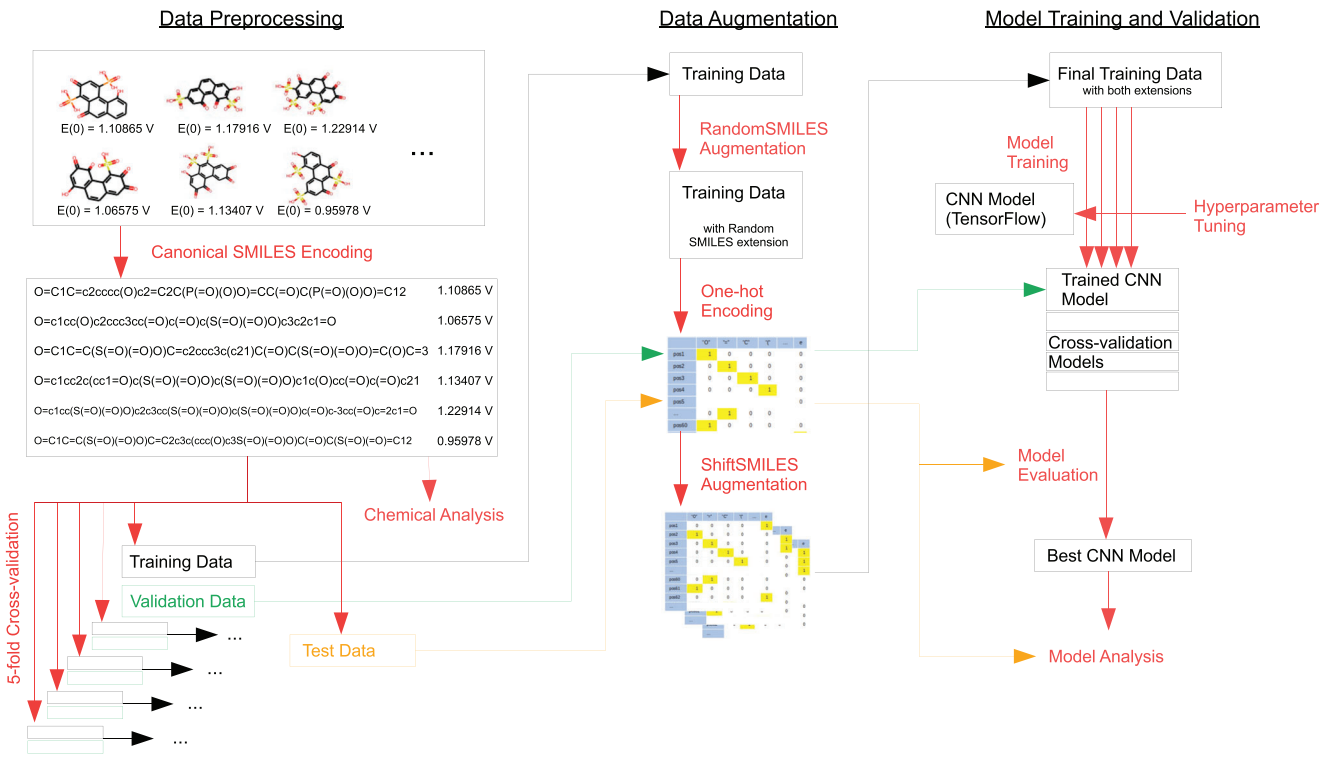

FIGURE 1 Schematic workflow illustrating the three main steps: data preprocessing, data augmentation, as well as model training and validation. Black objects represent data or model structures. Red arrows and labels represent individual methods and steps. Green and orange objects refer to subprocesses regarding validation and test data, respectively.

TABLE 1 Chemical profiles of the three data sets used in this study. The values represent average numbers of atoms, rings, functional groups, or the average oxygen/carbon ratio (O/C) for all molecules in the associated data sets.

Data Set	Atoms	Rings	Keto	Alkyl	Hydroxyl	Carboxyl	Amine
Tabor_all	26.5	3.0	3.6	0.0	4.5	0.7	–
Tabor_nosulf	26.0	3.0	4.0	0.0	4.8	1.1	–
Kristensen	25.8	2.9	2.4	6.8	1.7	0.1	0.0

	Amide	Nitrile	Sulfide	Sulfonic	O/C	RP, V
Tabor_all	–	–	–	0.7	0.7	0.6
Tabor_nosulf	–	–	–	–	0.7	0.6
Kristensen	0.1	0.0	0.0	–	0.3	0.2

from CNN model training and validation. After extraction of this test set, *k*-fold cross-validation is applied a common standard practice for machine learning applications to detect and avoid overfitting of the trained models during the optimization of hyper-parameters, as described in Section 2.4.[61,62] We use the method provided in the Python library scikit-learn,[63] set the number of consecutive folds (*k*) to 5, and use 80% of data for training and 20% for validation in this study. The smaller Kristensen data set is used to test and compare data augmentation methods. We apply double cross-validation,[64,65] where the splitting of data into test, validation, and training set is performed in two nested loops, with five folds each. The individual test sets contain 64 molecules, the validation sets 51, and the training sets 206 molecules. Data splits are applied before data augmentation and identical for all tested degrees of augmentation. Note that augmentation methods are not applied to the validation or test data. Additionally, a larger data set (referred to as "Tabor_Kristensen") is obtained by merging the two original data sets, similarly excluding 5000 molecules randomly as test data and applying fivefold cross-validation on the remaining data.

To transform the SMILES strings into numerical data as input for any CNN model, one-hot encoding is applied. To ensure a consistently sized input matrix for molecules with SMILES strings of different lengths, we fix the input matrix to the size of the largest SMILES string and extend smaller SMILES strings with additional empty characters. The resulting sparse binary matrix representing a single molecule is of size 30×166, where 30 is the number of different SMILES characters and 166 the number of characters in the longest SMILES string. Note that the Tabor data set the longest SMILES string has only 92 characters. To make a CNN model applicable for both data sets, the larger input data shape is selected.

Data augmentation: RandomSMILES and ShiftSMILES

We test two augmentation methods to artificially increase the training data set size in an application with limited training data, using only the Kristensen data set. On the level of SMILES encoding, we

apply the function RandomSMILES, provided in the Python library RDKit,[59] to obtain multiple strings for individual molecules. After one-hot encoding, when molecules are represented by binary matrices, we apply another augmentation method, ShiftSMILES, which has not been described previously. In the ShiftSMILES data augmentation method, the molecule-encoding strings are shifted along the empty characters within the reading frame that is determined by the largest string in the data set. Thus, instead of containing empty characters only at the end, strings may also contain empty characters at the beginning, column-shifting the binary matrix encoding the molecular structure. Optionally, in order to allow ShiftSMILES augmentation of particularly long SMILES strings, all SMILES may be expanded by a number of fields determined by an additional parameter, *ef* (extend frame). *ef* is set to 5 in this study. Note that the consecutive application of the two methods allows the generation of ShiftSMILES augmentations based on molecule representations that are generated by the RandomSMILES method.

As the number of possible augmentations of both methods depends on the size of the molecule, we add augmentations in two steps for each method, to ensure a similar number of augmentations for each molecule in the original training data. For each molecule, m augmented representations are obtained. From all representations of all molecules, n samples are then added to the original training data. m and n for RandomSMILES and ShiftSMILES will be referred to as m_r and n_r as well as m_s and n_s, respectively. More detailed descriptions of the augmentation methods and their joint application are provided in Supporting Information Note 2 and Figure S1. In the augmentation experiment presented in this study, m_r and m_s are set individually to ensure that n_r and n_s can be satisfied in any case. A more detailed explanation of this process and the numerical values used in this study are provided in Supporting Information Note 2.

Convolutional neural networks

Artificial neural networks can be described as a large number of computational nodes or "neurons" that are organized, partly interconnected, and can be activated individually based on signals from other nodes.[43] The resulting network or model represents a nonlinear function that transforms an input to an output vector. Each node itself represents a function, mapping a weighted sum of its inputs to an output, which, in turn, is passed to following nodes, forming a so-called fully connected layer. Training refers to the adjustment of weights by application of an optimization algorithm minimizing a loss function based on the final model output. In their entirety, the determination of the weights and their adaptation represent the "learning" process.

CNN contain not only fully connected layers, but also so-called convolutional layers and pooling layers. Convolutional layers contain neurons that calculate the scalar product of their weights and parts of the output of the previous layer. They allow a trained CNN model to detect subpatterns in the input data, which are associated with the model output. Convolutional layers are often followed by pooling layers, which perform downsampling along the spatial dimensionality of the given input. To achieve downsampling, the presence of features is summarized in patches of the output of the previous convolutional layer, or "feature map." This can be achieved either by summarizing the largest activation for the presence of a feature or by summarizing the average presence of a feature. One or multiple fully connected layers then translate from the information passed from the pattern detecting section of the network to the desired classification or regression output.[44,66]

In this study, CNN models with three convolutional layers are trained to predict the RP based on encoded molecular structures. Each convolutional layer is followed by a pooling layer and the network architecture is concluded by a global pooling layer and a fully connected layer. Molecular structures, represented by two-dimensional matrices, serve as a model input. Accordingly, convolution operations are executed over two axes. For more details regarding convolutional layers and the architecture of the CNN models presented in this work, see Supporting Information Note 3 and Figure S2.

Model training and hyperparameter tuning

The implementation of a CNN usually requires the providing of hyperparameters that control model architecture and learning process. Variation of these parameters may lead to significant effects on training success and model performance. We perform hyperparameter tuning individually for each CNN model for the different training data sets, as elaborated in Table S1. The best set of hyperparameters is selected under consideration of the average validation and average test set mean squared error of the cross-validation models, as their comparison allows for detection of potential overfitting. Of the cross-validation models, the one trained on the last subset of data is arbitrarily and continuously selected for further evaluation. Note that we did not choose the one with the lowest validation or test error to ensure comparability across different applications of the same model or across different models in the evaluation. Names, descriptions, and tested ranges for relevant hyperparameters provided by the Keras library[67] in Python are presented in Table S1. A set of hyperparameters that performed well on large data sets (Tabor_all, Tabor_nosulf, or Tabor_Kristensen data set) is given in the following:

Conv_layers: 3, filters: (128, 256, 512), window_size_c: (8, 4, 2), stride_size_c: (1, 1, 1), padding_c: ("valid," "valid," "valid"), pooling_type: ("average," "average," "average"), window_size_p: (3, 4, 5), stride_size_p: (1, 2, 2), padding_p: ("valid," "valid," "valid"), activation_function: ("relu," "relu," "relu"), global_pooling: 1, batch_size: 4, batch_norm: 0, epochs: 40, learning_rate: 0.0005, decay: 1, dropout: 0.1.

For the data augmentation experiment based on the much smaller Kristensen data set, we perform very basic hyperparameter tuning with a decreased model complexity to avoid overfitting. For the overall best performing model, the number of filters is reduced to 32, 64, and 128, respectively, the number of training epochs is set to 24, and the learning rate increased to 0.001.

Linear regression model

The CNN approach for the prediction of RP values based on SMILES representations is compared with linear models (LMs) that are trained on multiple molecular descriptors: MACCS structural keys (MACCS), topological fingerprints (TopoFP), and Morgan fingerprints (MorganFP). The descriptors are obtained directly from SMILES representations by application of the RDKit library.[59] The approach resembles traditional group contribution approaches, where predictive models are based on a molecule's functional groups and their interactions.[41] We apply an identical data split into training and test data for the CNN models and LMs. From a number of statistical tools to map the structural parameters to the RP of the molecules in the data, including LinearRegression, SGDRegressor, KernelRidge, and ElasticNet from the Python library scikit-learn,[63] the multiple linear regression model (LinearRegression) performed best and is referred to as "LM" in this work.

RESULTS AND DISCUSSION

Evaluation and comparison of CNN and linear model

Figure 2(a) presents test set predictions of the CNN model using the full Tabor_all data set (CNN_Tabor_all). The CNN performs well in this regression task, showing a high correlation coefficient R^2 of 0.872 between CNN model and test set. Average validation RMSE of the five cross-validation models at finalization of the training is 0.117 V. The variance of the validation error is very low at 1.23×10^{-6} V^2, indicating a very similar performance of the five cross-validation models. As we select only one model for further evaluation, validation error variance must be sufficiently low to ensure that our results can be generalized and are not associated with the data composition of the individual fold. The very similar RMSE of predictions for the test data (0.115 V) of the selected model and validation error (0.119 V) indicates no relevant overfitting. Models that are overfitted are usually associated with a significantly larger test set error in comparison with the validation or training set error, as they lack generalization to data not encountered in the training process. Exclusion of the quinones with sulfate groups leads to a CNN model average RMSE of 0.120 V for the validation set and 0.120 V for the test set.

Figure 2(b) shows the correlation of the LM fitted to topological fingerprints of the Tabor_all set (LM_Tabor_all), scoring a significantly higher RMSE of 0.140 V. Fitting an LM to Tabor_nosulf (LM_TopoFP_Tabor_nosulf) results in a test set RMSE of 0.137 V. A comparison and overview of the trained models in this work can be found in Table 2. Overall, model errors associated with the Tabor_nosulf data are very similar in comparison to the Tabor_all data set.

Learning curves for the CNN, obtained by training models on data subsets of various sizes, are presented in Figure 3. As an increase from 50,000 to 98,040 samples in training and validation data still significantly decreases model error and error variance, we expect that additional data derived from quantum mechanical calculations would allow for more accurate CNN models.

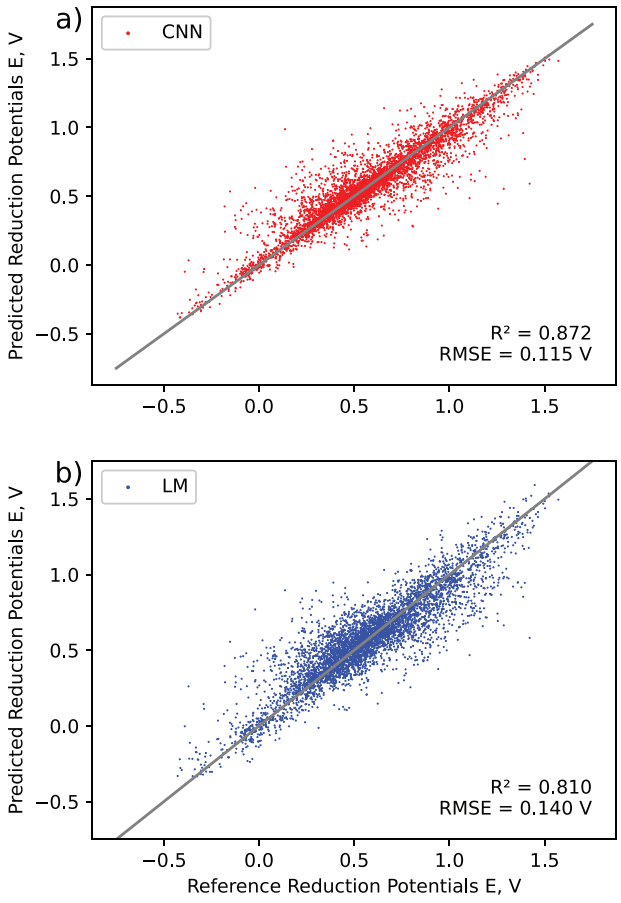

FIGURE 2 Correlation scatter plots for (a) the CNN model and (b) the linear model fitted to topological fingerprints for the 5000 quinones in the separated test data, originating from the Tabor_all data set. The models are both trained on or fitted to the full remaining data in the Tabor_all data set after exclusion of the test data to predict the associated RP.

Transferability to other data sets

To test the transferability of prediction models between data sets, we apply models that are trained on the Tabor data sets to predict the full Kristensen data, in this section referred to as "transfer set" (Figure 4). Note that the chemical profiles of the molecules differ between the Tabor and Kristensen data sets as summarized in Table 1. Multiple functional groups (amine, amide, nitrile, and sulfide) occur exclusively in the Kristensen data. The average O/C ratio, number of keto groups, as well as the RP are much larger for the molecules in the Tabor data.

The resulting prediction errors of the models to the Kristensen data are significantly larger compared to test data sets of the Tabor data: the RMSE of the CNN model trained on Tabor_all is 0.267 V and the model trained on the Tabor_nosulf data is 0.270 V. With correlation coefficients (R^2) < 0.3, the predictive capability of the CNN models in this transferability experiment is far below standard requirements of predictive regression models (> 0.6).[68] The linear models LM_MorganFP and LM_MACCS fail to accurately predict the RP of most quinones in this data set as RP predictions of some quinones are far outside a phys-

TABLE 2 Summary of the CNN and linear models (LM) trained in this work, including the number of samples in the different sets and the RMSE and R^2 of the models at the end of the training for validation and test set. For models that are trained only on Tabor_all or Tabor_nosulf data, an RMSE for their application on the Kristensen data is further displayed. The linear models are just fitted to the full training data, and consequently, no validation sample number and validation error are provided.

Model name	Training samples	Validation samples	Validation set RMSE, V	
CNN_Tabor_all	78,427	19,607	0.119	
CNN_Tabor_nosulf	51,674	12,919	0.120	
CNN_Tabor_Kristensen	79,012	19,754	0.118	
LM_TopoFP_Tabor_all	98,034	–	–	
LM_TopoFP_Tabor_nosulf	64,593	–	–	
LM_MorganFP_Tabor_all	98,034	–	–	
LM_MorganFP_Tabor_nosulf	64,593	–	–	
LM_MACCS_Tabor_all	98,034	–	–	
LM_MACCS_Tabor_nosulf	64,593	–	–	
	Test set RMSE, V	**Test set R^2**	**Transfer set RMSE, V**	**Transfer Set R^2**
CNN_Tabor_all	0.115	0.872	0.267	0.281
CNN_Tabor_nosulf	0.120	0.862	0.270	0.262
CNN_Tabor_Kristensen	0.116	0.870	–	–
LM_TopoFP_Tabor_all	0.140	0.810	0.389	−0.524
LM_TopoFP_Tabor_nosulf	0.137	0.822	0.362	−0.323
LM_MorganFP_Tabor_all	0.145	0.795	8.45×10^{9}	−0.109
LM_MorganFP_Tabor_nosulf	0.144	0.802	1.63×10^{10}	0.077
LM_MACCS_Tabor_all	0.250	0.392	8.31×10^{10}	−0.274
LM_MACCS_Tabor_nosulf	0.260	0.357	1.38×10^{11}	−0.094

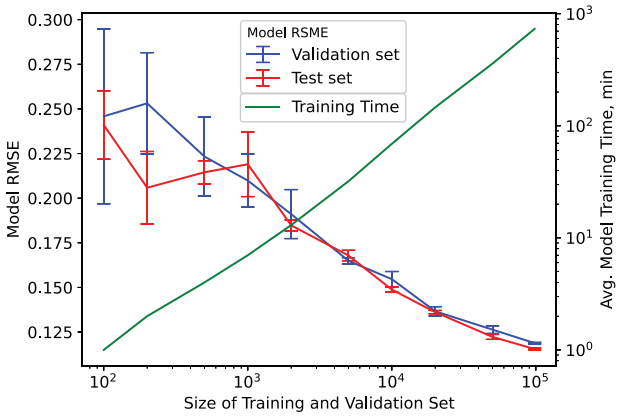

FIGURE 3 Learning curve of the CNN model trained on subsets of the Tabor_all data set. The RMSE is the average of the five cross-validation models, evaluated against the associated validation set (blue solid line) and the test set of 5000 quinones (red solid line) that is identical for all models. Error bars represent standard deviations. Training time is the average time required on a single NVIDIA GeForce GTX 1080 Ti to train the model of a single cross-validation fold.

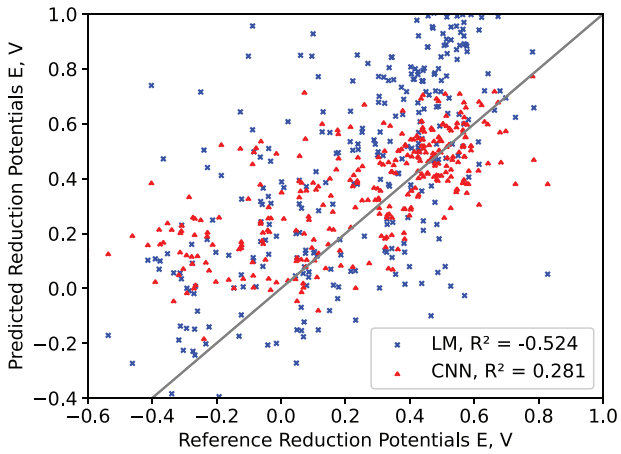

FIGURE 4 The CNN model (red) and LM (LM_TopoFP_Tabor_all, blue) are applied on the quinones in the Kristensen data set to predict their RP. RMSE is 0.267 V for the CNN model and 0.389 V for the LM.

ically sensible range (RMSE > 1×10^9. The reason here is likely that some functional groups or even chemical elements in the Kristensen data set are not present in the Tabor data set. The LM_TopoFP model does not predict RP far outside realistic ranges, but still performs worse than the CNN, resulting in a negative R^2 of -0.524 for the Tabor_all data set and -0.323 for Tabor_nosulf.

Both CNN and LMs appear not only less accurate, but also biased, with predictions generally overestimating the RP compared to the reference data. Notably, for the CNN model, the bias mainly applies to the quinone species with low reference RP values (<0.2 V), whereas for higher RP values, no significant bias is found. This might be due to

the very low number of quinones with a small or negative RP in the Tabor data set. From molecules that are contained in both Tabor_all and Kristensen data ($N = 10$), we find an average difference of 0.092 V for RP(Tabor)-RP(Kristensen).

Note that the developed models are subject to the bias and limitations of the quantum chemical methods used to estimate RP.[69,70] The bias toward a higher RP in the Tabor data set may thus be in part a result of systematic differences in the quantum-mechanical calculations underlying the two data sets. While the comparison between both data sets suggests relative biases, accurate experimental measurements are needed for quantification of the absolute bias of the underlying data.

The overall poor performance of the LM indicates that they are mostly incapable of extrapolation to structural motifs that are not represented in the training data; however, topological fingerprints (TopoFP) appear to be the molecular representation most suited for this multilinear regression task. The CNN models, however, are generally associated with a much lower prediction error in comparison with the LM. Extrapolation to larger and more complex structures works to some extent, but the absence of some functional groups in the training data set still has a significant effect on their performance.

Adaptation of CNN to literature data with solvent effects

Another CNN model is trained on the aggregate of Tabor_all and Kristensen data sets (CNN_Tabor_Kristensen) for the purpose of comparing the model output with experimental data of the RP of 11 quinones obtained from literature (Table 3).[32,71–82] These quinone species have been found in atmospheric samples and may be relevant with regard to public health or atmospheric chemistry.

We ensure that none of these quinones are included in the training and validation data by removing six of them. Average validation RMSE of the trained models from the five cross-validation folds is 0.118 V with an error variance of 5.23×10^{-7} V . The average RMSE for the test set of 5000 quinones is slightly lower at 0.115 V (error variance: 7.00×10^{-7} V), indicating no overfitting. From these five models, a model that achieved a validation RMSE of 0.118 V and a test set RMSE of 0.116 V (Table 2) is selected for further calculations.

The Tabor and Kristensen data sets use RP obtained from quantum-mechanical calculations and thus refer to gas-phase RP without solvent effects, whereas the experimental data are derived from electrochemical measurements using two different solvent/electrode combinations. In the first group (blue markers in Figure 5), water served as solvent and a normal hydrogen electrode (NHE) as a reference electrode. In the second group (red markers), dimethylformamide (DMF) was used as a solvent and a saturated calomel electrode (SCE) as a reference electrode. To compare the CNN-predicted RP to the measurements from two different chemical systems, CNN predictions are linearly shifted using measurements and predictions of 1,4-NQ. This species is well studied as indicated by the large number of experiments in this literature compilation (eight measurements in water/NHE, three

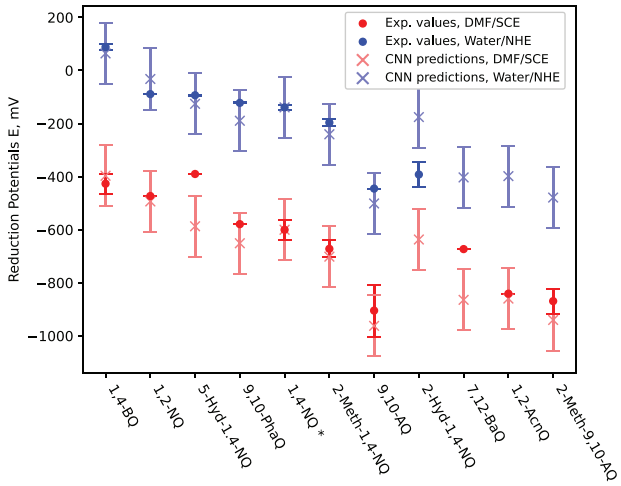

FIGURE 5 Comparison of CNN predictions and experimentally determined RP values for a selection of quinone species in two different solvent/electrode-systems. Experimental values are collected from literature and averaged. Solvent effects are considered by shifting CNN predictions so that the experimental data for 1,4-NQ are reproduced in both solvent/electrode systems. Quinone species, abbreviations, and associated values are presented in Table 3. Error bars of experimental values represent standard deviations from multiple measurements. Error bars of CNN predictions represent RMSE of the test set predictions. Correlation coefficients (r) for the two groups of predictions and experimental values are 0.858 for the DMF/SCE system and 0.762 for the water/NHE system.

in DMF/SCE). Linear relation of the RP of compounds in different solvents is common, but can be inaccurate for certain compound groups, for example, hydroxylated quinones.[32,76] The resulting offsets from the original CNN predictions are -0.576 V for the water/NHE system and -1.037 V for the DMF/SCE electrode system. The difference of 0.461 V for these offsets is consistent with values found in the literature of 0.473 V[76] and 0.6 V[32] between the RP in the two different solvent/electrode systems.

In Figure 5, predictions of the selected CNN model are displayed in comparison with experimental data. Apart from few outliers, RP experimental values are well captured by the CNN predictions (Figure 5), independent of the solvent/electrode system. Larger deviations in one solvent/electrode system compared to the other may point toward a shortcoming of the linear shift method for a specific substance. While this method is commonly used to transfer computationally derived RP to experimental measurements,[83,84] the true shift value depends on the specific solvation free energy of the molecule.[85] This may apply to 5-Hyd-1,4-NQ, where the associated CNN prediction accurately matches the experimental value obtained in the water/NHE system, but not the one shifted to fit the DMF/SCE system. Predictions for 2-Hyd-1,4-NQ and 7,12-BaQ, which have only been measured in one solvent/electrode system, are coincidentally associated with the largest deviation. Large differences between RP measurements in different solvent/electrode systems are often associated with quinone species containing prototropic functions such as hydroxyl groups,[76] which is the case for 2-Hyd-1,4-NQ and 5-Hyd-1,4-NQ. Prototropic

TABLE 3 Comparison of the CNN model predictions of the RP of quinones with literature values. The CNN predictions are linearly shifted using measurements and predictions of 1,4-naphthoquinone, resulting in one set matching the DMF/SCE system (-1037.1 mV) and one matching measurements of the water/NHE system (-575.8 mV).

Compound	Abbreviation	Literature average DMF/SCE, V	CNN prediction DMF/SCE, V
1,4-Benzoquinone	1,4-BQ	−0.427	−0.397
1,2-Naphthoquinone	1,2-NQ	−0.473	−0.493
5-Hydroxy-1,4-naphthoquinone	5-Hyd-1,4-NQ	−0.39	−0.587
9,10-Phenanthrenequinone	9,10-PhaQ	−0.579	−0.651
1,4-Naphthoquinone	1,4-NQ	−0.6	−0.6
2-Methyl-1,4-naphthoquinone	2-Meth-1,4-NQ	−0.672	−0.702
9,10-Anthraquinone	9,10-AQ	−0.673	−0.962
Benz(a)anthracene-7,12-dione	7,12-BaQ	−0.841	−0.864
1,2-Acenaphthenequinone	1,2-AcnQ	−0.869	−0.856
2-Methyl-9,10-Anthraquinone	2-Meth-9,10-AQ	−0.905	−0.940
		Literature average H_2O/NHE, V	**CNN prediction H_2O/NHE, V**
1,4-Benzoquinone	1,4-BQ	0.089	0.064
1,2-Naphthoquinone	1,2-NQ	−0.089	−0.032
5-Hydroxy-1,4-naphthoquinone	5-Hyd-1,4-NQ	−0.093	−0.126
9,10-Phenanthrenequinone	9,10-PhaQ	−0.122	−0.19
1,4-Naphthoquinone	1,4-NQ	−0.139	−0.139
2-Methyl-1,4-naphthoquinone	2-Meth-1,4-NQ	−0.196	−0.240
9,10-Anthraquinone	9,10-AQ	−0.392	−0.501
2-Hydroxy-1,4-naphthoquinone	2-Hyd-1,4-NQ	−0.445	−0.176

groups can, depending on the solvent, influence the stability of the quinone and the semiquinone due to tautomeric effects and internal hydrogen bonds. Similarly, their RP can be strongly pH-dependent.

RP predictions for compounds detected in atmospheric samples or as metabolites in living systems

For many quinones that are commonly detected in the atmosphere or as metabolites in living systems, experimental measurements of RP are not available. To estimate the potential impact of these substances on atmospheric chemistry or public health, we apply our model CNN_Tabor_Kristensen for target-oriented screening of quinone structures reported in the literature. Using Equation (3), as presented by Roginsky et al.,[32] we translate the RP predictions into the effective kinetic rate constant of oxygen consumption in an ascorbate assay, k_{eff}, which may be closely related to their OP[26] and is used as proxy for OP in this study.

$$\log(k_{eff}) = (3.91 \pm 90.07) + (0.01439 \pm 0.0004) \times \mathrm{RP}. \quad (3)$$

Roginsky et al.[32] found that quinone species with an RP between −250 mV and +50 mV are able to perform catalytic redox cycling in their experimental system mimicking biological fluids. Above 50 mV, however, the RP becomes too large to allow for backreaction of the semiquinone (Equation (2)). Hence, Equation (3) applies only to an RP range of −250 to 50 mV in the water/NHE system.

An overview over RP predictions and estimates for the OP proxy k_{eff} of quinones detected in atmospheric samples (Table S2), found as metabolites in living systems (Table S3), and investigated regarding their association with lipid peroxidation[86] (Table S4), is given in the Supporting Information.

Quinone species from atmospheric samples in Table S2 that correspond to high k_{eff} include well-known redox cyclers such as 1,2-naphthoquinone, 1,4-naphthoquinone, 9,10-phenanthrenequinone,[29,87–89] 5-hydroxy-1,4-naphtoquinone[30,90] 4,5-pyrenequinone, and 1,6-pyrenequinone.[91] Furthermore, we find a large estimated OP for the following substances: 5-methyl-1,2-chrysenequinone, benzo[a]anthracene-3,4-dione, benzo[c]phenanthrene-1,4-dione, 1,2-triphenylenequinone, 2-hydroxy-1,4-benzoquinone, 1,4-chrysenequinone, 2,6-dimethyl-1,4-benzoquinone, tetramethyl-1,4-benzoquinone, benzo[ghi]perylene-7,8-dione, and dibenzo[b,n]perylene-15,16-dione. These compounds may be of interest for future research regarding their adverse health effects. Note that the error in the estimation of the RP is fairly large at 116 mV. Hence, some compounds in this compilation may have an RP above 50 mV and thus not be redox cyclers in solutions containing ascorbate.

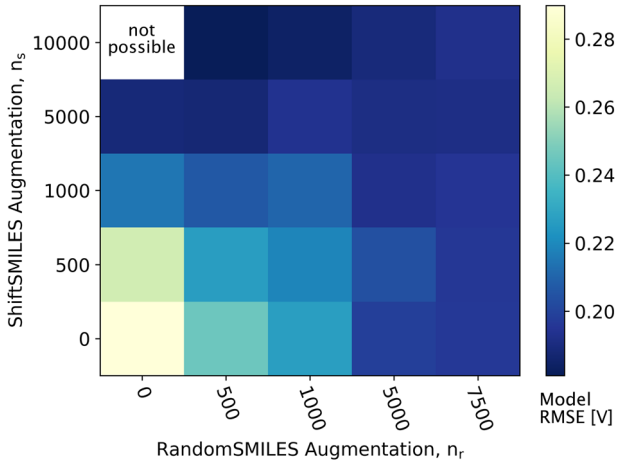

FIGURE 6 Heat map of test set RMSE of CNN models as a function of the extent of data augmentation. The RMSE values are average values of the 25 models trained under double cross-validation with 5 ×5 folds. CNN models are trained and tested on data in the Kristensen data set of 321 quinones. A specified number of synthetic samples are obtained by application of the two augmentation methods RandomSMILES and ShiftSMILES on the training data. No model is trained for $n_r = 0$ and $n_s = 10{,}000$, as there are not enough samples in the training set to generate 10,000 ShiftSMILES without any RandomSMILES enhancement. The model with the largest error in this experiment is the one trained only on the original data. Any addition of augmented data leads to a reduction of error. Overall, best model performance enhancements are achieved with the addition of more than 5000 augmentations with a large n_s/n_r ratio.

Likewise, compounds with an estimated RP above 50 mV, not explicitly mentioned here, may have a high OP.

Effect of data augmentation methods

CNN performs best with large amounts of training data. Here, we test whether augmentation methods can be used to increase the performance of a CNN model when only limited data are available. For this purpose, CNN models are trained and tested on the Kristensen data of 321 quinones, which is extended by increasing numbers of augmented data. Figure 6 shows a 5× 5 grid of the trained model RMSE as a function of ShiftSMILES and RandomSMILES augmentation samples added to the training data.

The average validation error of the 600 individual models trained in this experiment is 0.202 V, which is very similar to 0.209 V for the average test set error. This is an indicator of little to no overfitting despite the very low training set size, probably a result of decreased model complexity (Section 2.4).

We find that any degree of data augmentation improves the model performance compared to the original model, that is, the model trained only on the original data (206 quinone species in the training set, 51 in the validation set) achieves the lowest ability in predicting the test set (RMSE = 0.290 V). The overall best performing models are augmented by $n_r = 500$ and $n_s = 10{,}000$ (RMSE = 0.181 V), decreasing the average model error in association with the original data set by 37.5%. Notably, the combination of both augmentation methods outperforms their separate application, whereas large n_s/n_r ratios appear preferable. Note that each method individually only allows the computation of less than 8000 augmentations. Given their consecutive application, both methods theoretically allow the generation of data sets in the order of magnitude 10^5 from the Kristensen data.

Enhancement of training data sets for CNN models with data augmentation facilitates regularization and avoids overfitting.[92] Different SMILES representations of the same molecule may offer more possibilities for the model to learn how subpatterns in the molecule interact, leading to increased generalization of the model. This can compensate or trump a potential initial performance decrease that could result from the relaxation of underlying rules (e.g., the set order of functional groups in canonical SMILES quickly indicating compound class or the number of empty columns at the end of the input matrix indicating molecule size). At very large numbers of augmentations, model bias may counteract increases in the performance, as described in Supporting Information Note 2. In short, larger molecules allow more RandomSMILES and less ShiftSMILES augmentations, introducing bias in the randomized selection of augmented molecules. Therefore, we suggest to use similar numbers of augmentations per molecule for each method. As ShiftSMILES is applied on both the original data and the RandomSMILES augmentations, we suggest $n_s > n_r$.

CONCLUSIONS AND OUTLOOK

In this study, a machine learning approach for the prediction of quinone RP values using CNN is presented and applied. Depending on the availability of data, this method can similarly be applied to other chemical species and properties. The principle of encoding molecular structures in canonical SMILES, one-hot encoding, and application of a CNN had previously been applied successfully in classification tasks.[48] The findings in this study underline the feasibility of this method and extend it to regression tasks, where a continuous variable is predicted. In our experiment, the novel CNN approach significantly outperformed multiple linear regression models fitted to common molecular descriptors in terms of prediction error. Figure 7 shows how the CNN can be used in a target-oriented screening approach to determine the relevance of quinones for atmospheric chemistry or human health. Together with a parameterization translating RP into OP, a number of quinone structures with unknown chemical properties are screened for their ability to act as redox cyclers in the presence of molecular oxygen and the antioxidant ascorbic acid.

The representation of molecular structures plays an important role for the performance of a predictive model and there are many alternatives to the SMILES and one-hot-encoding approach in this study. For example, CNN,[93] Deep Tensor Neural Networks,[94] or molecular embedding[95] can be used as pretraining methods to create information-rich representations of molecules. We suggest that, in future research, a comparison (and potentially combination) of methods for feature extraction of molecular structures should be

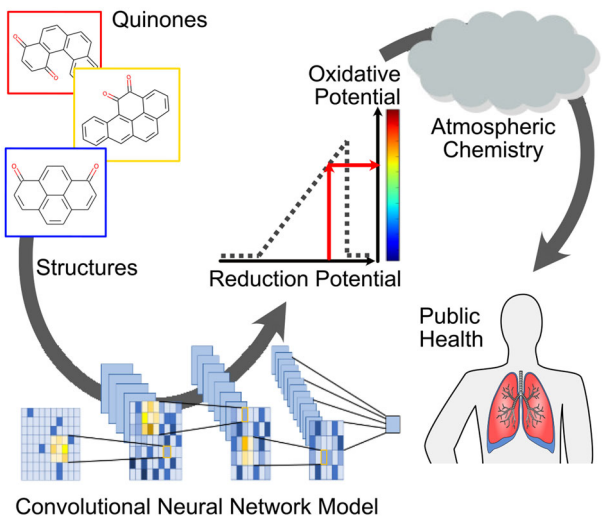

FIGURE 7 Application of the CNN model in target-oriented screening of quinones for relevance in atmospheric chemistry or human health. The CNN model predicts the RP based on molecular structures of quinones with unknown properties. By transforming RP predictions to OP proxies, a set of potential redox cyclers can be obtained to direct further investigation.

performed to achieve optimal performance for a given model system. For traditional cheminformatics representations, this was done by Lumiaro et al.[96] who predicted gas-particle partitioning coefficients of atmospheric molecules using kernel ridge regression and tested several methods for the encoding of molecular structure. Of the tested molecule representations, many-body tensors and topological fingerprints performed best in their application.

For SMILES and one-hot encoding as molecule representation, two augmentation methods are presented: RandomSMILES, which was successfully applied by Arús-Pous et al.[55] and Bjerrum[54] to increase model accuracy in machine learning applications based on SMILES-encoded inputs. And ShiftSMILES, which is applied not on a semantic level of SMILES-logic like RandomSMILES, but on a data-structural level, inspired by common applications of augmentation in images. In a broad experiment based on a small data set of 321 quinone species, application of both augmentation methods reduced test data errors of the associated CNN models. The use of such augmentation methods possibly allows the feasible training of similar models on data sets that would otherwise be too small to result in sufficient model accuracy.

Transferability of purely data-centric machine learning models across data sets with different distributions of molecular substructures, like functional groups, remains an issue. In Figure 4, prediction errors of the transferred CNN are much larger compared to the test set that originated from the same data set (Figure 2). This is in part because molecules in both data sets differ structurally and some functional groups are newly introduced. The problem, however, is not inherent to CNN models but applies to many, if not all models that are fitted or trained on data. For instance, the LM performed much worse in the same experiment.

As a potential remedy, the augmentation methods could be applied to only a fraction of the training data, for instance, molecules containing a specific functional group or associated with a specific range of the target property. In certain applications, this intentional modification of model bias could improve model performance for certain chemical compound classes or for a specific range of the target property. For example, if the model were applied to estimate the ability of specific quinones to induce oxidative stress in the lung-lining fluid of the human body,[19,21] one could apply the augmentation methods only on those compounds in the data set whose RP falls within the range associated with efficient redox cycling in the human body, specializing the model for applications in physiological chemistry.

When training and applying a CNN model, we usually have little insight into the underlying mechanism that leads to successful mapping of patterns in the input data to the output. The functionality of a CNN, however, allows an investigation of intermediate data of the trained model, linking the encoded molecular structure to the predicted property of interest. These feature maps, as introduced in Equation (S4), are information passed from a convolutional layer to the following layer. In a trained and successful model, they contain information regarding the data substructures that are of relevance for the CNN model with regard to the predicted value. For the purpose of visualization, we present a method in Supporting Information Note 5 that maps areas on the feature map to the original molecular structure, and thus, shows how substructures are parsed through the CNN layers and affect the model prediction. Such activation scheme patterns may be linked to a substructure's effect on the model prediction, the model's bias toward substructures or molecule classes, or the overall predictive accuracy for individual molecules. In our preliminary testing of applying basic mathematical operations on the activation schemes of the three convolutional layers, we found weak but significant correlations supporting these hypotheses. Establishing a systematic methodology for such a process is beyond the scope of this study and will be the subject of future work.

ACKNOWLEDGMENTS

We thank Daniel P. Tabor for providing training data in tabulated form. Parts of this research were conducted using the supercomputer Mogon and/or advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate, www.ahrp.info) and the Gauss Alliance e.V. The authors gratefully acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (hpc.uni-mainz.de). We would like to thank two anonymous reviewers for their helpful contributions during peer review.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ETHICS STATEMENT

The authors confirm that they have followed the ethical policies of the journal.

AUTHOR CONTRIBUTIONS

MK, JW, BS, and TB designed research. MK wrote the model code and performed the simulations. BAMB and MW provided data regarding atmospherically-relevant quinones and literature reduction potential values. All authors discussed and interpreted calculation results. MK and TB prepared the manuscript with contributions from all coauthors.

DATA AVAILABILITY STATEMENT

All data used in this study as well as the source code are available on gitlab. https://gitlab.mpcdf.mpg.de/mkruege/quin_redpot_cnn

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/ntls.20220016

ORCID

Matteo Krüger https://orcid.org/0000-0003-0191-2637
Jake Wilson https://orcid.org/0000-0002-2342-6966
Marco Wietzoreck https://orcid.org/0000-0001-9007-3067
Benjamin A. Musa Bandowe https://orcid.org/0000-0003-0605-2285
Gerhard Lammel https://orcid.org/0000-0003-2313-0628
Bertil Schmidt https://orcid.org/0000-0003-2597-8331
Ulrich Pöschl https://orcid.org/0000-0003-1412-3557
Thomas Berkemeier https://orcid.org/0000-0001-6390-6465

REFERENCES

1. Mazzei F, D'Alessandro A, Lucarelli F, et al. Characterization of particulate matter sources in an urban environment. *Sci Total Environ*. 2008;401(1-3):81-89. https://doi.org/10.1016/j.scitotenv.2008.03.008
2. Mukherjee A, Agrawal M. World air particulate matter: sources, distribution and health effects. *Environ Chem Lett*. 2017;15(2):283-309. https://doi.org/10.1007/s10311-017-0611-9
3. Shah AS, Langrish JP, Nair H, et al. Global association of air pollution and heart failure: a systematic review and meta-analysis. *Lancet*. 2013;382(9897):1039-1048. https://doi.org/10.1016/S0140-6736(13)60898-3
4. Liu L, Poon R, Chen L, et al. Acute effects of air pollution on pulmonary function, airway inflammation, and oxidative stress in asthmatic children. *Environ Health Perspect*. 2009;117(4):668-674. https://doi.org/10.1289/ehp11813
5. Sint T, Donohue JF, Ghio AJ. Ambient air pollution particles and the acute exacerbation of chronic obstructive pulmonary disease. *Inhal Toxicol*. 2008;20(1):25-29. https://doi.org/10.1080/08958370701758759
6. Cohen A, Pope CA. Lung cancer and air pollution. *Lancet*. 1978;311(8078):1366. https://doi.org/10.1016/S0140-6736(78)92444-3
7. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. *Lancet*. 2014;383(9928):1581-1592. https://doi.org/10.1016/S0140-6736(14)60617-6
8. Franklin BA, Brook R, Pope AC. Air pollution and cardiovascular disease. *Curr Probl Cardiol*. 2015;40(5):207-238. Air Pollution and Cardiovascular Disease https://doi.org/10.1016/j.cpcardiol.2015.01.003
9. Li N, Hao M, Phalen RF, Hinds WC, Nel AE. Particulate air pollutants and asthma: a paradigm for the role of oxidative stress in PM-induced adverse health effects. *Clin Immunol*. 2003;109(3):250-265. https://doi.org/10.1016/j.clim.2003.08.006
10. Shiraiwa M, Selzle K, Pöschl U. Hazardous components and health effects of atmospheric aerosol particles: reactive oxygen species, soot, polycyclic aromatic compounds and allergenic proteins. *Free Radic Res*. 2012;46(8):927-939. https://doi.org/10.3109/10715762.2012.663084
11. Pöschl U, Shiraiwa M. Multiphase chemistry at the atmosphere-biosphere interface influencing climate and public health in the Anthropocene. *Chem Rev*. 2015;115(10):4440-4475. PMID: 25856774 https://doi.org/10.1021/cr500487s
12. Ayres JG, Borm P, Cassee FR, et al. Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential - a workshop report and consensus statement. *Inhal Toxicol*. 2008;20(1):75-99. https://doi.org/10.1080/08958370701665517
13. Lodovici M, Bigagli E. Oxidative stress and air pollution exposure. *Int J Toxicol*. 2011;2011. https://doi.org/10.1155/2011/487074
14. Kelly FJ, Fussell JC. Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter. *Atmos Environ*. 2012;60:504-526. https://doi.org/10.1016/j.atmosenv.2012.06.039
15. Gehling W, Dellinger B. Environmentally persistent free radicals and their lifetimes in PM2. 5. *Environ Sci Technol*. 2013;47(15):8172-8178.
16. Borrowman CK, Zhou S, Burrow TE, Abbatt JP. Formation of environmentally persistent free radicals from the heterogeneous reaction of ozone and polycyclic aromatic compounds. *Phys Chem Chem Phys*. 2016;18(1):205-212.
17. Vejerano EP, Rao G, Khachatryan L, Cormier SA, Lomnicki S. Environmentally persistent free radicals: insights on a new class of pollutants. *Environ Sci Technol*. 2018;52(5):2468-2481. https://doi.org/10.1021/acs.est.7b04439
18. Charrier JG, McFall AS, Richards-Henderson NK, Anastasio C. Hydrogen peroxide formation in a surrogate lung fluid by transition metals and quinones present in particulate matter. *Environ Sci Technol*. 2014;48(12):7010-7017. https://doi.org/10.1021/es501011w
19. Lakey PS, Berkemeier T, Tong H, et al. Chemical exposure-response relationship between air pollutants and reactive oxygen species in the human respiratory tract. *Sci Rep*. 2016;6:1-6. https://doi.org/10.1038/srep32916
20. Fang T, Lakey PS, Weber RJ, Shiraiwa M. Oxidative potential of particulate matter and generation of reactive oxygen species in epithelial lining fluid. *Environ Sci Technol*. 2019;53(21):12784-12792. https://doi.org/10.1021/acs.est.9b03823
21. Lelieveld S, Wilson J, Dovrou E, et al. Hydroxyl radical production by air pollutants in epithelial lining fluid governed by interconversion and scavenging of reactive oxygen species. *Environ Sci Technol*. 2021;55(20):14069-14079. https://doi.org/10.1021/acs.est.1c03875
22. Gao D, Fang T, Verma V, Zeng L, Weber RJ. A method for measuring total aerosol oxidative potential (OP) with the dithiothreitol (DTT) assay and comparisons between an urban and roadside site of water-soluble and total OP. *Atmos Meas Tech*. 2017;10(8):2821-2835. https://doi.org/10.5194/amt-10-2821-2017
23. Jiang H, Sabbir Ahmed CM, Canchola A, Chen JY, Lin YH. Use of dithiothreitol assay to evaluate the oxidative potential of atmospheric aerosols. *Atmosphere*. 2019;10(10):1-21. https://doi.org/10.3390/atmos10100571
24. Ng NL, Tuet WY, Chen Y, et al. Cellular and acellular assays for measuring oxidative stress induced by ambient and laboratory-generated aerosols. *Res Rep Health Eff Inst*. 2019;197:1-57.
25. Pietrogrande MC, Bertoli I, Manarini F, Russo M. Ascorbate assay as a measure of oxidative potential for ambient particles: evidence for the importance of cell-free surrogate lung fluid composition. *Atmos Environ*. 2019;211:103-112.
26. Wietzoreck M, Filippi A, Wilson J, et al. Oxidative potential of polycyclic aromatic compounds (PACs) in particulate matter: measurement

and prediction by chemical structure and reduction potential. In preparation. 2022.
27. Monks TJ, Hanzlik RP, Cohen GM, Ross D, Graham DG. Quinone chemistry and toxicity. *Toxicol Appl Pharmacol*. 1992;112(1):2-16. https://doi.org/10.1016/0041-008X(92)90273-U
28. Bolton JL, Trush MA, Penning TM, Dryhurst G, Monks TJ. Role of quinones in toxicology. *Chem Res Toxicol*. 2000;13(3):135-160.
29. Charrier J, Anastasio C. On dithiothreitol (DTT) as a measure of oxidative potential for ambient particles: evidence for the importance of soluble transition metals. *Atmos Chem Phys*. 2012;257(12):9321-9333. https://doi.org/10.5194/acpd-12-11317-2012.On
30. Verma V, Wang Y, El-Afifi R, et al. Fractionating ambient humic-like substances (HULIS) for their reactive oxygen species activity - assessing the importance of quinones and atmospheric aging. *Atmos Environ*. 2015;120:351-359. https://doi.org/10.1016/j.atmosenv.2015.09.010
31. Gao D, Ripley S, Weichenthal S, Godri PKJ. Ambient particulate matter oxidative potential: chemical determinants, associated health effects, and strategies for risk management. *Free Radic Biol Med*. 2020;151:7-25. https://doi.org/10.1016/j.freeradbiomed.2020.04.028
32. Roginsky VA, Barsukova TK, Stegmann HB. Kinetics of redox interaction between substituted quinones and ascorbate under aerobic conditions. *Chem Biol Interact*. 1999;121(2):177-197. https://doi.org/10.1016/S0009-2797(99)00099-X
33. Testa B, Pedretti A, Vistoli G. Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Discov Today*. 2012;17(11-12):549-560.
34. Huskinson B, Marshak MP, Suh C, et al. A metal-free organic-inorganic aqueous flow battery. *Nature*. 2014;505(7482):195-198.
35. Schwan S, Schröder D, Wegner H, Janek J, Mollenhauer D. Substituent pattern effects on the redox potentials of quinone-based active materials for aqueous redox flow batteries. *ChemSusChem*. 2020;13(20):5480-5488.
36. Kissinger PT, Heineman WR. Cyclic voltammetry. *J Chem Educ*. 1983;60(9):702-706. https://doi.org/10.1021/ed060p702
37. Geerlings P, De Proft F, Langenaeker W. Conceptual density functional theory. *Chem Rev*. 2003;103(5):1793-1873. https://doi.org/10.1021/cr990029p
38. Magnussen T, Rasmussen P, Fredenslund A. Unifac parameter table for prediction of liquid-liquid equilibria. *Ind Eng Chem Process Des Dev*. 1981;20(2):331-339. https://doi.org/10.1021/i200013a024
39. Zuend A, Marcolli C, Luo BP, Peter T. A thermodynamic model of mixed organic-inorganic aerosols to predict activity coefficients. *Atmos Chem Phys*. 2008;8(16):4559-4593.
40. Compernolle S, Ceulemans K, Müller JF. Evaporation: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions. *Atmospheric Chem Phys*. 2011;11(18):9431-9450. https://doi.org/10.5194/acp-11-9431-2011
41. Fredenslund A. *Vapor-Liquid Equilibria Using UNIFAC: A Group-Contribution Method*. Elsevier; 2012.
42. LeCun Y, Bengio Y, Hinton, G. Deep learning. *Nature* 2015;521:436-444. https://doi.org/10.1038/nature14539
43. Kröse B, Smagt v. dP. *An Introduction to Neural Networks*. Citeseer; 1993.
44. Heaton J. *Ian goodfellow, Yoshua Bengio, and Aaron Courville: Deep Learning*. Springer; 2018.
45. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem Inf Model*. 1988;28(1):31-36.
46. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Model*. 1989;29(2):97-101. https://doi.org/10.1021/ci00062a008
47. Potdar KSTDC. A comparative study of categorical variable encoding techniques for neural network classifiers. *Int J Comput*. 2017;175(4):7-9. https://doi.org/10.5120/ijca2017915495
48. Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinform*. 2018;19(Suppl 19). https://doi.org/10.1186/s12859-018-2523-5
49. Sanchez-Lengeling B, Outeiral C, Guimaraes GL, Aspuru-Guzik A. Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). 2017.
50. Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*. 2017.
51. Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In: *IIPhDW 2018*. 2018:117-122. https://doi.org/10.1109/IIPHDW.2018.8388338
52. Mosolova AV, Fomin VV, Bondarenko IY. Text augmentation for neural networks. *CEUR Workshop Proc*. 2018;2268:104-109.
53. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*. 2020:6382-6388. https://doi.org/10.18653/v1/d19-1670
54. Bjerrum EJ. SMILES enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076*. 2017.
55. Arús-Pous J, Johansson SV, Prykhodko O, et al. Randomized SMILES strings improve the quality of molecular generative models. *J Cheminformatics*. 2019;11(1):1-13. https://doi.org/10.1186/s13321-019-0393-0
56. Tabor DP, Gómez-Bombarelli R, Tong L, Gordon RG, Aziz MJ, Aspuru-Guzik A. Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries. *J Mater Chem A*. 2019;7(20):12833-12841. https://doi.org/10.1039/c9ta03219c
57. Kristensen SB, Mourik vT, Pedersen TB, Sørensen JL, Muff J. Simulation of electrochemical properties of naturally occurring quinones. *Sci Rep*. 2020;10(1). https://doi.org/10.1038/s41598-020-70522-z
58. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Model*. 2002;42(6):1273-1280.
59. Landrum G. RDKit: open-source cheminformatics software. https://www.rdkit.org. 2010.
60. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J Chem doc*. 1965;5(2):107-113.
61. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Series B Stat Methodol*. 1974;36(2):111-133.
62. Wong TT, Yeh PY. Reliable accuracy estimates from <italic>k</italic>-fold cross validation. *IEEE Trans Knowl Data Eng*. 2020;32(8):1586-1594. https://doi.org/10.1109/TKDE.2019.2912815
63. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.
64. Mosteller F, Tukey JW. Data analysis, including statistics. *Handb Soc Psychol*. 1968;2:80-203.
65. Baumann D, Baumann K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform*. 2014;6(1):1-19.
66. O'Shea K, Nash R. An Introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*. 2015:1-11.
67. Gulli A, Pal S. *Deep Learning with Keras*. Packt Publishing Ltd.; 2017.
68. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK. Validation of QSAR models-strategies and importance. *Int J Drug Des Discov*. 2011;3:511-519.
69. Wodrich MD, Corminboeuf C, Schleyer PvR. Systematic errors in computed alkane energies using B3LYP and other popular DFT functionals. *Org Lett*. 2006;8(17):3631-3634.

70. Csonka GI, French AD, Johnson GP, Stortz CA. Evaluation of density functionals and basis sets for carbohydrates. *J Chem Theory Comput*. 2009;5(4):679-692.
71. Ilan YA, Czapski G, Meisel D. The one-electron transfer redox potentials of free radicals. I. The oxygen/superoxide system. *Biochim BioPhys Acta Biomembr*. 1976;430(2):209-224. https://doi.org/10.1016/0005-2728(76)90080-3
72. Prince RC, Leslie Dutton P, Malcolm Bruce J. Electrochemistry of ubiquinones. *FEBS Lett*. 1983;160(1-2):273-276. https://doi.org/10.1016/0014-5793(83)80981-8
73. Mukherjee T. One-electron reduction of juglone (5-hydroxy-1,4-naphthoquinone): a pulse radiolysis study. *Int J Radiat Appl Instrum Part C*. 1987;29(6):455-462. https://doi.org/10.1016/1359-0197(87)90024-5
74. Wardman P. Reduction potentials of one electron couples involving free radicals in aqueous solution. *J Phys Chem Ref Data*. 1989;18(4):1637-1755. https://doi.org/10.1063/1.555843
75. Öllinger K, Buffinton GD, Ernster L, Cadenas E. Effect of superoxide dismutase on the autoxidation of substituted hydro- and semi-naphthoquinones. *Chem Biol Interact*. 1990;73(1):53-76. https://doi.org/10.1016/0009-2797(90)90108-Y
76. Wardman P. Bioreductive activation of quinones: redox properties and thiol reactivity. *Free Radic Res*. 1990;8(4-6):219-229. https://doi.org/10.3109/10715769009053355
77. Bironaite DA, Čenas NK, Kulys JJ. The rotenone-insensitive reduction of quinones and nitrocompounds by mitochondrial NADH:ubiquinone reductase. *Biochim BioPhys Acta - Bioenerg*. 1991;1060(2):203-209. https://doi.org/10.1016/S0005-2728(09)91008-8
78. O'Brien P. Molecular mechanisms of quinone cytotoxicity. *Chem Biol Interact*. 1991;80(1):1-41. https://doi.org/10.1016/0009-2797(91)90029-7
79. Cenas N, Anusevicius Z, Bironaite D, Bachmanova G, Archakov A, Ollinger K. The electron-transfer reactions of NADPH-cytochrome P450 reductase with nonphysiological oxidants. *Arch BioChem Biophys*. 1994;315(2):400-406. https://doi.org/10.1006/abbi.1994.1517
80. Livertoux MH, Lagrange P, Minn A. The superoxide production mediated by the redox cycling of xenobiotics in rat brain microsomes is dependent on their reduction potential. *Brain Res*. 1996;725(2):207-216. https://doi.org/10.1016/0006-8993(96)00251-X
81. Rath M, Pal H, Mukherjee T. Pulse-radiolytic one-electron reduction of anthraquinone and chloro-anthraquinones in aqueous-isopropanol-acetone mixed solvent. *Radiat Phys Chem*. 1996;47(2):221-227. https://doi.org/10.1016/0969-806X(95)00003-G
82. Trumpower B. *Function of Quinones in Energy Conserving Systems*. Elsevier; 2012.
83. Er S, Suh C, Marshak MP, Aspuru-Guzik A. Computational design of molecules for an all-quinone redox flow battery. *Chem Sci*. 2015;6(2):885-893.
84. Huynh MT, Anson CW, Cavell AC, Stahl SS, Hammes-Schiffer S. Quinone 1 e–and 2 e–/2 H+ reduction potentials: identification and analysis of deviations from systematic scaling relationships. *J Am Chem Soc*. 2016;138(49):15903-15910.
85. Hruska E, Gale A, Liu F. Bridging the experiment-calculation divide: machine learning corrections to redox potential calculations in implicit and explicit solvent models. *J Chem Theory Comput*. 2022;18:1096-1108.
86. Zhao Y, Xia Q, Yin JJ, Yu H, Fu PP. Photoirradiation of polycyclic aromatic hydrocarbon diones by UVA light leading to lipid peroxidation. *Chemosphere*. 2011;85(1):83-91.
87. Chung MY, Lazaro RA, Lim D, et al. Aerosol-borne quinones and reactive oxygen species generation by particulate matter extracts. *Environ Sci Technol*. 2006;40(16):4880-4886.
88. Charrier JG, Anastasio C. Rates of hydroxyl radical production from transition metals and quinones in a surrogate lung fluid. *Environ Sci Technol*. 2015;49(15):9317-9325.
89. Visentin M, Pagnoni A, Sarti E, Pietrogrande MC. Urban PM2. 5 oxidative potential: importance of chemical species and comparison of two spectrophotometric cell-free assays. *Environ Pollut*. 2016;219:72-79.
90. Xiong Q, Yu H, Wang R, Wei J, Verma V. Rethinking dithiothreitol-based particulate matter oxidative potential: measuring dithiothreitol consumption versus reactive oxygen species generation. *Environ Sci Technol*. 2017;51(11):6507-6514.
91. Okubo R, Kameda T, Tohno S. Evaluation of oxidative potential of pyrenequinone isomers by the dithiothreitol (DTT) assay. *Polycycl Aromat Compd*. 2021;0(0):1-8.
92. Hernández-García A, König P. Further advantages of data augmentation on convolutional neural networks. In: *International Conference on Artificial Neural Networks*. Springer; 2018:95-103.
93. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF. Convolutional embedding of attributed molecular graphs for physical property prediction. *J Chem Inf Model*. 2017;57(8):1757-1772.
94. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. Schnet-a deep learning architecture for molecules and materials. *J Chem Phys*. 2018;148(24):241722.
95. Galeazzo T, Shiraiwa M. Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings. *Environ Sci: Atmos*. 2022;2:362-374.
96. Lumiaro E, Todorović M, Kurten T, Vehkamäki H, Rinke P. Predicting gas-particle partitioning coefficients of atmospheric molecules with machine learning. *Atmos Chem Phys*. 2021;21(17):13227-13246. https://doi.org/10.5194/acp-21-13227-2021

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Krüger M, Wilson J, Wietzoreck M, et al. Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects. *Nat Sci*. 2022;2:e20220016. https://doi.org/10.1002/ntls.20220016

2.3. Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^{2}NN)

This chapter presents a research article published in the journal Geoscientific Model Development. I am the first author and the main contributor to this paper. I received the data and code to apply and evaluate a basic graph convolutional neural network (GCNN) model by Tommaso Galeazzo. In the following, I extended and refined the data set assisted by Ivan Erements, developed and implemented more advanced GCNN architectures including the group contribution module and adaptive-depth approach, and carried out model training and optimization using high performance computing. I extended the model evaluation and benchmarking, designed all figures, and wrote and revised the manuscript together with Thomas Berkemeier. More detailed information on the author contributions are provided at the end of the manuscript.

Krüger, M., Galeazzo, T., Eremets, I., Schmidt, B., Pöschl, U., Shiraiwa, M., Berkemeier, T.: Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^{2}NN), *Geosci. Model Dev.*, doi: 10.5194/gmd-18-7357-2025, (2025).

A novel approach was developed to predict the saturation vapor pressures of atmospherically relevant compounds using group contribution-assisted graph convolutional neural networks (GC^{2}NN). The models use molecular descriptors like molar mass alongside molecular graphs containing atom and bond features as representations of molecular structure. Best results were achieved with an adaptive-depth GC^{2}NN, where the number of evaluated graph layers depends on molecular size. The adaptive-depth GC^{2}NN models clearly outperformed existing methods, including parameterizations and group-contribution methods, demonstrating that graph-based ML techniques are powerful tools for the estimation of physicochemical properties, even when experimental data are scarce. The supplement to this work can be found in appendix B2.

Geosci. Model Dev., 18, 7357–7371, 2025
https://doi.org/10.5194/gmd-18-7357-2025

Geoscientific Model Development
Open Access
EGU

Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^{2}NN)

Matteo Krüger[1,★], Tommaso Galeazzo[2,★], Ivan Eremets[1], Bertil Schmidt[3], Ulrich Pöschl[1], Manabu Shiraiwa[2], and Thomas Berkemeier[1]

[1]Multiphase Chemistry Department, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany
[2]Department of Chemistry, University of California Irvine, Irvine, California, USA
[3]Department of Computer Science, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany
★These authors contributed equally to this work.

Correspondence: Manabu Shiraiwa (m.shiraiwa@uci.edu) and Thomas Berkemeier (t.berkemeier@mpic.de)

Received: 13 March 2025 – Discussion started: 20 March 2025
Revised: 15 July 2025 – Accepted: 25 August 2025 – Published: 15 October 2025

Abstract. The vapor pressures (p_{vap}) of organic molecules play a crucial role in the partitioning of secondary organic aerosol (SOA). Given the vast diversity of atmospheric organic compounds, experimentally determining p_{vap} of each compound is unfeasible. Machine Learning (ML) algorithms allow the prediction of physicochemical properties based on complex representations of molecular structure, but their performance crucially depends on the availability of sufficient training data. We propose a novel approach to predict p_{vap} using group contribution-assisted graph convolutional neural networks (GC^{2}NN). The models use molecular descriptors like molar mass alongside molecular graphs containing atom and bond features as representations of molecular structure. The model's group contribution component is a shallow fully-connected neural network which processes numerical molecular descriptors and complements the model's graph component. Molecular graphs allow the ML model to better infer molecular connectivity compared to methods using other, non-structural embeddings. We achieve best results with an adaptive-depth GC^{2}NN, where the number of evaluated graph layers depends on molecular size. We present two vapor pressure estimation models that achieve strong agreement between predicted and experimentally-determined p_{vap}. The first is a general model with broad scope that is suitable for both organic and inorganic molecules and achieves a mean absolute error (MAE) of 0.69 log-units ($R^2 = 0.86$). The second model is specialized on organic compounds with functional groups often encountered in atmospheric SOA, achieving an even stronger correlation with the test data (MAE = 0.37 log-units, $R^2 = 0.94$). The adaptive-depth GC^{2}NN models clearly outperform existing methods, including parameterizations and group-contribution methods, demonstrating that graph-based ML techniques are powerful tools for the estimation of physicochemical properties, even when experimental data are scarce.

1 Introduction

Secondary organic aerosols (SOA) account for a substantial mass fraction (20 %–90 %) of tropospheric aerosols (Jimenez et al., 2009). They affect the atmosphere's radiative budget and serve as nuclei in cloud droplet and ice crystal formation (Kanakidou et al., 2005; Shrivastava et al., 2017). Furthermore, SOA play a major role in the context of air quality and have been linked to adverse health effects (Pöschl and Shiraiwa, 2015). Understanding SOA formation and evolution is complicated by the large number and variety of involved organic species and associated reactions and properties, making SOA a source of large uncertainties in climate and air quality modelling (Intergovernmental Panel on Climate Change, 2023).

The saturation vapor pressure (p_{vap}) of a compound determines its partitioning equilibrium between the condensed and gas phase. In the following, we will classify compounds into volatility ranges based on their saturation mass concentrations over the pure liquid (C_0)

as proposed by Donahue et al. (2009). The classes are extremely low-volatility organic compounds (ELVOC, $C_0 < 3\times10^{-6}\,\mu g\,m^{-3}$), low-volatility organic compounds (LVOC, $3\times10^{-6} < C_0 < 3\times10^{-4}\,\mu g\,m^{-3}$), semi-volatile organic compounds (SVOC, $3\times10^{-4} < C_0 < 300\,\mu g\,m^{-3}$), intermediate-volatility organic compounds (IVOC, $300 < C_0 < 3\times10^{6}\,\mu g\,m^{-3}$) and volatile organic compounds (IVOC, $C_0 > 3\times10^{6}\,\mu g\,m^{-3}$). In the atmosphere, saturation vapor pressure governs new particle formation and gas-particle partitioning, such that SOA mass yield is largely determined by p_{vap} (Pankow, 1987; Kulmala and Kerminen, 2008). However, due to the large number of atmospherically-relevant compounds, exhaustive experimental determination of p_{vap} is not feasible (Goldstein and Galbally, 2007; Bilde et al., 2015).

Various quantitative structure-activity relationship (QSAR) methods for the approximation of thermodynamic properties like p_{vap} or reactivity have been developed to address this limitation: empirical structure-property relationship models often map a sum formula to a thermodynamic property of interest, using algebraic equations with parameters that are fitted to experimental data (Donahue et al., 2011; Li et al., 2016). Group contribution models such as SIMPOL (Pankow and Asher, 2008) and EVAPORATION (Compernolle et al., 2011) can be classified as semi-empirical (Gani, 2019) as they incorporate existing theoretical knowledge about the relationships of structural features and chemical behavior into mathematical equations. This often includes the consideration the occurrences, positions, or interactions of functional groups, while also determining fit parameters using experimental data (Nannoolal et al., 2004; Moller et al., 2008). The consideration of specific functional groups limits group contribution models to certain compound classes, possibly leading to significant errors when applied to molecules outside their applicable range (Tahami et al., 2019). Quantum-mechanical calculation (QM) models based on density functional theory are a common non-empirical approach to property determination (Geerlings et al., 2003), and can be combined with empirical approaches (Ratcliff et al., 2017). Such quantum-mechanical calculations have been used for the generation of large data sets (Wang et al., 2017; Tabor et al., 2019; Besel et al., 2023), facilitating the development of machine learning (ML)-based QSAR models (Lumiaro et al., 2021; Krüger et al., 2022). When categorising ML-based QSAR models, we can distinguish the actual algorithm and the molecular representation that encodes molecular structures into suitable model input, which together majorly determine a ML model's performance in deriving properties from molecular structures (Lumiaro et al., 2021). Combinations successfully applied in previous studies include one-hot encoded Simplified Molecular Input Line Entry System (SMILES) strings with convolutional neural networks (OHE-CNN; Krüger et al., 2022), specific molecular descriptors with decision trees (Armeli et al., 2023) or topological fingerprints with Gaussian process regression (Besel et al., 2024). Galeazzo and Shiraiwa (2022) developed a method to predict glass transition temperature and melting points of small molecules using Extreme Gradient Boosting (XGBoost) and a neural network, respectively, in combination with derived molecular embeddings as molecular fingerprints. The transformation of molecular structures into such machine-readable molecular representations requires the ML models to learn the representation principles along with the physicochemical principles that determine the target property, to the detriment of limiting their application to the prediction of properties with extensive amounts of data (von Lilienfeld and Burke, 2020). This limitation can be mitigated using foundation models, pre-trained networks that are fine-tuned on relatively small data sets for a specific property (Burns et al., 2025). Data curation techniques can improve model accuracy, e.g., through identification and deletion of data points associated with large experimental uncertainty (Gadaleta et al., 2018; Ulrich et al., 2021). Within atmospheric chemistry, only few ML-based QSAR models have been trained exclusively on experimental measurements, as they generally require a large quantity of training data for sufficient model generalization, and a careful and computationally expensive error estimation when only limited amounts of data are available (Galeazzo and Shiraiwa, 2022; Armeli et al., 2023). The overall moderate to poor accuracy of existing QSAR models for p_{vap} prediction exemplifies the need for more accurate, publicly available models (Longnecker et al., 2025).

Graph neural networks (GNNs) are a class of algorithms within the domain of geometric deep learning which have emerged as a powerful addition to machine learning methods in computational chemistry and material sciences in the last decade (von Lilienfeld and Burke, 2020; Reiser et al., 2022). GNNs can be interpreted as an extension of convolutional neural networks beyond fixed dimension grids of data to include irregularly shaped structures (Kipf and Welling, 2017; Bronstein et al., 2017), such as graph-based representations of molecules (Duvenaud et al., 2015; Atz et al., 2021). Molecular graph representations and algorithms that operate on such graphs omit an additional representation learning step and can directly infer intramolecular spatial relations along with properties assigned to graph elements. Furthermore, in contrast to sum formula-based methods, structure-based methods can distinguish structural isomers, which may differ significantly in their properties (Isaacman-VanWertz and Aumont, 2021). Lumiaro et al. (2021) compared a variety of molecular fingerprints in combination with Kernel Ridge Regression, finding graph-based representations to be advantageous compared to canonical descriptive chemical features based methods. For the prediction of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, Xiong et al. (2021) employed a multi-task graph attention framework addressing classification and regression tasks.

In this work, we propose group contribution-assisted graph convolutional neural network (GC^{2}NN) models that are simultaneously trained on lists of molecular descriptors as well as graph representations of molecules, in which atom features are mapped to nodes, and bond features mapped to edges of a graph structure. We test model performance on data sets from experimental measurements and QM calculations (Besel et al., 2023), and compare our models with established methods for the determination of p_{vap}: one ML approach, where convolutional neural networks are trained on one-hot encoded SMILES representations (Krüger et al., 2022), two parameterizations, where p_{vap} are derived only from the compounds' elemental composition (Donahue et al., 2011; Li et al., 2016), and SIMPOL (Pankow and Asher, 2008), EVAPORATION (Compernolle et al., 2011), and EPI-Suite (EPI, 2024), which are commonly used semi-empirical group-contribution methods.

2 Methods

2.1 Vapor pressure data

We assembled a data set of SMILES representations of 6042 compounds with experimental saturation vapor pressure (p_{vap}) measurements at 298 K by crawling data from pubchem (Kim et al., 2016). In addition, we retrieved the data set published in Naef and Acree (2021), comprised of 2070 compounds. After removal of species present in both data sets, and species that contain elements that occur in fewer than 30 compounds, a total of 6178 unique compounds with experimental p_{vap} measurements are obtained and referred to as broad data. An overview of molecular substructures in the broad data set is displayed in Fig. 1a. It encompasses various compound types, such as aromatics, alcohols, carboxylic acids, esters, amines, amides, carbonyls, sulfides and nitriles. As the broad data set also contains $\sim 5\,\%$ inorganic compounds, we refer to compounds in this data set more generally as extremely low-volatility compounds (ELVOC), low-volatility compounds (LVOC), semi-volatile compounds (SVOC), intermediate-volatility compounds (IVOC) and volatile compounds (VOC), thus keeping the same acronyms and vapor pressures bins as Donahue et al. (2009) established for organic compounds. Experimental p_{vap} measurements range from 10^{-10} to 10^{7} Pa. The distribution of saturation concentrations and the number of ELVOC, LVOC, SVOC, IVOC and VOC are summarized in Fig. 1e. For a comparison with established methods for p_{vap} prediction, and to test the method on a data set of compounds that are relevant for the atmosphere, we extract all compounds that lie within the scope of these methods (Pankow and Asher, 2008; Compernolle et al., 2011; Donahue et al., 2011; Li et al., 2016), confining the data set to molecules only consisting of C, H, and O atoms and belonging to the following compound classes: alkanes, (non-aromatic) alkenes, aldehydes, ketones, ethers, esters, peroxides, nitrates, peroxy acyl nitrates, alcohols, acids, hydroperoxides and peracids. This subset of the broad data, referred to as confined data, contains a total of 1349 compounds with much smaller variety of compound classes, including carboxyl, hydroxyl, ester and carbonyl functional groups (Fig. 1b). While the overall p_{vap} range is very similar, the confined data set exhibits a smaller fraction of ELVOC, LVOC and SVOC than the broad data set (Fig. 1c, d, e, f). This skew towards higher vapor pressures in the confined data can be attributed to smaller molecules that contain fewer heavy atoms, as indicated by its lower average molecular mass of $154.8\,\mathrm{g\,mol^{-1}}$, compared to $205.8\,\mathrm{g\,mol^{-1}}$ in the broad data set. Both data sets are available for download, as specified in the data availability statement.

In addition to the experimental data, we train and evaluate GC^{2}NN models based on the quantum-mechanical (QM) data set GeckoQ (Besel et al., 2023). This data set contains a total of 31 637 compounds with calculated p_{vap}. Compounds in this data are carbon backbones derived from decane, toluene and α-pinene with various functional groups (including C, O, H). These structures were generated by the GECKO-A mechanism generator following Isaacman-VanWertz and Aumont (2021). GECKO-A simulates the atmospheric oxidation of hydrocarbons (Aumont et al., 2005), ensuring the atmospheric relevance of the compounds in this data set. Besel et al. conducted a conformer search using the COSMO*conf* program, calculated individual conformer p_{vap} values with COSMO*therm*, and determined a single p_{vap} accounting for the population of conformers according to the Boltzmann distribution (Wang et al., 2017; Kurtén et al., 2018; Hyttinen et al., 2022).

From each data set, we sample test sets (10 % of compounds) that are fully withheld from model training and used to evaluate the trained GC^{2}NN models. The remaining compounds in each data set (90 %) are used for training of the GC^{2}NN models, applying 5-fold cross-validation with 80 % of data in the training and 20 % in the validation set. The resulting data set sizes are the following: broad training: 4449, broad validation: 1112, broad test: 617, confined training: 972, confined validation: 243, confined test: 134, GeckoQ training: 22 778, GeckoQ validation: 5695, and GeckoQ test: 3164. p_{vap} measurements in Pa are logarithmized and scaled to a [0, 1]-interval using min-max scaling.

Of the 1349 molecules in the confined data set, 474 are also contained in the EVAPORATION training data (Compernolle et al., 2011). We ensure that no EVAPORATION training data are present in the test set that is used for comparison between the methods. Note that this only applies to EVAPORATION due to data availability and practicability; any other pre-trained or fitted method is likely to contain some fraction of the test set used in this study in their training data, including Donahue et al. (2011), Li et al. (2016), SIMPOL, and EPI-Suite.

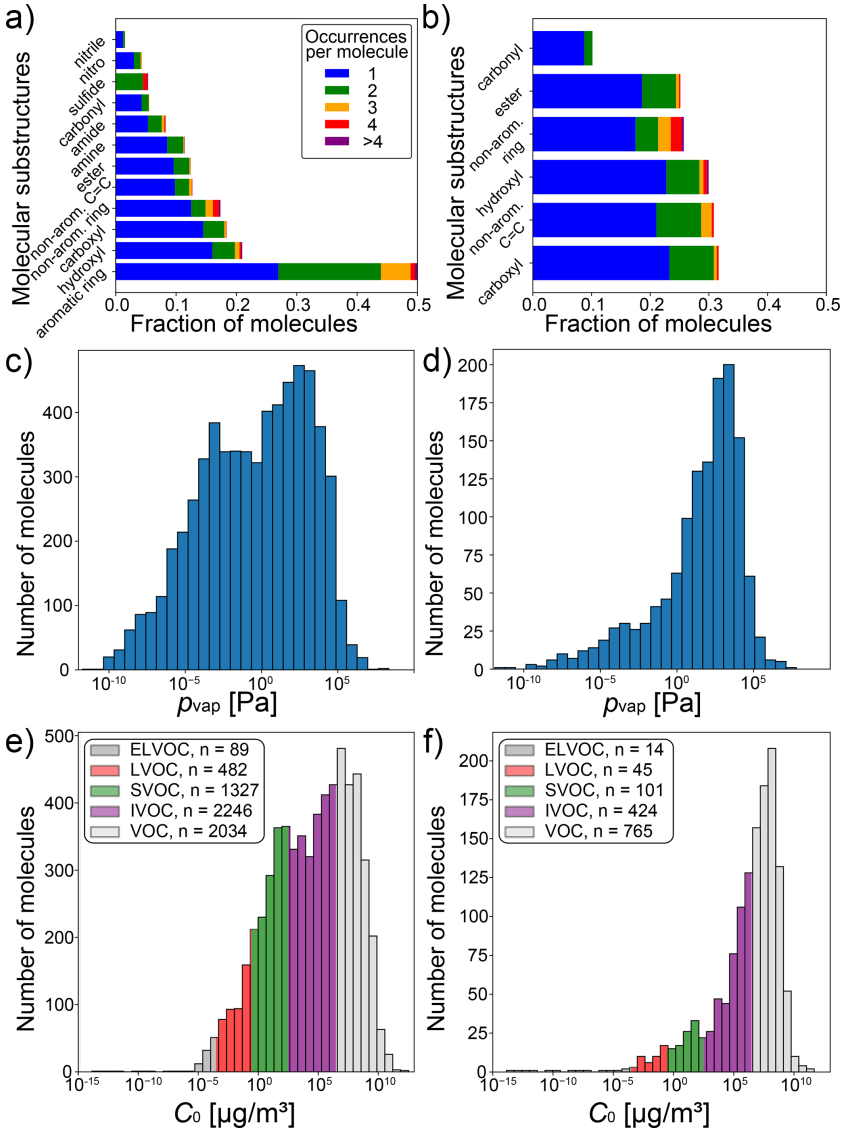

Figure 1. Overview of the two experimental data sets used in this study: broad data set ($n = 6178$; **a**, **c**, **e**) and confined subset ($n = 1349$; **b**, **d**, **f**). Panels **(a)** and **(b)** show all substructures which are present in more than 1 % of molecules in the respective data set (not shown: **a**: nitrate, sulfo, peroxide, organosulfate, peroxy acyl nitrate; **b**: peroxide). Panels **(c)** and **(d)** display histograms of experimental vapor pressure measurements in each data set, whereas Panels **(e)** and **(f)** show the same data as saturation mass concentrations (C_0). The volatility classes are adopted from Donahue et al. (2009).

2.2 Molecular representation

For the graph convolution component of the GC^{2}NN, we transform SMILES representations of molecular structures into graph-representations where atom features are mapped to node features, and bond features to edge features (Tables S1 and S2 in the Supplement). The final graph structure is comprised of three tensors. Each node and bond in the graph is associated with a vector of atom features and bond features, respectively. An adjacency matrix indicates the connectivity of atoms in the molecule. Graph convolution layers receive the adjacency matrix indicating which nodes (i.e., atoms) are connected, as well as the node feature matrix as inputs, graph attention layers additionally receive edge features. While the adjacency matrix remains unmodified to allow deduction of the connectivity for the following layers, each graph layer alters the feature matrix or matrices by aggregating features from neighboring nodes or edges, using the adjacency matrix to guide the aggregation.

For the model's group contribution component, a list of molecular descriptors (including molar mass, number of atoms for each element, and the number of common functional groups) are derived directly from the SMILES representation of the molecule. The descriptors are specific to each

data set and are summarized in Table S3. All descriptors and features are one-hot encoded or normalized to a [0, 1] interval.

2.3 Model architecture and training

We test and compare two group contribution-assisted graph convolutional neural networks (GC^{2}NN) models in this work: a fixed-depth GC^{2}NN (fdGC^{2}NN) model with a fixed number of graph layers, and an adaptive-depth GC^{2}NN (adGC^{2}NN) model where the number of graph layers is dynamically adapted based on a compound's size. Schematic overviews of the adGC^{2}NN and fdGC^{2}NN models are shown in Figs. 2 and S1, respectively. All GC^{2}NN models encompass two components with separate inputs that are derived from the SMILES-encoded molecular structure. The graph convolution component is comprised of multiple graph convolution layers and graph attention layers. Graph convolution layers apply convolution operations on each node, deriving information from the current node's properties, as well as its neighbors (Kipf and Welling, 2017; Zhang et al., 2019). Graph attention layers utilize attention mechanisms, enabling them to weigh convoluted nodes and features by their importance (Veličković et al., 2017; Withnall et al., 2020; Tang et al., 2020). This capability allows the assessment of feature importances by evaluating attention weights (Sanchez-Lengeling et al., 2020). Furthermore, graph attention layers enable the model to also derive information from edge attributes (Battaglia et al., 2018). Each graph attention or convolution layer increases the nodes' receptive fields, i.e. the distance between two nodes (and hence atoms) that still affect each other. To account for variable molecule sizes, we use the maximum distance between two atoms of a compound (maxdist) to determine the number of processing graph layers in the adGC^{2}NN, with a maximum of five layers for molecules with maxdist > 4. In the fdGC^{2}NN, all compounds are indiscriminately passed through five graph layers. The models' group contribution component is comprised of fully connected hidden layers that process additional molecular descriptors in parallel. Graph layer-specific merging layers map the information obtained from both model components to the output layer and a vapor pressure prediction. We use the Python packages RDKit and PyTorch (and PyTorch_Geometric) to generate the graph representations of molecular species from SMILES and train GC^{2}NN models (Landrum, 2013; Paszke et al., 2019).

The Python package Optuna (Akiba et al., 2019) is used to efficiently optimize hyperparameters of each GC^{2}NN model, using 5-fold cross-validation to mitigate variability due to the small data sets. We select mean absolute error (MAE) as loss function for model training, as well as model evaluation and comparison with established methods, due to its robustness over methods that give more weight to outliers such as root mean squared error (RMSE). This is particularly important given that the training data consist of experimental measurements that may possess high uncertainty and could be subject to systematic biases originating from different experimental setups. Measurements in the ELVOC range are particularly susceptible to higher experimental uncertainties, which would receive disproportionate weighting under RMSE-based training and consequently degrade model performance on other ranges. MAE allows for a reliable and interpretable evaluation of model accuracy without being overly influenced by extreme values. Hyperparameters are optimized by minimizing average validation loss across all cross-validation folds, but we reject models if the MAE standard deviation is larger than 0.08, to ensure robust model architectures. All models are trained to a maximum of 400 training epochs, unless validation loss does not decrease for 20 consecutive epochs. If so, model parameters are reset to the state of the epoch where the last validation loss decrease occurred, and training is terminated to avoid over-fitting. After the selection of suitable hyperparameters, a single model is trained by merging training and validation data to a single training data set, referred to as T+V model. To account for the additional training data, we locally optimize the number of training epochs around the number determined during hyperparameter tuning. A summary of the relevant hyperparameters including descriptions and tested ranges is displayed in Table S4. Hyperparameter optimization and model training are conducted on the Raven high-performance computing (HPC) system of Max Planck Computing and Data Facility (MPCDF), which provides GPU-accelerated compute nodes, each with four Nvidia A100-SXM4 GPUs and 160 GB HBM2. Each model is trained on a single Nvidia A100-SXM4 GPU using up to 24 GB of memory and PyTorch version 2.4.0 with CUDA version 12.1 support.

3 Results and discussion

We train and evaluate group contribution-assisted graph convolutional neural network (GC^{2}NN) models on two sets of experimental vapor pressure (p_{vap}) data and the GeckoQ data set where p_{vap} was derived from quantum-mechanical calculations (Besel et al., 2023). We distinguish between models trained on experimental data sets with different scopes: the GC^{2}NN-*confined* are trained on a confined data set that only contains compounds relevant in the atmosphere within the scope of the methods used for benchmarking, i.e. only containing C, H, and O, and excluding aromatics and some additional functional groups (Fig. 1b, d, f). GC^{2}NN-*broad* are trained on the full experimental data set (Fig. 1a, c, e).

3.1 GC^{2}NN-confined

Figure 3a shows that the adGC^{2}NN model exhibits excellent agreement with the experimental measurements in the independent test set, except from a small number of outliers (MAE = 0.37 log-units). Average training time of the

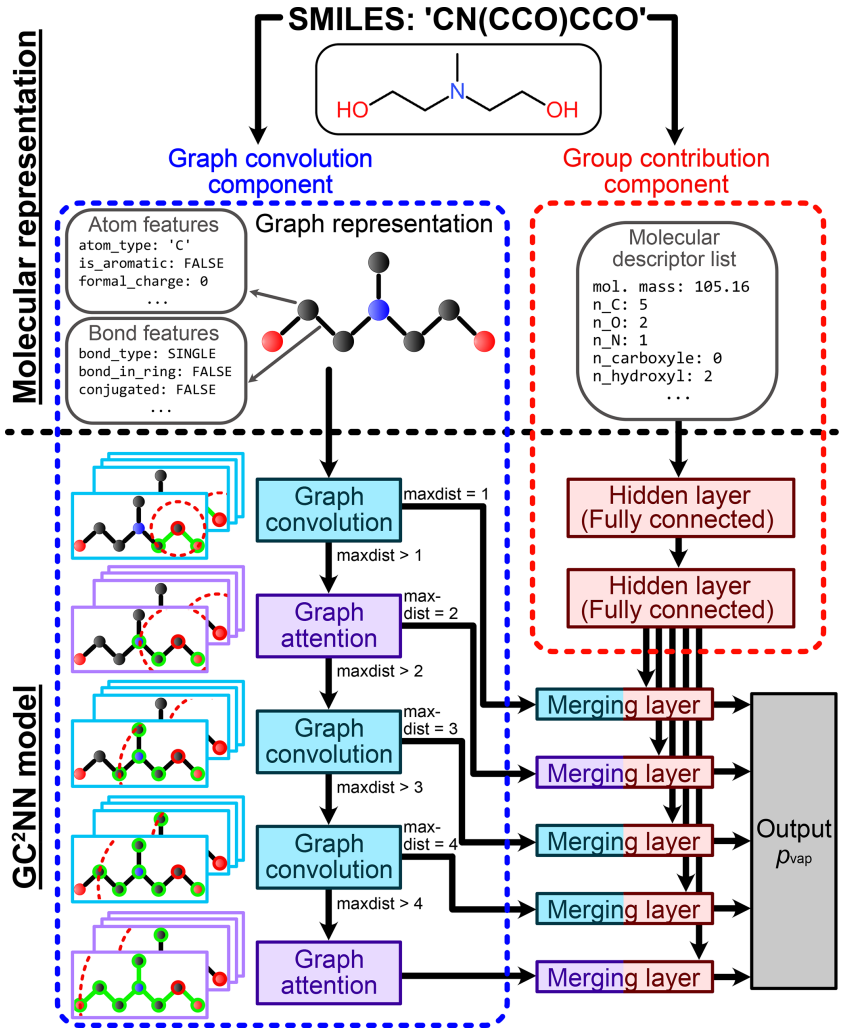

Figure 2. Schematic overview of molecular representation and model functionality in the adaptive-depth GC^{2}NN models. Right: for the group contribution component, Simplified Molecular Input Line Entry System (SMILES) strings are used to derive holistic information on the molecule, such as its molar mass and the presence of atoms and functional groups (Table S3). Left: for the model's graph convolution component, SMILES strings are transformed into graph representations, encoded as adjacency matrices, node features, and edge features. This molecular representation is transformed using graph attention and graph convolution layers. The maximum distance (maxdist) between two nodes in the input graph determines the number of utilized graph layers, matching the nodes' receptive fields with the respective compound's size. After passing all graph layers applicable to a compound, the convoluted and flattened node and edge feature matrices are concatenated with the processed data from the group contribution component. Fully-connected merging layers process these vectors and map them to the single-node output layer, the p_{vap} prediction.

five adGC^{2}NN cross-validation models is 55 min on a Nvidia A100, and the average test set mean absolute error (MAE) is 0.40 log-units with a standard deviation of 2.04×10^{-2}. The T+V fdGC^{2}NN performs worse with an MAE of 0.47 log-units. Average training time of the five fdGC^{2}NN cross-validation models is 22 min on a Nvidia A100, and the average test set mean absolute error (MAE) is 0.46 log-units with a standard deviation of 3.0×10^{-2}. The selected hyperparameters for all fdGC^{2}NN models are summarized in Table S5. The adGC^{2}NN model is more robust regarding the choice of hyperparameters, which permits the use of a single model architecture for all data sets. All adGC^{2}NN models possess two hidden layers with each 32 nodes in the group contribution component and a single merging layer with eight nodes for each graph convolution layer. The graph component of the adGC^{2}NN models is comprised of a total of five layers with 32, 16, 64, 16 and 32 nodes, using "LeakyReLU", "LeakyReLU", "ReLU", "ReLU" and "LeakyReLU" activation functions, respectively. Among these, the second and fifth layers are graph attention layers with six attention heads each, processing additional edge information. Training is conducted with a learning rate of 6.25×10^{-4}, a learning rate

decay of 0.985 per training epoch, no weight decay and a batch size of four (Fig. 2, Table S6). The adGC^{2}NN significantly outperforms the Krüger et al. (2022) one hot-encoding convolutional neural network approach (OHE-CNN; MAE = 0.79 log-units; average MAE = 0.93 log-units for five cross-validation folds), the Donahue et al. (2011) (MAE =1.61 log-units) and Li et al. (2016) (MAE =1.05 log-units) parameterizations, as well as EPI-Suite (MAE = 0.43 log-units), SIMPOL (MAE = 0.61 log-units) and EVAPORATION (MAE = 0.54 log-units) group contribution methods (Fig. 3). Note that the exclusion of a large fraction of molecules (>30 %) from the test set biases the populations of chemical species in the training and test set for the GC^{2}NN and OHE-CNN models (Fig. S2). This may be disadvantageous for the GC^{2}NN models, however, separate calculations with unbiased test set sampling show that the choice of the test set does not have a strong effect on the test set error of the GC^{2}NN models.

Figure 4 shows the distributions of the individual errors for chemical species in the test set for all methods. The fdGC^{2}NN-confined, SIMPOL and EVAPORATION methods exhibit near-identical error distributions where the majority of predictions are very accurate (MAE < 0.5 log-units), and few predictions fall within the range of 0.5 to 1.5 log-units. Only the adGC^{2}NN model has a larger density of very accurate predictions with only few compounds exceeding an MAE of 1.0. EPI-Suite shows an hour-glass shaped profile with a large fraction of very accurate predictions, as well as a large fraction of outliers. This is likely due to the presence of EPI-Suite training data in our test set. Methods for which this is likely the case are marked with an asterisk in Fig. 4. All methods generally perform better at higher p_{vap} (Fig. S3). This behavior correlates with a similar, but weaker bias with regards to molar mass (Fig. S4). The parameterization methods (Li et al., 2016; Donahue et al., 2011), which are solely based on elemental composition without considering functional group and molecular structure, exhibit the highest percentage of significant outliers.

For a general feature attribution analysis, we investigate attention scores of the second layer (graph attention) of the trained model's graph component. The attention weights, which are trained parameters of the model, are applied to each chemical compound and graph node (i.e., atom) to compute attention scores. They represent the calculated importances, quantifying the contribution of each node to the p_{vap} prediction relative to its neighboring nodes for a specific compound. For functional groups, importances of all associated atoms are averaged. With regards to single atoms, oxygen (0.36) scores a slightly larger attention score than carbon (0.32) in the confined test set (Fig. S5). Among functional groups, hydroxyl groups achieve the highest score.

To investigate the effect of experimental error in the low volatility range, we train fdGC^{2}NN models on a subset of the confined data with $\log_{10}(p_{vap}/[\mathrm{Pa}]) > 0$, encompassing only VOC and IVOC, resulting in 1057 compounds. The average test set MAE of the cross-validation folds of this high-volatility fdGC^{2}NN model is 0.32 log-units. This suggests that not only does experimental uncertainty of ELVOC and LVOC lead to model uncertainty in this low-volatility range, but it impedes the accuracy of fdGC^{2}NN models in general. To assess model uncertainty, we analyze ensemble predictions from the 5-fold cross-validation models on both confined and broad test data sets with regards to their prediction errors and standard deviations (Fig. S6). While the ensemble mean error represents model bias, the ensemble standard deviation can serve as an indicator for overall model uncertainty.

We use the trained adGC^{2}NN-confined model to review the concept of molecular corridors, following Shiraiwa et al. (2014), where the chemical evolution of molecules constituting SOA is contextualized through their vapor pressure, molar mass, and oxygen-to-carbon (O : C) ratio. The tight inverse correlation between volatility and molar mass mostly holds for the confined test set (Fig. 5a) as well as a data set of atmospherically-relevant compounds from Shiraiwa et al. (2014) (Fig. 5b). For the confined test set, the adGC^{2}NN predictions even tend to fall more strictly into these molecular corridors than the experimental measurements, a potential indicator for experimental uncertainties. When applied to the data from Shiraiwa et al. (2014), we observe a few compounds that appear to deviate from the molecular corridors by exceeding the upper boundary line corresponding to n-alkanes (O : C = 0). This deviation is either due to a mismatch between the adGC^{2}NN and the EVAPORATION model that was used to determine the boundary lines established in Shiraiwa et al. (2014), or could be due to a systematic error of the adGC^{2}NN as a result of the sparsity of ELVOC data in the training set (Fig. S2b). Furthermore, the difficulties of accurately determining vapor pressures of ELVOC experimentally (Huisman et al., 2013; Bilde et al., 2015) may contribute to this error. In atmospheric context, the accurate determination of ELVOC vapor pressure is not critical with regards to SOA formation, as such compounds condense anyway. Note however, that the accurate determination of ELVOC may be relevant in the context of nucleation, as recent experimental studies found ultra-low-volatility organic compounds (ULVOC) to nucleate, but not LVOC or ELVOC (Kirkby et al., 2023). Attempts have thus been undertaken previously to increase the representation of ELVOC molecules in training data sets for vapor pressure estimation models (Besel et al., 2024).

3.2 GC^{2}NN-broad

Compared to the confined data set, the broad data set encompasses a much larger range of molecular complexity, going far beyond molecules relevant for atmospheric SOA. Thus, despite a much larger training set size, the adGC^{2}NN-broad model achieves a lower test set accuracy than the adGC^{2}NN-confined model, with an MAE of 0.69 log-units for the T+V

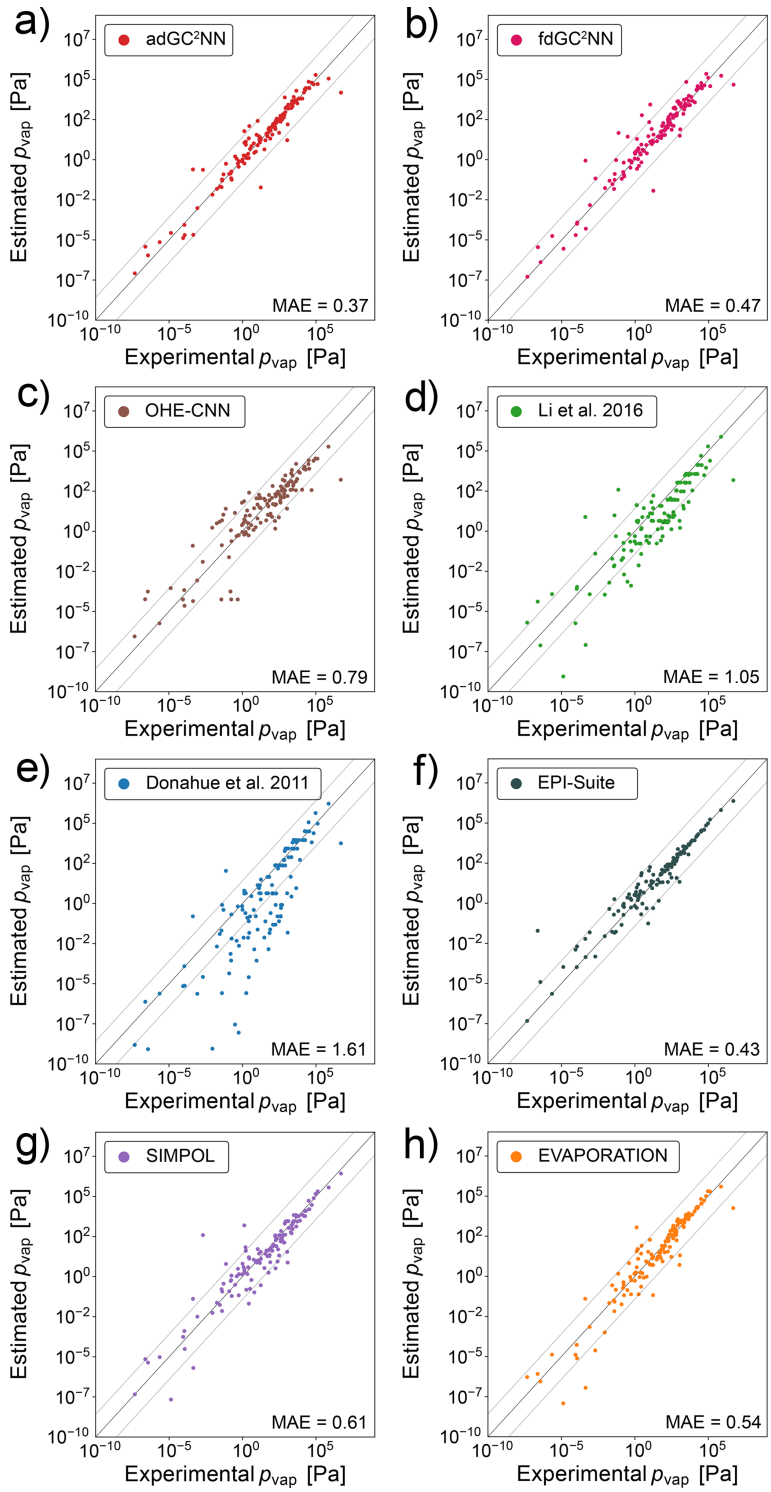

Figure 3. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the confined data set. Displayed are data from the independent test set only. The adGC^{2}NN-confined **(a)** and fdGC^{2}NN-confined **(b)** models are compared with established methods: **(c)** shows the results using a convolutional neural network approach on one-hot encoded SMILES strings following Krüger et al. (2022). **(d)** Li et al. (2016) and **(e)** Donahue et al. (2011) are empirical parameterizations, whereas **(f)** EPI-Suite (EPI, 2024), **(g)** Pankow and Asher (2008) and **(h)** Compernolle et al. (2011) are group contribution methods. All molecules present in the EVAPORATION training data have been excluded from the test data set. Mean absolute error (MAE) values are in $\log_{10}(p_{vap}/[Pa])$. The dashed lines (± 1.5 log-units from the 1 : 1 line) are used to indicate significant outliers.

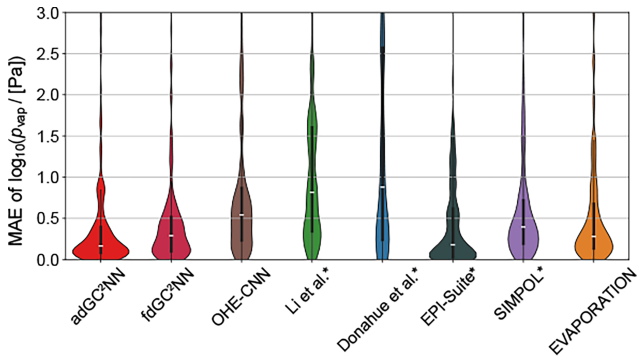

Figure 4. Violin plots representing confined test set error distribution of models shown in Fig. 3. Medians are displayed as white markers, interquartile ranges as vertical wide black lines and 1.5× interquartile ranges as vertical narrow black lines. Outliers with an MAE > 3 log-units are not shown. Methods marked with an asterisk likely used a fraction of our test data in their training.

model (Fig. 6). Average training time of the cross-validation models is 4.4 hours on a Nvidia A100 GPU, and the average test set mean absolute error (MAE) is 0.71 log-units with a standard deviation of 3.02×10^{-2}. The T+V fdGC^{2}NN model performs worse with an MAE of 0.77 log-units. Cross-validation fdGC^{2}NN models have an average test set MAE of 0.78 with a standard deviation of $2.36 \times 10^{-}2$ and an average training time of 2.4 h. Both GC^{2}NN models outperform the OHE-CNN approach from Krüger et al. (2022) (MAE = 0.99 log-units; average MAE = 0.96 log-units for five cross-validation folds), but have a similar test set error than EPI-Suite (EPI, 2024) (MAE = 0.69 log-units). Error distributions for the broad test set are displayed in Fig. S7. Note that EPI-Suite was trained on larger data sets that are not publicly available. As discussed above, the MAE that EPI-Suite achieves in our test set is likely biased through overlap of training and test data and thus not fully representative for unknown molecules.

We also train a fdGC^{2}NN model on a subset of the broad data with $\log_{10}(p_{\mathrm{vap}}/[\mathrm{Pa}]) > 0$ to investigate the effect of experimental uncertainty in the low-volatility range. Due to the large fraction of low-volatile compounds in the broad data, the high-volatility subset only contains roughly 50 % of the original compounds ($n = 3116$). The cross-validation models achieve an average MAE of 0.37 log-units, greatly reducing the error by nearly 50 % and outperforming EPI-Suite (Figs. S8, S9). An uncertainty analysis based on ensemble predictions for the broad test data of the 5-fold cross validation adGC^{2}NN-broad models is shown in Fig. S10. Calculated attention scores for single atoms and functional groups are summarized in Fig. S11. Notably, we observe a good agreement between the attention score orders of functional groups between adGC^{2}NN-confined and adGC^{2}NN-broad, with hydroxyl groups having the highest scores, followed by carbonyl groups, ester groups and finally non-aromatic C=C double bonds. The importance of hydroxyl groups may be attributed to their ability to form hydrogen bonds that reduce the compound's vapor pressure. Note that feature importances assigned by trained models are not exclusively governed by chemical principles, but also the prevalence and distribution of substructures in the training data. Rarity and commonness of certain substructures may both decrease associated feature importances, as high importances are attributed to relevant features that enable the model to distinguish compounds of the training population. To differentiate between chemistry-governed and prevalence-governed importances, feature attribution analyses could be supported by generative sensitivity studies, where the effect of substructures on p_{vap} predictions is statistically tested through systematic substitution of substructures in template compounds. A molecular corridor plot following Shiraiwa et al. (2014) for the adGC^{2}NN-broad model is displayed in Fig. S12, exhibiting a much stronger overestimation of ELVOC vapor pressures than the confined model (Fig. 5). Thus, it appears that the higher diversity of molecular features in the broad data set exacerbates the problem of sparse data in the ELVOC range.

3.3 GC^{2}NN-GeckoQ

In addition to the experimental data sets, we train GC^{2}NN models on the GeckoQ data from Besel et al. (2023), which were derived from quantum-mechanical calculations. For the T+V adGC^{2}NN model, the average test set mean absolute error (MAE) is 0.66 log-units (Fig. 7). The five adGC^{2}NN cross-validation models achieve an MAE of 0.67 log-units, average training time is 13.77 h on a Nvidia A100. Again, the adGC^{2}NN model achieves a better result than the fdGC^{2}NN model (MSE = 0.71 log-units; average MAE = 0.74 log-units for five cross-validation folds with an average training time of 3.4 h on a Nvidia A100), as well as the model adapted from Krüger et al. (2022) for p_{vap} prediction (MAE = 0.77 log-units; average MAE = 0.77 log-units for five cross-validation folds). It also outperforms the Gaussian Process Regression model presented in Besel et al. (2023) which achieved a test set MAE of 0.82 log-units.

3.4 Learning curves

Figure 8 shows learning curves for the adGC2NN and fdGC2NN models for each of the three data sets (broad, confined, GeckoQ). Learning curves are obtained by training on data subsets of specific sizes, while consistently using the hyperparameter sets optimized for the full data sets (Tables S5, S6). Gradients and convergence rates of the learning curves significantly differ between the models and data sets. In general, the fdGC^{2}NN models exhibit steeper learning curves than the adGC^{2}NN models, demonstrating the superiority of the adG^{2}NN model architecture across various data set sizes and data sets. Note that only one adGC^{2}NN ar-

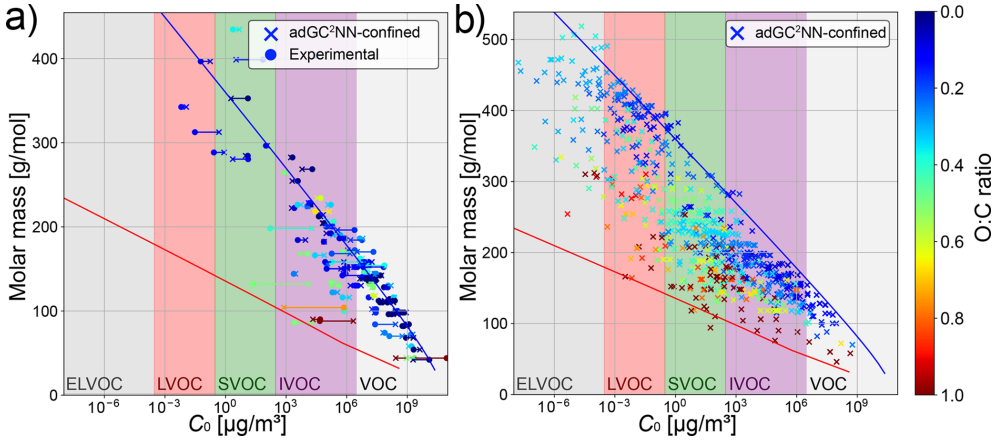

Figure 5. Molecular corridor plots following Shiraiwa et al. (2014). Left: comparison between adGC^{2}NN-confined predictions and experimental measurements in the confined test set. Right: application of the adGC^{2}NN-confined to a data set of atmospherically relevant compounds (Shiraiwa et al., 2014). Blue and red boundary lines correspond to the volatility of n-alkanes and sugar alcohols (as determined by EVAPORATION), respectively.

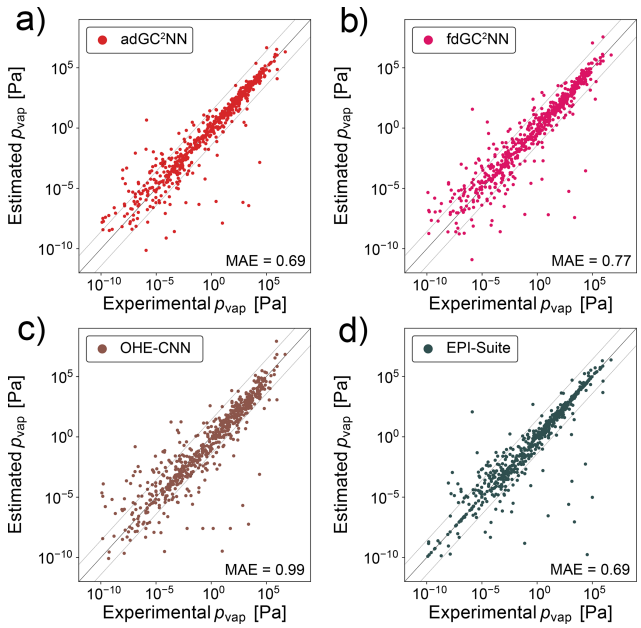

Figure 6. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the broad data set. Displayed are data from the independent test set only. **(a)** adGC^{2}NN-broad model, **(b)** fdGC^{2}NN-broad model, **(c)** OHE-CNN method presented in Krüger et al. (2022), and **(d)** EPI-Suite (EPI, 2024). Mean absolute error (MAE) values are in $\log_{10}(p_{vap}/[\text{Pa}])$. The dashed lines ($\pm 1.5$ log-units from the 1 : 1 line) are used to indicate significant outliers.

chitecture and hyper parameter set is consistently used across the study, while fdGC^{2}NN models are optimized individually for each of the three data sets. We observe that significantly more data are needed to achieve the same accuracy if the data contain a large variety of compound classes, as the broad and GeckoQ data models show consistently higher MAE than the confined data models for data sets of similar size. In the broad and GeckoQ data, the high diversity of molecular features and, potentially, their complex interactions require much more data for accurate predictions. None of the learning curves appear to fully level-off for large data set sizes, which means that the models can be expected to improve significantly with additional training data.

In addition to the adGC^{2}NN and fdGC^{2}NN models, we tested graph-only models without the additional input layer to obtain holistic molecular information (group-contribution component). These pure GCNN models are associated with significantly larger errors for nearly all data sets and sizes (Fig. S13), despite data set size-specific hyperparameter tuning. This can be attributed to graph convolutions which, in principle, are merely a succession of local operations on subgraphs. In other words, a pure graph convolutional neural network performs local operations on the input graph that are independent and unaware of operations and interpretations that occur in distant areas of the molecular graph. Deducting and learning holistic molecular information only from local convolutions on the graph structure is difficult, especially for the larger molecules. As each additional convolution layer increases the distance allowed for two nodes (and hence atoms) to influence each other, setting the number of graph convolution layers to the largest distance between two nodes in the data set would enable the model to derive information from each molecule as a whole. However, intramolecular interactions are usually not long ranged. Furthermore, this is detrimental for most model training because it would result in very deep neural networks which would likely over-fit on most data sets. Therefore, since the graph neural network training might not effectively capture whole-molecule properties, the lack of information on general molecular proper-

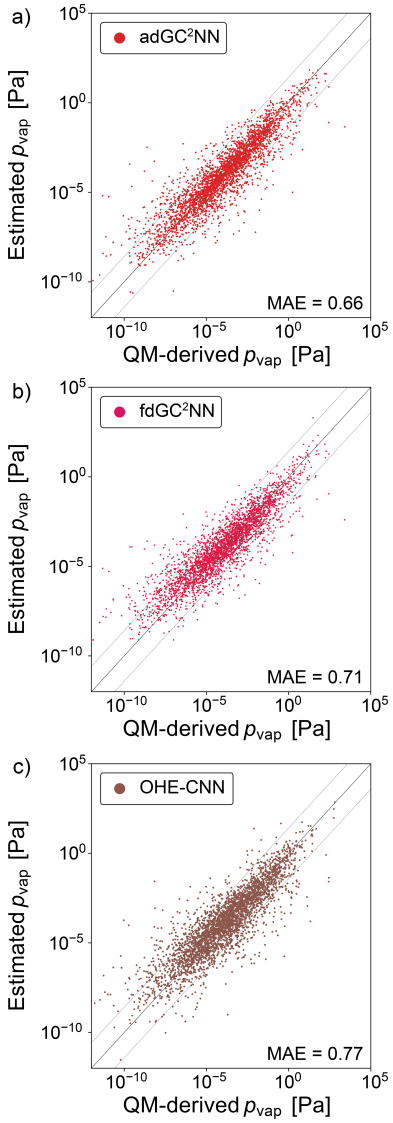

Figure 7. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the GeckoQ data set. Displayed are data from the independent test set only. **(a)** adGC^{2}NN-GeckoQ model **(b)** fdGC^{2}NN-GeckoQ model, and **(c)** OHE-CNN method presented in Krüger et al. (2022). Mean absolute error (MAE) values are in $\log_{10}(p_{vap}/[\text{Pa}])$. The dashed lines ($\pm 1.5$ log-units from the 1 : 1 line) are used to indicate significant outliers.

ties, like molar mass, inhibits the graph-only models to generalize between molecules of different size. We observe that the addition of molar mass as an input is crucial for the performance of GC^{2}NN, while additional descriptors like element and functional group counts lead to further, but minor improvements.

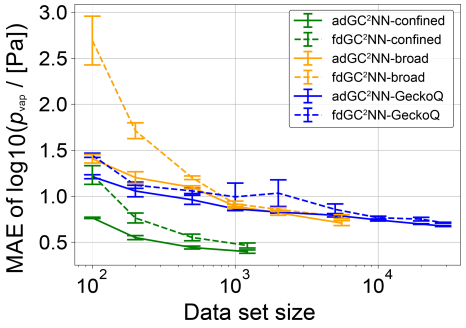

Figure 8. Mean absolute error (MAE) for independent test sets (confined: $n = 134$; broad: $n = 617$; GeckoQ: $n = 3163$), as a function of training data set size of adGC^{2}NN and fdGC^{2}NN models trained on subsets of the three data sets. The experiment is performed by sampling subsets of various size from each of the respective data sets and training adGC^{2}NN and fdGC^{2}NN models on these. Shown are the average test set log unit MAE of five cross-validation models in each subset. Error bars represent standard deviations across the cross-validation folds.

4 Summary and conclusions

Our findings suggest that group contribution-assisted graph convolutional neural networks (GC^{2}NN) and graph representations of molecules are a promising approach for quantitative structure-activity relationship (QSAR) models. Despite the challenging scarcity of experimental data available for atmospherically relevant compounds, the GC^{2}NN models surpass established methods, including parameterizations, group contribution methods, and machine learning (ML) approaches. Graph representations are a natural and unambiguous representation of molecular structures, encoding additional information related to individual atoms (graph nodes) or bonds (graph edges), and making spatial relations between molecular substructures directly interpretable by ML models suitable for graph processing. With that, graph representations are advantageous over molecular representations in which spatial information are lost or not easily retrievable, such as one-hot encoded (OHE) SMILES strings, which we used previously in conjunction with convolutional neural networks (CNN) for the determination of quinone redox potentials Krüger et al. (2022). In this study, OHE-CNN models performed worse than GC^{2}NN models for every tested data set. Note, however, that we only performed a very basic tuning of the hyperparameters from the original study and correlation of the OHE-CNN model may improve with more extensive optimization.

We find that models that combine graph convolution with the direct interpretation of molecular properties like molar mass, element, and functional group occurrences outperform models that only process one of the two. The accuracy of graph-only GCNN models, without the additional input layer, falls behind pure group contribution models that process information on functional groups under consideration of

known principles governing their effect on molecular properties. The provision of holistic information on the molecular structure, especially molar mass, is crucial for the performance of GC^{2}NN models, as graph convolutions only process structural information locally. The difficulty in the application of graph convolutional neural networks is their dependence on the size of the input graphs. Therefore, specialized fdGC^{2}NN models for narrow vapor pressure ranges achieved excellent results, given sufficient training data, in this study. Our adaptive-depth approach, however, enables the GC^{2}NN to make use of the full training data, while matching the individual nodes' receptive fields with the compound size dynamically.

In general, the application of machine learning with few data is challenging, and learning curves suggest that additional data would significantly improve model accuracy for all compound ranges. We hypothesize that ML QSAR models may furthermore improve through prediction of multiple related molecular properties at a time. For instance, vapor pressure-predicting models may benefit from the simultaneous prediction of melting points or glass transition temperature, as the addition of such properties in the training data possibly makes physical principles more accessible by the model. Additional molecular parameters that are known to affect vapor pressure, such as polarity and representations of secondary intermolecular bonding, might also increase prediction performances with a similar architecture in the future. However, this may pose further restrictions on the training data available while highlighting how the application of machine learning methods in atmospheric chemistry is currently limited by the scarcity of comprehensive experimental data sets involving atmospheric compounds. The problem of data scarcity is very evident for compounds in the ELVOC range, which are comparably rare and underrepresented in our data set. This may be due to greater difficulties in the experimental determination of saturation vapor pressures of ELVOCs. To accurately extend QSAR models to the ELVOC range, possible strategies may include the utilization of quantum mechanical-derived data instead of experimental data, or potentially the application of more advanced machine learning models that include heuristic rules or physics-informed modules (Bilde et al., 2015), transfer learning to enable extrapolation outside of the training domain (Lansford et al., 2023) or pre-trained models that can be fine-tuned using small data sets (Burns et al., 2025). Our adaptive-depth model, however, achieved overall good results given relatively few training data, making the architecture a promising candidate for QSAR models addressing other molecular properties with relevance for atmospheric chemistry and physics, such as Henry's law solubility coefficients or reaction rate coefficients. Furthermore, the multiple component approach to QSAR modelling permits the utilization of far more advanced group contribution components alongside the graph convolution component. While the shallow neural networks in our study can indiscriminately be applied to various molecular descriptors and data sets, the utilization of advanced group contribution methods like SIMPOL or EVAPORATION alongside the graph convolution component, or the utilization of additional molecular descriptors may significantly increase model accuracy. In a similar fashion, QSAR models can likely be improved through integration of physics-informed models or hybrid quantum-mechanical/machine learning models (Zhang et al., 2018).

By using data sets of differing molecular complexity, a broad data set using most web-crawled data and a data set confined for atmospherically-relevant compounds, we find that the more specialized model can achieve a higher test set accuracy. In turn, while the models training on the broad data set have the largest error of all GC^{2}NN models in this study, they are applicable to a large population of compounds with a diverse elemental composition and variety of functional groups, encompassing both organic and inorganic species. It is therefore recommended to train QSAR models that are specific to certain molecule scopes and applications. We also find that model accuracy significantly differs between models that are trained on subsets of the p_{vap} range, and that models that are trained on smaller ranges can outperform more general models despite training data scarcity. In practice, an ensemble approach with multiple models, e.g., specifically for the low and high volatility range may be a viable approach for ML methods, similarly to the ensemble utilization of the Modified Grain, Antoine and Mackay methods (EPI, 2024; Li et al., 2016). Further improvements may be achievable through data curation techniques, as common outliers between various methods indicate data points with large experimental uncertainty.

The data sets (broad and confined) as well as the associated trained models are published along with this study. The compiled experimental vapor pressure data can be used for future benchmarking or training of vapor pressure estimation methods. Furthermore, our trained adGC^{2}NN models can be downloaded as easy-to-use executables, enabling researchers in various fields to obtain accurate vapor pressure predictions for their research, e.g., in the fields of SOA modeling or climate simulations. To run the models, no knowledge on machine learning or programming is required.

Code and data availability. The data and source code, as well as a model executable, are openly available at https://doi.org/10.17617/3.GIKHJL (Krüger and Berkemeier, 2025).

Supplement. The supplement related to this article is available online at https://doi.org/10.5194/gmd-18-7357-2025-supplement.

Author contributions. MS and TG conceived the study. All authors designed research. MK and TG wrote the code and performed

model simulations. All authors discussed and interpreted calculation results. MK and TB wrote the manuscript with contributions from all authors.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Also, please note that this paper has not received English language copy-editing. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We thank Steven Compernolle for providing the list of molecules contained in EVAPORATION training data in machine-readable format. We thank Nadin Ulrich for helpful discussions.

Financial support. This work was funded by the U.S. Department of Energy (grant no. DE-SC0022139), the U.S. National Science Foundation (grant no. AGS-2246502) and the Max Planck Society (MPG). Matteo Krüger is supported by the Max Planck Graduate Center with the Johannes Gutenberg University Mainz (MPGC).

The article processing charges for this open-access publication were covered by the Max Planck Society.

Review statement. This paper was edited by Jason Williams and reviewed by Patrick Rinke and one anonymous referee.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage AK USA, 2623–2631, ISBN 978-1-4503-6201-6, https://doi.org/10.1145/3292500.3330701, 2019.

Armeli, G., Peters, J.-H., and Koop, T.: Machine-Learning-Based Prediction of the Glass Transition Temperature of Organic Compounds Using Experimental Data, ACS Omega, 8, 12298–12309, https://doi.org/10.1021/acsomega.2c08146, 2023.

Atz, K., Grisoni, F., and Schneider, G.: Geometric deep learning on molecular representations, Nat. Mach. Intell., 3, 1023–1032, https://doi.org/10.1038/s42256-021-00418-8, 2021.

Aumont, B., Szopa, S., and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos. Chem. Phys., 5, 2497–2517, https://doi.org/10.5194/acp-5-2497-2005, 2005.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., Dyer, C., Heess, N., Wierstra, D., Kohli, P., Botvinick, M., Vinyals, O., Li, Y., and Pascanu, R.: Relational inductive biases, deep learning, and graph networks, arXiv [preprint], https://doi.org/10.48550/ARXIV.1806.01261, 2018.

Besel, V., Todorović, M., Kurtén, T., Rinke, P., and Vehkamäki, H.: Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules, Sci. Data, 10, 450, https://doi.org/10.1038/s41597-023-02366-x, 2023.

Besel, V., Todorović, M., Kurtén, T., Vehkamäki, H., and Rinke, P.: The search for sparse data in molecular datasets: Application of active learning to identify extremely low volatile organic compounds, J. Aerosol Sci., 179, 106375, https://doi.org/10.1016/j.jaerosci.2024.106375, 2024.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., Topping, D., Ziemann, P., Barley, M., Clegg, S., Dennis-Smither, B., Hallquist, M., Hallquist, A. M., Khlystov, A., Kulmala, M., Mogensen, D., Percival, C. J., Pope, F., Reid, J. P., Ribeiro da Silva, M. A. V., Rosenoern, T., Salo, K., Soonsin, V. P., Yli-Juuti, T., Prisle, N. L., Pagels, J., Rarey, J., Zardini, A. A., and Riipinen, I.: Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures, Chem. Rev., 115, 4115–4156, https://doi.org/10.1021/cr5005502, 2015.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P.: Geometric Deep Learning: Going beyond Euclidean data, IEEE Signal Process. Mag., 34, 18–42, https://doi.org/10.1109/MSP.2017.2693418, 2017.

Burns, J., Zalte, A., and Green, W.: Descriptor-based Foundation Models for Molecular Property Prediction, arXiv [preprint], https://doi.org/10.48550/ARXIV.2506.15792, 2025.

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmos. Chem. Phys., 11, 9431–9450, https://doi.org/10.5194/acp-11-9431-2011, 2011.

Donahue, N. M., Robinson, A. L., and Pandis, S. N.: Atmospheric organic particulate matter: From smoke to secondary organic aerosol, Atmos. Environ., 43, 94–106, https://doi.org/10.1016/j.atmosenv.2008.09.055, 2009.

Donahue, N. M., Epstein, S. A., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics, Atmos. Chem. Phys., 11, 3303–3318, https://doi.org/10.5194/acp-11-3303-2011, 2011.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P.: Convolutional Networks on Graphs for Learning Molecular Fingerprints, arXiv [preprint], https://doi.org/10.48550/arXiv.1509.09292, 2015.

EPI: EPI Suite™-Estimation Program Interface, https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface, last access: 13 October 2024.

Gadaleta, D., Lombardo, A., Toma, C., and Benfenati, E.: A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications, J. Cheminform., 10, 60, https://doi.org/10.1186/s13321-018-0315-6, 2018.

Galeazzo, T. and Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds *via* machine learning and molecular embeddings, Environ. Sci. Atmos., 2, 362–374, https://doi.org/10.1039/D1EA00090J, 2022.

Gani, R.: Group contribution-based property estimation methods: advances and perspectives, Curr. Opin. Chem. Eng., 23, 184–196, https://doi.org/10.1016/j.coche.2019.04.007, 2019.

Geerlings, P., De Proft, F., and Langenaeker, W.: Conceptual Density Functional Theory, Chem. Rev., 103, 1793–1874, https://doi.org/10.1021/cr990029p, 2003.

Goldstein, A. H. and Galbally, I. E.: Known and unexplored organic constituents in the earth's atmosphere, Environ. Sci. Technol., 41, 1514–1521, 2007.

Huisman, A. J., Krieger, U. K., Zuend, A., Marcolli, C., and Peter, T.: Vapor pressures of substituted polycarboxylic acids are much lower than previously reported, Atmos. Chem. Phys., 13, 6647–6662, https://doi.org/10.5194/acp-13-6647-2013, 2013.

Hyttinen, N., Pullinen, I., Nissinen, A., Schobesberger, S., Virtanen, A., and Yli-Juuti, T.: Comparison of saturation vapor pressures of α-pinene $+ O_3$ oxidation products derived from COSMO-RS computations and thermal desorption experiments, Atmos. Chem. Phys., 22, 1195–1208, https://doi.org/10.5194/acp-22-1195-2022, 2022.

Intergovernmental Panel on Climate Change: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, 1 edn., Cambridge University Press, ISBN 978-1-00-915789-6, https://doi.org/10.1017/9781009157896, 2023.

Isaacman-VanWertz, G. and Aumont, B.: Impact of organic molecular structure on the estimation of atmospherically relevant physicochemical parameters, Atmos. Chem. Phys., 21, 6541–6563, https://doi.org/10.5194/acp-21-6541-2021, 2021.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., E., Dunlea, J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525–1529, https://doi.org/10.1126/science.1180353, 2009.

Kanakidou, M., Seinfeld, J. H., Pandis, S. N., Barnes, I., Dentener, F. J., Facchini, M. C., Van Dingenen, R., Ervens, B., Nenes, A., Nielsen, C. J., Swietlicki, E., Putaud, J. P., Balkanski, Y., Fuzzi, S., Horth, J., Moortgat, G. K., Winterhalter, R., Myhre, C. E. L., Tsigaridis, K., Vignati, E., Stephanou, E. G., and Wilson, J.: Organic aerosol and global climate modelling: a review, Atmos. Chem. Phys., 5, 1053–1123, https://doi.org/10.5194/acp-5-1053-2005, 2005.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J., and Bryant, S. H.: PubChem substance and compound databases, Nucleic Acids Research, 44, D1202–D1213, https://doi.org/10.1093/nar/gkv951, 2016.

Kipf, T. N. and Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks, arXiv [preprint], https://doi.org/10.48550/arXiv.1609.02907, 2017.

Kirkby, J., Amorim, A., Baltensperger, U., Carslaw, K. S., Christoudias, T., Curtius, J., Donahue, N. M., Haddad, I. E., Flagan, R. C., Gordon, H., Hansel, A., Harder, H., Junninen, H., Kulmala, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Lelieveld, J., Möhler, O., Riipinen, I., Stratmann, F., Tomé, A., Virtanen, A., Volkamer, R., Winkler, P. M., and Worsnop, D. R.: Atmospheric new particle formation from the CERN CLOUD experiment, Nat. Geosci., 16, 948–957, https://doi.org/10.1038/s41561-023-01305-0, 2023.

Krüger, M. and Berkemeier T.: Code and data for 'Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC2NN)', Edmond, V2 [code and data], https://doi.org/10.17617/3.GIKHJL, 2025.

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B. A. M., Lammel, G., Schmidt, B., Pöschl, U., and Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, Nat. Sci., 2, e20220016, https://doi.org/10.1002/ntls.20220016, 2022.

Kulmala, M. and Kerminen, V.-M.: On the formation and growth of atmospheric nanoparticles, Atmos. Res., 90, 132–150, https://doi.org/10.1016/j.atmosres.2008.01.005, 2008.

Kurtén, T., Hyttinen, N., D'Ambro, E. L., Thornton, J., and Prisle, N. L.: Estimating the saturation vapor pressures of isoprene oxidation products $C_5H_{12}O_6$ and $C_5H_{10}O_6$ using COSMO-RS, Atmos. Chem. Phys., 18, 17589–17600, https://doi.org/10.5194/acp-18-17589-2018, 2018.

Landrum, G.: RDKit: Open-source cheminformatics, Release, 1, 4, https://www.rdkit.org (last access: 1 October 2025), 2013.

Lansford, J. L., Jensen, K. F., and Barnes, B. C.: Physics-informed Transfer Learning for Out-of-sample Vapor Pressure Predictions, Propellants Explos. Pyrotech., 48, e202200265, https://doi.org/10.1002/prep.202200265, 2023.

Li, Y., Pöschl, U., and Shiraiwa, M.: Molecular corridors and parameterizations of volatility in the chemical evolution of organic aerosols, Atmos. Chem. Phys., 16, 3327–3344, https://doi.org/10.5194/acp-16-3327-2016, 2016.

Longnecker, E. R., Bakker-Arkema, J. G., and Ziemann, P. J.: Comparison of Vapor Pressure Estimation Methods Used to Model Secondary Organic Aerosol Formation from Reactions of Linear and Branched Alkenes with OH/NO_x, ACS Earth Space Chem., https://doi.org/10.1021/acsearthspacechem.4c00285, 2025.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., and Rinke, P.: Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning, Atmos. Chem. Phys., 21, 13227–13246, https://doi.org/10.5194/acp-21-13227-2021, 2021.

Moller, B., Rarey, J., and Ramjugernath, D.: Estimation of the vapour pressure of non-electrolyte organic compounds via group

contributions and group interactions, J. Mol. Liq., 143, 52–63, https://doi.org/10.1016/j.molliq.2008.04.020, 2008.

Naef, R. and Acree, W. E.: Calculation of the Vapour Pressure of Organic Molecules by Means of a Group-Additivity Method and Their Resultant Gibbs Free Energy and Entropy of Vaporization at 298.15 K, Molecules [data], 26, 1045, https://doi.org/10.3390/molecules26041045, 2021.

Nannoolal, Y., Rarey, J., Ramjugernath, D., and Cordes, W.: Estimation of pure component properties, Fluid Phase Equilibria, 226, 45–63, https://doi.org/10.1016/j.fluid.2004.09.001, 2004.

Pankow, J. F.: Review and comparative analysis of the theories on partitioning between the gas and aerosol particulate phases in the atmosphere, Atmos. Environ. (1967), 21, 2275–2283, https://doi.org/10.1016/0004-6981(87)90363-5, 1987.

Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmos. Chem. Phys., 8, 2773–2796, https://doi.org/10.5194/acp-8-2773-2008, 2008.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst., 32, https://doi.org/10.48550/arXiv.1912.01703, 2019.

Pöschl, U. and Shiraiwa, M.: Multiphase Chemistry at the Atmosphere–Biosphere Interface Influencing Climate and Public Health in the Anthropocene, Chem. Rev., 115, 4440–4475, https://doi.org/10.1021/cr500487s, 2015.

Ratcliff, L. E., Mohr, S., Huhs, G., Deutsch, T., Masella, M., and Genovese, L.: Challenges in large scale quantum mechanical calculations, WIREs Comput. Mol. Sci., 7, e1290, https://doi.org/10.1002/wcms.1290, 2017.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., and Friederich, P.: Graph neural networks for materials science and chemistry, Commun. Mater., 3, 241722, https://doi.org/10.1038/s43246-022-00315-6, 2022.

Sanchez-Lengeling, B., Wei, J., Lee, B., Reif, E., Wang, P., Qian, W., McCloskey, K., Colwell, L., and Wiltschko, A.: Evaluating Attribution for Graph Neural Networks, in: Adv. Neural Inf. Process. Syst., edited by: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., vol. 33, Curran Associates, Inc., 5898–5910, https://proceedings.neurips.cc/paper_files/paper/2020/file/417fbbf2e9d5a28a855a11894b2e795a-Paper.pdf (last access: 1 October 2025), 2020.

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmos. Chem. Phys., 14, 8323–8341, https://doi.org/10.5194/acp-14-8323-2014, 2014.

Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing, Rev. Geophys., 55, 509–559, https://doi.org/10.1002/2016RG000540, 2017.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., and Aspuru-Guzik, A.: Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries, J. Mater. Chem. A, 7, 12833–12841, https://doi.org/10.1039/C9TA03219C, 2019.

Tahami, S., Movagharnejad, K., and Ghasemitabar, H.: Estimation of the critical constants of organic compounds via a new group contribution method, Fluid Ph. Equilibria, 494, 45–60, https://doi.org/10.1016/j.fluid.2019.04.022, 2019.

Tang, B., Kramer, S. T., Fang, M., Qiu, Y., Wu, Z., and Xu, D.: A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility, J. Cheminformatics, 12, 1–9, 2020.

Ulrich, N., Goss, K.-U., and Ebert, A.: Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation, Commun. Chem., 4, 90, https://doi.org/10.1038/s42004-021-00528-9, 2021.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph Attention Networks, arXiv [preprint], https://doi.org/10.48550/arxiv.1710.10903, 2017.

von Lilienfeld, O. A. and Burke, K.: Retrospective on a decade of machine learning for chemical discovery, Nat. Commun., 11, https://doi.org/10.1038/s41467-020-18556-9, 2020.

Wang, C., Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q., and Wania, F.: Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products, Atmos. Chem. Phys., 17, 7529–7540, https://doi.org/10.5194/acp-17-7529-2017, 2017.

Withnall, M., Lindelöf, E., Engkvist, O., and Chen, H.: Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction, J. Cheminformatics, 12, https://doi.org/10.1186/s13321-019-0407-y, 2020.

Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., Yin, M., Zeng, X., Wu, C., Lu, A., Chen, X., Hou, T., and Cao, D.: ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Research, 49, W5–W14, https://doi.org/10.1093/nar/gkab255, 2021.

Zhang, S., Tong, H., Xu, J., and Maciejewski, R.: Graph convolutional networks: a comprehensive review, Comput. Soc. Netw., 6, 11, https://doi.org/10.1186/s40649-019-0069-y, 2019.

Zhang, Y.-J., Khorshidi, A., Kastlunger, G., and Peterson, A. A.: The potential for machine learning in hybrid QM/MM calculations, J. Chem. Phys., 148, 241740, https://doi.org/10.1063/1.5029879, 2018.

2.4. Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks

This chapter presents a research article published in the journal Geoscientific Model Development. I am the second author of this paper and contributed to it significantly. I explored the concept of using neural networks as surrogate models prior to the collaboration and participated in research design. I wrote all code for NN surrogate modelling, including data pre-processing, model training and evaluation. I also implemented the inverse modelling test including the Metropolis-Hastings algorithm and carried out the experiment using high performance computing. I also prepared all figures related to the NN surrogate model, significantly contributed to the manuscript preparation and assisted during the revision. More detailed information on the author contributions are provided at the end of the manuscript.

Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., Krieger, U.K.: Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks, *Geosci. Model Dev.*, doi: 10.5194/gmd-16-2037-2023, (2023).

Artificial neural networks (NN) and polynomial chaos expansion were used to generate inexpensive surrogate models for the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) to predict reaction times in multiphase chemical systems. The surrogate models were fast, accurate, and robust, which suggests their applicability as sub-modules in large-scale atmospheric models. Furthermore, it was shown that NN surrogate models could be used to enable or accelerate inverse-modeling applications. These qualities make them suitable supporting tools for laboratory work in the interpretation of data and the design of future experiments. The supplement to this work can be found in appendix B3.

Geosci. Model Dev., 16, 2037–2054, 2023
https://doi.org/10.5194/gmd-16-2037-2023

Geoscientific Model Development
Open Access
EGU

Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks

Thomas Berkemeier[1], **Matteo Krüger**[1,★], **Aryeh Feinberg**[2,3,4,a,★], **Marcel Müller**[2,★], **Ulrich Pöschl**[1], **and Ulrich K. Krieger**[2]

[1]Multiphase Chemistry Department, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany
[2]Institute for Atmospheric and Climate Science, ETH Zürich, 8092 Zürich, Switzerland
[3]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland
[4]Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland
[a]currently at: Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
★These authors contributed equally to this work.

Correspondence: Thomas Berkemeier (t.berkemeier@mpic.de)

Received: 13 October 2022 – Discussion started: 20 October 2022
Revised: 15 February 2023 – Accepted: 20 March 2023 – Published: 14 April 2023

Abstract. The heterogeneous chemistry of atmospheric aerosols involves multiphase chemical kinetics that can be described by kinetic multi-layer models (KMs) that explicitly resolve mass transport and chemical reactions. However, KMs are computationally too expensive to be used as sub-modules in large-scale atmospheric models, and the computational costs also limit their utility in inverse-modeling approaches commonly used to infer aerosol kinetic parameters from laboratory studies. In this study, we show how machine learning methods can generate inexpensive surrogate models for the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) to predict reaction times in multiphase chemical systems. We apply and compare two common and openly available methods for the generation of surrogate models, polynomial chaos expansion (PCE) with UQLab and neural networks (NNs) through the Python package Keras. We show that the PCE method is well suited to determining global sensitivity indices of the KMs, and we demonstrate how inverse-modeling applications can be enabled or accelerated with NN-suggested sampling. These qualities make them suitable supporting tools for laboratory work in the interpretation of data and the design of future experiments. Overall, the KM surrogate models investigated in this study are fast, accurate, and robust, which suggests their applicability as sub-modules in large-scale atmospheric models.

1 Introduction

An accurate description of the heterogeneous chemistry of atmospheric particles requires explicit coupling of mass transport with chemical reactions (Pöschl et al., 2007; Kolb et al., 2010; Shiraiwa et al., 2014). Especially for particles containing secondary organic matter, field and laboratory experiments during the last decade showed severe transport limitations that affect chemical reactivity (Shiraiwa et al., 2011; Kuwata and Martin, 2012; Berkemeier et al., 2016). While the elementary processes are well understood, kinetic multi-layer models (KMs) describing mass transport and chemical reactions at the gas–particle interface and throughout the particle bulk are computationally expensive due to the need for spatial resolution within the particles (Pöschl et al., 2007; Shiraiwa et al., 2012; Roldin et al., 2014; Berkemeier et al., 2017; Semeniuk and Dastoor, 2020; Dou et al., 2021). For use in global or regional models, the KMs would have to be evaluated for every grid cell, time step, and particle class (size and composition). This computational volume makes

the application of KM extremely costly, if not outright impossible.

A second complicating factor for KMs is the multitude of chemical and physical input parameters, such as transport parameters or chemical reaction rate coefficients, which are often poorly constrained or unknown. Thus, in a laboratory setting, KMs are often used in an inverse-modeling approach, in which model parameters are deduced or constrained with experimental data using global optimization (Berkemeier et al., 2017; Tikkanen et al., 2019; Berkemeier et al., 2021; Wei et al., 2021; Milsom et al., 2022). However, due to the inherently coupled nature of the underlying physical and chemical processes, input parameters are often ill constrained; i.e., their numerical value cannot be uniquely determined (Berkemeier et al., 2017). This is particularly problematic when extrapolating the KMs to conditions outside the calibration range, where the calculation outcome can depend strongly on previously insensitive and thus unconstrained parameters (or combinations of parameters). Fit ensembles, i.e., arrays of multiple solutions from repeated execution of a global optimization algorithm, can be utilized to propagate the uncertainty of the global fit to conditions outside the calibration range (Berkemeier et al., 2021). Solving the inverse problem is a complex task that becomes computationally more expensive with an increasing number of uncertain model input parameters, often requiring $> 10^5$ model simulations (Xu et al., 2018). In some cases, this can be prohibitively expensive to do with a full model, and the problem is exacerbated when acquiring or evaluating fit ensembles.

Computationally inexpensive surrogate models can replace KMs in specialized tasks and help solve the issue of computational cost. These surrogate models are trained on a dataset consisting of a wide range of kinetic input parameters and the associated calculated outputs until they reproduce the KM output with the desired accuracy. Surrogate-based optimization methods are an active field of research (Booker et al., 1999; Vu et al., 2017; Xu et al., 2018). Some studies use an iterative approach, wherein the surrogate model is used to constrain the likely parameter space, and the full model is run within this likely parameter space to refine the surrogate model. Here, we illustrate the generation of surrogate models by introducing two suitable machine learning methods, namely artificial neural networks (NNs) through the Python package Keras (Gulli and Pal, 2017) and polynomial chaos expansion (PCE) with UQLab (Marelli and Sudret, 2014).

Artificial NNs represent a group of common machine learning algorithms. Their functionality is inspired by biological brains, where complex computational processes are based on comparably simple interactions of large numbers of interconnected nodes or neurons (Kröse and van der Smagt, 1996). Neural networks are commonly organized in layers, where an individual neuron obtains signals from neurons in the previous layer and maps them to a single new signal that is passed to neurons of the following layer (Almeida, 2001; Popescu et al., 2009). By systematic variation of the numerical weights of individual neuron operations, the so-called training, an NN can increase its predictive accuracy. The exact mathematical operations that are performed by neurons in specific layers and the arrangement of such layers (architecture of the NN) are determined by so-called hyperparameters. Hyperparameters can be adapted to obtain an NN that is specialized for a specific task, input data structure, or output type (Bishop, 1994; Sadeeq and Abdulazeez, 2020).

In the atmospheric sciences, NNs are used for air quality prediction, function approximation, and pattern recognition tasks (Gardner and Dorling, 1998), but their application as surrogate models for computationally expensive KMs is less well researched. Recently, popular applications of machine learning in atmospheric chemistry and physics include quantitative structure–activity relationship (QSAR) models that map molecular structures to compound properties as an alternative to time-consuming laboratory experiments or quantum mechanical calculations (Lu et al., 2021; Lumiaro et al., 2021; Galeazzo and Shiraiwa, 2022; Krüger et al., 2022; Xia et al., 2022). Holeňa et al. (2010) used surrogate models in computationally costly evolutionary optimization and successfully enhanced this approach with the application of NNs. Tripathy and Bilionis (2018) used an NN to create surrogate models for expensive high-dimensional uncertainty quantification. Other recent applications of NNs as surrogate models address chemical and process engineering (Cavalcanti et al., 2021; Esche et al., 2022) or materials science (Allotey et al., 2021). Machine-learning-based surrogate models have also found application as modules in geoscientific models, including large-scale atmospheric chemistry, transport, and climate models, to reduce computational cost in very demanding tasks such as atmospheric convection (O'Gorman and Dwyer, 2018), gas-phase and heterogeneous chemistry (Keller and Evans, 2019; Kelp et al., 2020; Sturm and Wexler, 2022), or aerosol and cloud microphysics (Rasp et al., 2018; Harder et al., 2022). These surrogate models function either as parameterizations for subgrid processes or replace the chemical integrator.

The second method applied in this work is polynomial chaos expansion (PCE), a method commonly used for uncertainty quantification (Sudret, 2008). In the PCE approach, the full model is represented as a series of suitably built, multivariate, and orthonormal polynomial functions (Marelli and Sudret, 2014). Surrogate models using PCE methods have been developed mainly within engineering fields (Ghanem and Spanos, 2003; Sudret, 2008). Several recent environmental chemistry investigations have applied PCE surrogate modeling, particularly because of its suitability for global sensitivity analysis problems (Thackray et al., 2015; Feinberg et al., 2020). The goal of global sensitivity analysis is to apportion the uncertainty in model output into contributions from the uncertainties of different model input variables, additionally considering interacting effects between input parameter uncertainties (Saltelli et al., 2008). The results from

the sensitivity analysis indicate which are the most influential input parameters that should be further constrained and may therefore be a useful tool in designing or prioritizing laboratory experiments.

2 Methods

The surrogate-modeling workflow employed in this study is shown in Fig. 1. To acquire a fast-computing surrogate model for the computationally expensive KMs, training data are first acquired by sampling outputs of the full model from the possible model parameter space. The surrogate models are trained with Keras and UQLab on this data and are validated by comparison with a test dataset of full model output.

2.1 Kinetic multi-layer model KM-SUB

In this study, we employ the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB; Shiraiwa et al., 2010), but the statistical methods could be used with any process model. KM-SUB describes mass transport and chemical reaction at the surface and in the bulk of aerosol particles by solving a set of ordinary differential equations. The model explicitly treats gas diffusion, surface and bulk accommodation of gas molecules, surface–bulk exchange, and bulk diffusion, as well as chemical reaction at the surface and in the bulk of aerosol particles. For a schematic depiction of the processes and compartments of KM-SUB, see Fig. B1.

For the model calculations in this study, we chose a general model scenario of a single volatile reactant X (e.g., OH, O_3, NO_3) reacting with a single non-volatile reactant Y at the surface and in the bulk of the aerosol particle. The input parameters of KM-SUB resulting from this scenario include initial concentrations, reaction rate coefficients, and diffusion coefficients (Table 1). The outputs of KM-SUB are concentration profiles over space and time, but in this study, we summarized KM-SUB output as the total number of Y in a single aerosol particle at time t ($N_{Y,t}$). To minimize data storage requirements, we reduce the full KM-SUB time series to three output values, the time required to reach 90 %, 50 % (i.e., the chemical half-life), and 10 % of $N_{Y,0}$ by interpolation of the primary model output. The inputs and outputs of KM-SUB are then log-transformed. For the NN application, all input parameters and model outputs are additionally normalized to the interval $[0:1]$. Outputs are normalized by dividing by the longest time recorded to reach 10 % of $N_{Y,0}$.

For each input parameter of KM-SUB, individual parameter boundaries are defined, representing a wide array of reactants and scenarios that can be found in either the atmosphere or in laboratory experiments (Table 1). As these ranges cover orders of magnitude, they are assumed to follow log-uniform probability distributions. The parameter space includes liquid to semisolid particles (as expressed by the reactant diffusivities) from 50 nm to 100 µm in size. Reaction rate coefficients range from reactivity close to the diffusion limit, typical for the OH radical (1×10^{11} cm^3 s^{-1}), down to reactions that are 9 orders of magnitude slower, and they may be associated with reactions involving ozone. The volatile reactant X is given a large variability in terms of partitioning properties (as expressed by surface accommodation coefficient $\alpha_{s,0}$ and desorption lifetime τ_d) and solubility properties (as expressed by the Henry's law coefficient), each varying over several orders of magnitude. The initial concentration of non-volatile reactant Y ranges from 10^{19} to 2×10^{21} cm^{-3}, which, for an organic substance with a molar mass of 250 g mol^{-1}, corresponds roughly to a molar fraction from 0.5 % in relation to pure particles. The concentration of X in the gas phase is held constant over a simulation and is varied between simulations from a few parts per billion (10^{11} molec. cm^{-3}) to about 200 parts per million (5×10^{15} molec. cm^{-3}). For the explicit treatment of gas diffusion, we assume a temperature of 298 K and a fixed diffusion coefficient of 0.14 cm^2 s^{-1}.

2.2 Acquisition of training data

The KM is used to generate a training dataset for the surrogate models by randomly sampling parameters in log-uniform space within their associated boundaries. The number of KM samples obtained in this study is about 4.3×10^6 and required supercomputing. A random set of 1000 samples is removed from the dataset and withheld from model training for the visualization and validation of fully trained surrogate models. We refer to this set of data as test data.

As not only the computational effort of sampling training data but also the time required for surrogate-model creation increases with the size of the training dataset, the surrogate-model performance is tested on different fractions of the total training dataset in order to find an optimal or sufficient computational expense for a given application (Table 2). Note that the PCE method is only applied to the first nine fractions (50–20 000) due to the computational expense of the method at higher training-set sizes.

2.3 Neural network (NN)

The neural network architecture employed in this study is a multi-layer perceptron (MLP), in which nodes are organized in consecutive layers. MLPs are characterized by a chosen number of so-called hidden layers that connect the visible input and output layers. Each node in a layer is connected with each node in the previous and following layers (fully connected layers). We test MLPs consisting of up to five hidden layers with variable numbers of neurons to determine a network architecture that suits the specified task. A detailed mathematical description of MLP functionality and architecture is given in Appendix A1. The processes of hyperparameter tuning, tested ranges, and suggested values for individual hyperparameters are described in Appendix A2. We apply 5-fold cross-validation to avoid over-fitting of the trained mod-

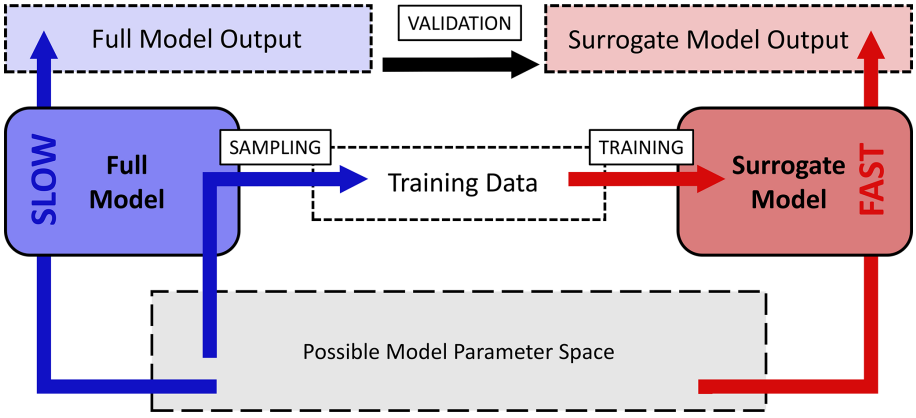

Figure 1. Workflow chart for the surrogate-modeling process employed in this study. The possible or desired model parameter space (gray) is sampled with the slow-computing full model (blue) to acquire training data consisting of model input–output pairs. Training data are used for training of a fast-computing surrogate model (red). Surrogate models are validated by comparison of full-model output and surrogate-model output.

Table 1. KM-SUB input parameters with lower and upper boundaries and fit parameters to the laboratory dataset.

Parameter	Lower boundary	Upper boundary	Description
k_{SLR}	1.0×10^{-15}	1.0×10^{-8}	Rate coefficient of X+Y surface reaction ($cm^2\ s^{-1}$)
k_{BR}	1.0×10^{-20}	1.0×10^{-11}	Rate coefficient of X+Y bulk reaction ($cm^3\ s^{-1}$)
$D_{b,X}$	1.0×10^{-11}	1.0×10^{-5}	Bulk diffusion coefficient of X ($cm^2\ s^{-1}$)
$D_{b,Y}$	1.0×10^{-12}	1.0×10^{-6}	Bulk diffusion coefficient of Y ($cm^2\ s^{-1}$)
$H_{cp,X}$	5.0×10^{-6}	5.0×10^{-3}	Henry's law solubility coefficient of X ($mol\ cm^{-3}\ atm^{-1}$)
$\tau_{d,X}$	1.0×10^{-9}	1.0×10^{-2}	Desorption lifetime of X (s)
$\alpha_{s,0,X}$	1.0×10^{-4}	1	Surface accommodation coefficient of X on an adsorbate-free surface (unitless)
r_p	2.5×10^{-6}	1.0×10^{-3}	Particle radius (cm)
$[X]_{g,0}$	1.0×10^{11}	1.0×10^{15}	Initial gas-phase number concentration of X (cm^{-3})
$[Y]_{b,0}$	1.0×10^{19}	2.0×10^{21}	Initial bulk number concentration of Y (cm^{-3})

els during hyperparameter tuning (Stone, 1974; Wong and Yeh, 2020).

2.4 Polynomial chaos expansion (PCE)

The PCE surrogate-modeling approach will be briefly summarized here. For more technical descriptions, the reader can refer to Sudret (2008) and Le Gratiet et al. (2017). The principle behind PCE is that the model output Z is decomposed into an infinite series as follows (Ghanem and Spanos, 2003):

$$Z = \sum_{\alpha \in \mathbb{N}^M} y_\alpha \psi_\alpha(X), \quad (1)$$

where M is the number of model input variables, α is a multi-index that defines the variable components of the polynomials, y_α refers to coefficients, and ψ_α refers to orthonormal polynomials of either one input variable (representing first-order effects) or multiple input variables (representing interacting effects). The type of orthonormal polynomial in Eq. (1) depends on the probability distribution of the input parameters, with uniform probability distributions being represented by Legendre polynomials and Gaussian probability distributions being represented by Hermite polynomials (Xiu and Karniadakis, 2002). In practice, Eq. (1) is truncated by restricting the maximum degree of the polynomials. We calculate PCE coefficients (y_α) using the implementation of least-angle regression (Blatman and Sudret, 2010) from the open-source MATLAB-based software UQLab (Marelli and Sudret, 2014). This software allows degree-adaptive calculation of the PCE, meaning that PCE models can be constructed from degree 1 to a maximum selected degree, which we set to 14. If the cross-validation error of the model does not decrease over two steps in degree, the algorithm stops, and the PCE with the lowest cross-validation error is selected. All PCEs calculated for this study are equal to or below degree 7 (Table A1).

2.5 Global sensitivity analysis

In global sensitivity analysis, Sobol' indices describe the contribution of uncertainty from each input parameter and interactions between input parameters (Sobol', 2001). The

variance D of the model output Z is decomposed into partial variances as follows:

$$D = \mathrm{Var}(Z) = \sum_{i=1}^{M} D_i + \sum_{1 \le i < j \le M} D_{ij} + \text{higher order terms}, \quad (2)$$

i.e., the sum of first-order partial variances (D_i), second order partial variances (D_{ij}), and higher order terms. Sobol' indices (S) are calculated by normalizing the partial variances by the total variances, e.g., $S_i = \frac{D_i}{D}$ for the first-order contribution of ith input parameter and $S_{ij} = \frac{D_{ij}}{D}$ for the contribution of the interaction between the ith and jth input parameters to the model uncertainty. In order to summarize the overall influence of a specific input parameter, including interactions, a total Sobol' index (S_i^T) can be calculated:

$$S_i^T = S_i + \sum_{j \neq i}^{M} S_{ij} + \sum_{j \neq i} \sum_{\substack{k \neq i \\ k \neq j}} S_{ijk} + \ldots + S_{ij \ldots M}. \quad (3)$$

Given the similarities between the PCE and Sobol' decompositions, the Sobol' sensitivity indices can be calculated analytically from the PCE coefficients rather than with Monte Carlo sampling (Sudret, 2008). This eliminates a potentially computationally expensive step of the sensitivity analysis process using other surrogate models.

2.6 Acquisition of fit ensembles

With the trained NN model, we illustrate and test the application of surrogate models in inverse-modeling approaches with KM-SUB. Six sets of experimental data of the well-studied oleic acid ozonolysis heterogeneous reaction system (Hearn and Smith, 2004; Ziemann, 2005; Gallimore et al., 2017; Berkemeier et al., 2021) are used to determine kinetic parameter sets that minimize the mean squared (absolute) logarithmic error (MSLE) between model and experiments. More details about the specific optimization problem can be found in Appendix B.

$$\mathrm{MSLE} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n} \sum_{j=1}^{n} (\log_{10}(z_{ij}) - \log_{10}(y_{ij}))^2, \quad (4)$$

where N is the number of experimental datasets, n is the number of data points in each set, z_{ij} is the model output, and y_{ij} is the value for experiment i and data point j. As this optimization problem does not offer a unique solution (Berkemeier et al., 2021), the aim is not to find a best-fitting parameter set but rather to find a fit ensemble, i.e., an array of parameter sets that all yield a sufficient agreement of the associated KM-SUB outputs with the experimental data. The fit ensemble then not only represents the ranges to which kinetic input parameters could be constrained but is also a means of assessing the uncertainty associated with the KM-SUB model fit when extrapolating the model to environmental conditions outside the calibration range (Berkemeier et al., 2021). For both purposes, the number of model fits in the ensemble must be sufficiently large to fully grasp the remaining model flexibility. The process of determining such a large set of fits can be computationally expensive. A surrogate model can either fully replace the KM or assist in the fitting process by suggesting sampling points.

In this study, we evaluate the benefits of surrogate-model-supported sampling by comparing the distribution of KM-SUB output MSLE for three different sampling approaches within the parameter boundaries presented in Table 1. These approaches are

- random log-uniform sampling,
- Metropolis–Hastings algorithm (MHA)-directed sampling,
- NN-suggested sampling.

We choose an MSLE of 0.016 as representing sufficient agreement between model and experiment. For NN-suggested sampling, we perform a random log-uniform screening of the NN surrogate model in batches of 10 000 samples until we find 5000 NN-suggested fits with MSLE < 0.016; we then feed these pre-sampled parameter sets into KM-SUB. We refer to KM-SUB outputs with an MSLE below 0.016 as fits.

As a directed-sampling approach, we apply the Metropolis–Hastings algorithm (MHA), a common Markov chain Monte Carlo method to sample multivariate distributions with high numbers of dimensions (Chib and Greenberg, 1995; Robert and Casella, 1999). We determine the maximum step size of the MHA by basic testing on smaller subsets, and we find that a step size of 0.1 is a good compromise between a high acceptance ratio and sufficient exploration of the entire parameter space. Here, step size is defined as the maximal parameter variation as a fraction of the total logarithmic parameter space. For comparability of the aggregate computational effort, each sampling is performed on an 11th Generation Intel(R) Core(TM) i5-1145G7 CPU with 2.6 GHz.

2.7 Hardware and software tools

Training-data acquisition with KM-SUB was performed in MATLAB on the high-performance computing system Cobra at the Max Planck Computing and Data Facility (MPCDF). Model training of the NN was performed in Python 3.6 using the packages Keras 2.3.0 (Chollet et al., 2015), TensorFlow 1.14.0 (Abadi et al., 2015), scikit-learn 0.22.1 (Pedregosa et al., 2011), NumPy 1.18.1 (Harris et al., 2020), and pandas 0.25.3 (McKinney et al., 2010). Each model training was conducted on one NVIDIA GeForce GTX 1080 Ti on the high-performance computing cluster Mogon of the Johannes

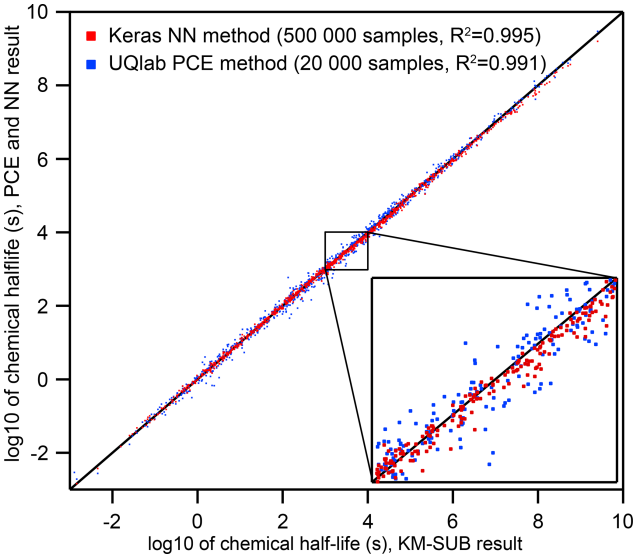

Figure 2. Comparison of the two surrogate models predicting the chemical half-life for heterogeneous chemistry on aerosol particles for a wide range of KM-SUB output ($N = 1000$ – test dataset not part of the training dataset). The surrogate models were trained on 20 000 (PCE) and 500 000 (NN) KM-SUB data samples, respectively. Training times of models with this complexity fall below an upper feasibility range on a personal computer within a few days of time. The inset shows a magnified section and spans from chemical half-lives of 10^3 s ($\approx$ 15 min) to 10^4 s ($\approx$ 3 h), a common range for laboratory experiments.

Gutenberg University Mainz. For the PCE and sensitivity analysis, we use the MATLAB-based software UQLab 1.3 (Marelli and Sudret, 2014), which provides a framework for surrogate modeling and uncertainty quantification. We performed PCE calculations on ETH Zurich's high-performance computing cluster Euler, using four CPUs per PCE calculation and up to 45 GB of memory for the largest sample size (20 000).

To determine training times of the NN and PCE models, the required time for sample loading and file writing is disregarded, and only the true training time is reported. For the PCE method, the time to reach 90 %, 50 %, and 10 % of the initial amount of Y, $N_{Y,0}$, is calculated by three separate models, and training times are added to yield a combined training time for each training sample size. For the NN method, one model can be set to return multiple values as output; thus, a single model is used for each dataset to predict all three output values collectively.

3 Results and discussion

3.1 Surrogate-model training, accuracy, and speed

Neural networks (NNs) and polynomial chaos expansion (PCE) are used to emulate the reaction time of a multiphase chemical system in KM-SUB. Table 2 displays the test set errors and training times of surrogate models with the NN and PCE methods as a function of training-dataset size. The best surrogate models achieve mean square errors (MSEs) for logarithmic reaction times of 0.0049 for the NN method and 0.0137 for the PCE method. This corresponds to correlation coefficients R^2 of 0.995 and 0.991, respectively. Figure 2 shows that these optimal versions for both surrogate models track the chemical half-life in the test dataset remarkably well. The MSE of test predictions is very similar between both approaches for the same training-dataset size. Error variance of the five cross-validation NN models for the unseen test data is very low at 2.98×10^{-6}, indicating little to no over-fitting. We found no significant correlations between surrogate-model error and the values of the 10 model input parameters (Fig. S1 in the Supplement).

For dataset sizes above 2000, the PCE model requires much more training time than the NN model. However, note that these training times of individual NN models disregard the necessity of hyperparameter tuning. While hyperparameter tuning is not required in an already established application, the total computation times of NN surrogate-model training and hyperparameter tuning can be 2 orders of magnitude larger, depending on the extent of the hyperparameter tuning that is performed. Hence, the use of an NN method is advisable when a large amount of training data are easily available and when model accuracy is of high importance.

The PCE method, on the other hand, is limited in terms of training-dataset size ($\leq$ 20 000) as a result of the calculation time and memory requirements in MATLAB. The PCE method is thus a good choice if the training dataset is small or if its acquisition is time limiting and when time-consuming hyperparameter tuning is not desired.

Both surrogate models calculate new output data orders of magnitude faster than the full model, KM-SUB. The computation time of KM-SUB lies on the order of a few seconds per model run, while both the PCE and NN methods can generate large arrays of 10 000 individual surrogate-model solutions in under 1 s.

3.2 Prediction of chemical loss and half-life

Figure 3 visualizes the accuracy of the surrogate models (training-set sizes 20 000 for PCE method and 500 000 for NN method) by generating five concentration–time curves from various input parameter combinations and comparing them to the full KM-SUB model. Input parameter sets were arbitrarily selected from the test set so that the results were spaced out homogeneously across KM-SUB chemical half-

Table 2. Training times of surrogate models with the NN and PCE method.

Training data set size	MSE of NN test predictions	NN training time (s)	MSE of PCE test predictions	PCE training time (s)
50	1.03	2	1.44	3
100	0.718	2	0.328	3
200	0.398	3	0.313	4
500	0.172	7	0.196	5
1000	0.144	14	0.132	20
2000	0.104	28	0.078	144
5000	0.049	102	0.039	4232
1×10^4	0.025	67	0.022	3.28×10^4
2×10^4	0.014	260	0.014	2.17×10^5
5×10^4	0.010	326		
1×10^5	8.6×10^{-3}	657		
2×10^5	6.7×10^{-3}	961		
5×10^5	4.9×10^{-3}	3250		
1×10^6	6.6×10^{-3}	4097		
2×10^6	7.3×10^{-3}	6477		
4.3×10^6	5.9×10^{-3}	1.64×10^4		

lives. We see that, over the wide range, both surrogate models closely represent the KM output, with the NN method slightly outperforming the PCE method as result of the larger training-set size.

Note that both methods are able to produce relatively good surrogate models (MSE ≈ 0.1) from only 1000 training-data samples (Table 2), which, depending on the user's application, may already be accurate enough. We conclude that KM-SUB is a rather well-behaved model and that it is suitable for these surrogate-modeling techniques.

3.3 Global sensitivity analysis with surrogate models

An advantage of using a PCE surrogate model is that the Sobol' sensitivity indices can be extracted analytically (Sudret, 2008). We present the global sensitivity analysis for the 50 % lifetime (i.e., the chemical half-life) PCE model in Fig. 4. We can differentiate between first-order effects of a model input parameter, wherein the parameter alone influences the output, and interaction effects, wherein combinations of parameter values influence the output. In Fig. 4, first-order effects dominate the total effect, accounting for 88 % of the model variance. Using the total Sobol' indices (S^T) as a metric, we can assess the overall influence of individual model parameters on the uncertainty of the model output. The input parameters with the largest influence on the chemical half-life of Y are the initial gas-phase concentration of X ($[X]_{g,0}$, $S^T = 0.36$) and the radius of the particle (r_p, $S^T = 0.22$). Certain parameters have a very low influence ($S^T \leq 0.05$) on the chemical half-life, including the accommodation coefficient ($\alpha_{s,0,X}$), the initial concentration of Y ($[Y]_{b,0}$), and the bulk diffusion coefficient of Y ($D_{b,Y}$). This means that variations in these parameters will, in many cases, not have a large effect on the chemical half-life, indicating that it will be difficult to constrain these parameters with measurements. Sensitivity analysis is thus a useful tool to understand model behavior and to identify parameters which have the largest influence on model output.

It has to be noted that a low global sensitivity across the entire input parameter space does exclude the possibility that pockets in the parameter space exist where either of these parameters are very influential. Constraining the input parameter space to smaller subsets can constrain the model to special kinetic regimes or limiting cases that exhibit characteristic profiles of parameter sensitivity (Berkemeier et al., 2013).

In most laboratory experiments, the particle radius and the initial concentration of X are known values. By fixing these parameters in the sensitivity analysis, a substantial fraction of the model variance is eliminated, and other unknown parameters account for a more significant fraction of the overall model variance. To demonstrate how the importance of parameters varies over different experimental conditions, we conducted sensitivity analyses by sampling the PCE surrogate model for specified values of $[X]_{g,0}$ and r_p (Fig. 5a). Certain input parameters are consistently important across the range of experimental conditions, e.g., oxidant diffusivity ($D_{b,X}$) and solubility ($H_{cp,X}$). Other parameters, including k_{BR} and $\tau_{d,X}$, have varying influences depending on the experimental conditions. For example, at a high $[X]_{g,0}$ and for large r_p, the total Sobol' index of $\tau_{d,X}$ is 0.14. Accordingly, the upper panel of Fig. 5b shows that the chemical half-life of Y only decreases slightly with increasing $\tau_{d,X}$. In contrast, at low $[X]_{g,0}$ and for small r_p, the total Sobol' index increases to 0.31. In the lower panel of Fig. 5b, the chemical half-life of Y shows a stronger dependence on $\tau_{d,X}$. This can be under-

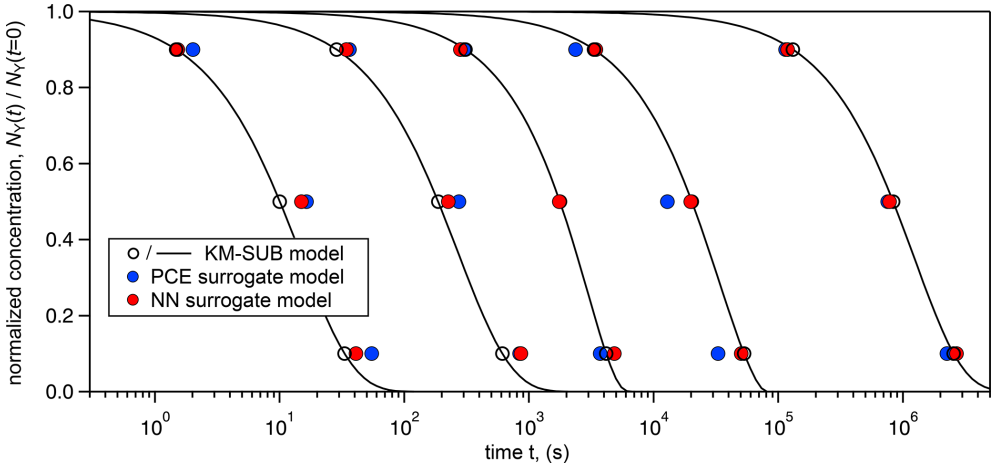

Figure 3. Comparison of time-dependent output of the surrogate models (PCE method – blue markers; NN method – red markers) with KM-SUB model output (solid black lines) for five arbitrarily chosen KM-SUB runs spanning seconds to weeks of reaction time. The surrogate models' predicted time for depletion of 10 %, 50 %, and 90 % of reactant Y in the aerosol phase. KM-SUB output at these three stages is highlighted with open black markers.

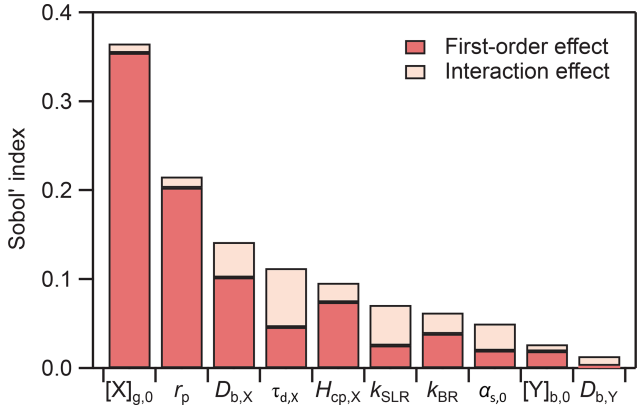

Figure 4. Results of global sensitivity analysis showing Sobol' sensitivity indices for the chemical half-life PCE model.

stood because, for small particles, surface processes are more important, and the surface concentration of X depends on its lifetime for desorption, especially at low gas-phase concentrations. This information could be potentially useful for an experimental researcher, as it shows that experiments at low $[X]_{g,0}$ and small r_p could be more helpful for constraining $\tau_{d,X}$ than experiments under other experimental conditions.

These calculations would have been very time consuming when carried out with the full KM. Hence, the combination of surrogate modeling and sensitivity analysis is a helpful yet underutilized tool for designing experiments that are best suited to constraining certain model parameters.

3.4 NN-supported global optimization

Utilizing the NN surrogate model, we illustrate the accelerated acquisition of parameter sets associated with KM-SUB outputs in good agreement with experimental data, which is the key step in inverse-modeling and optimization approaches. While uncertainty is introduced by surrogate models, their predictions can be obtained orders of magnitude faster than regular KM-SUB calculations. The uncertainty introduced by the NN method can be minimized by additional sampling of a much smaller number of parameter sets with the KM. Re-sampling of NN-suggested solutions with the KM can avoid collection of false-positive fits (i.e., meeting the conditions for a fit in the NN model but not in KM-SUB), and sampling in close vicinity of NN-suggested solutions might avoid false-negative fits (i.e., not meeting the conditions for a fit in the NN model but in KM-SUB).

We perform random parameter sampling in log-uniform space using the boundaries presented in Table 1, and we find about 5000 NN-suggested fits in 1.84×10^7 parameter sets (0.027 % acceptance), requiring a total of 13 847 s (< 4 h). A comparable calculation with KM-SUB would take years on a desktop computer or days on a supercomputer. In contrast, re-sampling of the NN-suggested fits with KM-SUB to avoid false-positive fits is time-consuming but feasible. The time required for sampling 5000 kinetic parameter sets (i.e., 5000×6 runs in KM-SUB) on a desktop computer ranges from 51 646 s (≈ 14 h) for NN-suggested sampling to 103 530 s (≈ 29 h) for random log-uniform sampling. The differences may be a result of the fraction of parameter sets where differential equation calculations of the KM require a very long time to terminate. They are often associated with very long reaction times and thus with large MSLEs.

Figure 6 shows the distributions of KM-SUB output MSLE for three different sampling methods: log-uniform random sampling, MHA-directed sampling, and NN-suggested sampling (Sect. 2.6). The NN-suggested sampling method greatly outperforms both random and MHA-

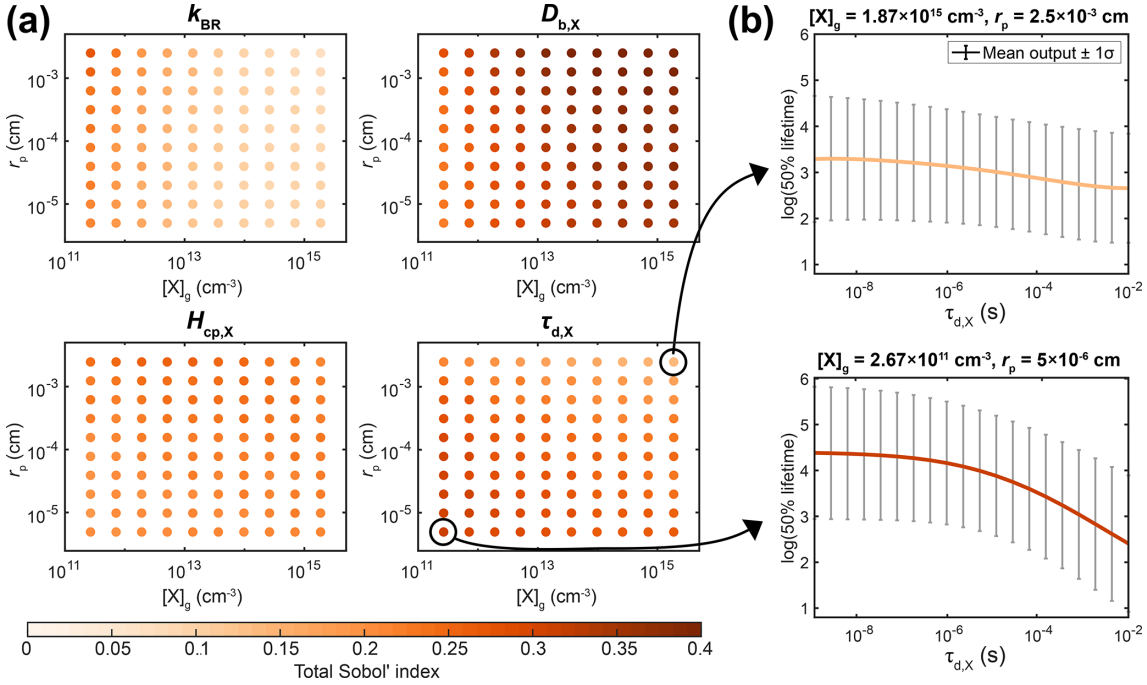

Figure 5. Detailed sensitivity analysis with the PCE method as a function of experimental conditions, i.e., the gas-phase concentration of X ($[X]_{g,0}$) and particle radius (r_p). **(a)** Total Sobol' indices of four KM input parameters: bulk reaction rate coefficient of X and Y (k_{BR}), bulk diffusion coefficient of X ($D_{b,X}$), solubility coefficient of X ($H_{cp,X}$), and desorption lifetime of X ($\tau_{d,X}$). **(b)** Relationship between the value of $\tau_{d,X}$ and the chemical half-life of Y for two selected experimental conditions.

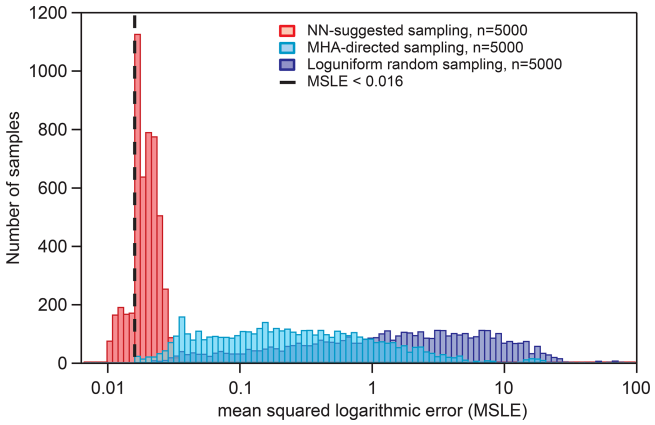

Figure 6. Distribution of KM-SUB output MSLEs for three different sampling methods in comparison with six sets of experimental data, as described in Sect. 2.6. The dashed vertical line represents the threshold used for the acquisition of NN-suggested fits (MSLE < 0.016). The maximum step size for the MHA-directed sampling is 0.1.

directed sampling. The number (fraction) of KM-SUB outputs with an MSLE < 0.016 is 1602 (32.04 %) for NN-suggested sampling, 21 (0.42 %) for directed KM-SUB sampling, and 3 (0.06 %) for random sampling.

Figure 7 compares the fitting parameter space of 5000 fits obtained with KM-SUB (panel a) and the NN surrogate model (panel b), exemplary for four kinetic parameters in a so-called scatter plot matrix. The off-diagonal elements in each matrix show bivariate scatter plots (top right) or density plots (bottom left) depicting the relationship of two kinetic parameters within the fit ensemble. The diagonal elements are histograms showing frequency distributions of the individual parameters. The two scatter plot matrices show a clear resemblance in terms of the fit parameter spaces between the surrogate model and the original KM. Much like the scatter plots of the original-model fits, the scatter plots of the surrogate-model fits can be used to identify areas that will not produce a fit to experimental data. For example, there are no fits with a slow surface reaction rate coefficient (k_{SLR}) and a high oxidant solubility ($H_{cp,X}$). However, some features in the scatter plots of the surrogate model deviate from those in the scatter plots of the original KM. We can visually identify areas in the scatter plots that indicate false-positive fits, i.e., areas that are only occupied in the plots for the surrogate model. An absence of density in other areas, compared to the plots for the original model, suggests the existence of false-negative fits.

Whether it is worthwhile to train a surrogate model for a given optimization task depends strongly on the complexity of the KM and the difficulty of the optimization problem. For every application, there is a break-even point where the computational expense of training a surrogate model is compensated for by the acceleration of the optimization task(s).

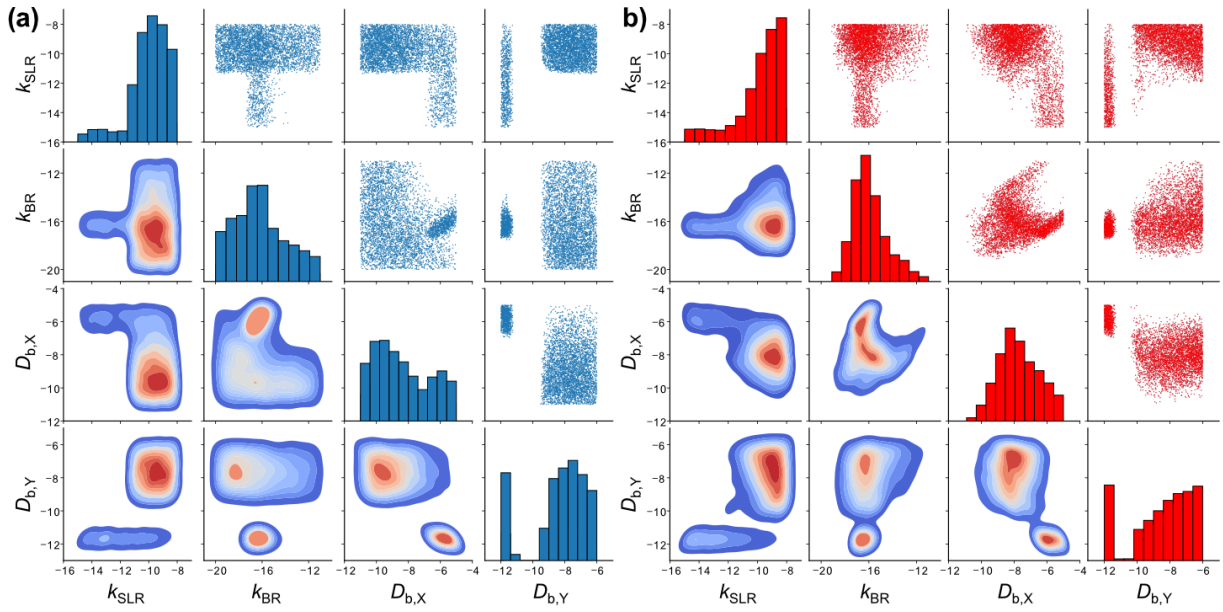

Figure 7. Scatter plot matrices of the fitting parameter space of 5000 fits to six experimental datasets of the ozonolysis of oleic acid aerosols (Appendix B) obtained with **(a)** KM-SUB and **(b)** the NN surrogate model. Shown are four out of seven optimized kinetic parameters. The diagonal elements are histograms showing the distributions of the individual fit parameter densities. The off-diagonal elements are scatter plots (top right) or densities (bottom left) of solutions for all possible combinations of two kinetic parameters. The KM-SUB fit ensemble originates from the application of the MHA with a step size of 0.1 and the NN fit ensemble from log-uniform random sampling.

In this study, the computational effort required to obtain the training data for the best-performing surrogate model (500 000 KM-SUB sample runs) would only find ∼ 350 fits if we had directed this initial sampling effort into fit acquisition using only KM-SUB. This is due to the very low fraction of fits (0.42 %) without the aid of surrogate models and because KM-SUB has to be evaluated six times, once for each laboratory dataset. Thus, if the uniqueness of an optimization result must be determined, large amounts of laboratory data are available, or simply, if global optimization of the same model is required on a regular basis, training of a surrogate model for this task quickly becomes worthwhile.

4 Conclusions

In this study, we illustrate the application of artificial neural networks (NNs) and polynomial chaos expansion (PCE) to generate fast surrogate models for computationally expensive kinetic models (KMs). As a template KM, we use the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB; Shiraiwa et al., 2010), but the presented methods can equally be applied to other process models. To reduce data storage requirements in sampling and to simplify emulation, the complex model output of KM-SUB, i.e., the concentration profiles of all reactants over space and time, is reduced to the reaction time of the system to reach a certain reaction progress, as this is a typical observable in laboratory experiments. We note that other derivatives of KM-SUB model output, such as the uptake coefficient of the reactant gas to the aerosol surface, could be chosen depending on the target application of the surrogate model. Emulation of the entire KM-SUB output may be feasible and could be facilitated by data compression methods such as auto-encoders, singular-value decomposition, or principal component analysis.

Our findings suggest that, after an initial investment of computational effort for training-data sampling and model training, both methods yield models with very good correlations to KM-SUB outputs ($R^2 > 0.99$). Furthermore, we provide examples for the application of such surrogate models for inverse modeling and kinetic parameter optimization: global sensitivity analysis with the PCE method and acceleration of global optimization with the NN method. The results indicate that surrogate models can aid in costly optimization tasks or help to select environmental system parameters for experiments that significantly constrain KM solution space and thus global fit uncertainty.

It is important to note that errors of surrogate models are not simply based on a random deviation of surrogate-model predictions from the values of the original KM but on a divergence of the predicted parameter hyper-surface in specific areas, for instance where training data are sparse. False-positive fits, i.e., parameter sets with associated surrogate-model predictions in better agreement with experimental data as the delineated KM output, can simply be eliminated by resampling the parameter sets in question with the KM (Fig. 6).

On the other hand, false-negative fits and their implications for inverse-modeling approaches are much more difficult to address. While optimization hyper-surfaces can be scanned relatively quickly with a surrogate model, this is not the case for the much slower KM. Scatter plot matrices of the fitting parameter space are a valid means of identifying areas that are occupied by false-negative fits, but a proper comparison (Fig. 7) requires computationally costly sampling with the KM.

Another potential application of surrogate models for KM is their utilization as modules in large-scale chemical transport models. As such models often require many calls of the respective module, direct use of models such as KM-SUB, where calculation time is on the order of seconds, is not feasible. Trained, predictive surrogate models, however, can easily be integrated into existing modeling programs. This potentially allows the coupling of small-scale kinetic process models with large-scale chemical transport models for the simulation of weather, pollution, and climate. Kelp et al. (2022) recently demonstrated acceleration of a global model with an online-learned NN as a chemistry module. The machine learning models presented in this study could be embedded in existing FORTRAN code in a similar fashion.

Appendix A: Neural networks

A1 Neural network architecture

A multi-layer perceptron (MLP) represents a complex, non-linear function that maps an input to an output vector. Each individual node in an MLP represents a non-linear function, mapping from the sum of its inputs to an output, which is passed to the following interconnected nodes. Connections between nodes are associated with weights that are optimized during training in order to reduce model output error in comparison with the dataset values. For this purpose, an optimization algorithm is used to minimize a previously defined loss function based on the final model output. In their entirety, these weights determine the output of the MLP based on a specific input, and their adaptation, based on the training data, represents the learning process. The following equations show the principal mathematical functionality of neurons in an MLP, as elaborated upon in Kröse and van der Smagt (1996):

$$s_k(t) = \sum_j w_{jk}(t) y_j(t) + \Theta_k(t), \tag{A1}$$

where $s_k(t)$ is the effective input of a neuron k at time t, w_{jk} is the weight between neuron j and k, and $y_j(t)$ is the activation of the previous neuron j. This equation represents the input of a single computational node in the NN, which is based on the activation of connected previous nodes and the associated (trained or initialized) weights. $\Theta_k(t)$ represents an offset term. Of this so-called propagation rule, different adaptations have been proposed (Feldman and Ballard, 1982).

Table A1. Employed polynomial degree of the three PCE models (90 %, 50 %, and 10 % lifetime) as function of training-dataset size.

Dataset size	PCE 90 % $N_{Y,0}$	PCE 50 % $N_{Y,0}$	PCE 10 % $N_{Y,0}$
50	3	3	3
100	2	2	2
200	3	3	3
500	3	3	3
1000	4	4	4
2000	5	5	5
5000	7	6	6
10 000	7	7	7
20 000	7	7	7

$$y_k(t+1) = F_k(y_k(t), s_k(t))) \tag{A2}$$

This equation introduces the activation function of neuron k (F_k) that maps the neuron input $s_k(t)$ and the current activation $y_k(t)$ of the neuron to a new activation value. A common type of the activation function is a sigmoid-like function, as shown in the following equation:

$$y_k = F(s_k) = \frac{1}{1 + e^{-s_k}}. \tag{A3}$$

The definition of the input and activation functions of neurons determines the output of any NN given a specific input and a set of weights. NN model training or learning describes the process of iterative modification of weights in order to shift the output in a desired way. In most cases, this desired shift is a reduction of error towards the associated predictable values in the underlying population associated with the training data. If the model is well fitted to the training data but predicts further data of the same population with much larger error, it is referred to as over-fitted. Over-fitting describes overall ill generalization of an NN model. A common learning rule for nodes, the so-called perceptron learning rule, is shown in the following equation:

$$w_i(t+1) = w_i(t) + \Delta w_i(t). \tag{A4}$$

In order to adjust the weights, the output of the NN is compared with the associated training-data values. If the prediction is inaccurate, the modification Δw_i is applied. For this iterative adjustment to be target-oriented, an optimizer is necessary to reduce the prediction error of the NN during training. Different optimizers are commonly used in machine learning applications, such as simple gradient methods like stochastic gradient descent (SGD), where an estimate of the gradient (the direction of the steepest descent)

Table A2. Descriptions and tested ranges for neural network hyperparameters used in the Python package Keras, as well as the recommendation based on our best-performing model.

Parameter	Lower boundary	Upper boundary	Recommended value	Description
Number of hidden layers (HL)	1	5	2	The number of hidden layers in the NN determines network size and strongly impacts computational cost
Activation functions[1]	"relu", "elu", or "sigmoid"		All "relu"	Activation function for the neurons in each of the hidden layers
Number of neurons[1]	4	4096[2]	(4096, 4096)	Also determines NN model size – large numbers are associated with increased computational cost and risk of over-fitting
Dropout rate[1]	0.1	0.9	0.5	The model ignores this fraction of all weights in this HL during training[3]
Optimizer	"Adam", "Nadam", "SGD", or "RMSprop"		"Adam"	Optimizer for training process
Batch size	4	128	16, depending on learning rate[4]	The number of training samples handled by model in a batch
Epochs	4	60	32, until model loss converges	Number of training epochs
Learning rate	10^{-5}	10^{-1}	0.0001	Extent of variation of weights in attempt to decrease error
Decay	0	0.9	0	Decrease of learning rate throughout training epochs

[1] Must be set for each individual HL. [2] Larger numbers of neurons per layer lead to over-fitting and, with the hardware setup in this study, memory limitations on the computational cluster. [3] A random fraction of weights obtained in previous training, determined in size by this parameter, is not considered during the current training. This handicap or restriction ensures that the model is not capable of just saving or learning all the inputs and associated outputs in the training dataset throughout multiple training epochs (as this would be over-fitting). [4] A larger batch size decreases training time and requires higher learning rates.

along with a selected step size determines the variation of input parameters in the current step. As information in a feed-forward NN, like an MLP, is only passed in one direction, a method called back propagation is used to determine the direction and amount of weight adjustment in previous NN layers based on the error of the final prediction. More in-depth explanations, definitions, and examples for back propagation and optimization throughout the learning process can be found in Rumelhart et al. (1995) and Hecht-Nielsen (1992); for further information regarding MLPs and NNs in general, see Almeida (2001) or Popescu et al. (2009).

A2 Hyperparameter tuning

Comprehensive hyperparameter tuning is conducted every time a surrogate model is trained on different training data. In this study, we focus on the investigation of dataset sizes and training times. For this reason and because our application of NN is not very common and only a small amount of information regarding successful model architectures and hyperparameters is available, only basic, plain network architectures are tested (i.e., MLPs with up to five fully connected hidden layers and up to 4096 neurons in each of the layers). We perform hyperparameter tuning in three steps, aiming for an optimization of number of layers, layer activation functions, learning rate, and batch size in the first step; number of neurons in each layer in the second step; and dropout rate in the third step. For each step, we apply an adapted grid search where multiple well-performing hyperparameter sets from the previous step are extended by variation of the additionally optimized hyperparameter of the current step.

We performed relatively comprehensive hyperparameter tuning with 60 to 120 hyperparameter sets for each data subset, with each tested set resulting in five models for the individual cross-validation folds. Sets of hyperparameters that lead to well-performing models can, to some extent, be adopted for approaches with similar preconditions regarding the number of inputs and outputs or training-dataset size. For a similar approach, we recommend a basic hyperparameter tuning with at least 10 hyperparameter sets and 5-fold cross-validation. The best models are selected by the average test set error of the five models for each of the cross-validation

folds using the mean squared error. The ranges of hyperparameters tested in this study are listed in Table A2 along with the hyperparameter values of the best-performing models for large datasets.

Besides NNs from the Keras package, other deep-learning algorithms tested for this study are the random forest regressor, the decision tree regressor, the SGD regressor, the ridge regressor, least absolute shrinkage and selection operator (LASSO), logistic regression, and the MLP regressor, provided by the Python library scikit-learn (Pedregosa et al., 2011). As most of the tested algorithms did not perform very well in basic tests, we focus on Keras as a common and versatile tool for neural network application.

Appendix B: Oleic acid ozonolysis datasets

In Sect. 3.4, KM-SUB and the NN surrogate model are applied to six experimental datasets of the ozonolysis of oleic acid aerosols – these are available in the literature (Hearn and Smith, 2004; Ziemann, 2005; Gallimore et al., 2017; Berkemeier et al., 2021). These datasets comprise flow tube, environmental chamber, and single-particle levitation techniques and are a subset of data investigated earlier by Berkemeier et al. (2021), omitting the studies that investigated particles with a sodium chloride core or in which the particle size was not measured. The experimental datasets are converted to normalized concentrations ($N_{Y,t}$ and $N_{Y,0}$) and are further simplified by fitting a mono-exponential decay ($A + B \cdot \exp(-\tau_e \cdot t)$) and evaluating the reaction time at which 10 %, 50 %, and 90 % of oleic acids are consumed. Table B1 shows the environmental parameters (particle radius r_p, ozone concentration $[X]_{g,0}$, and initial oleic acid concentration $[Y]_{b,0}$), the derived reaction times, and the mono-exponential fit parameters. The remaining seven KM-SUB input parameters listed in Table 1 are optimized. Figure B2 shows all datasets alongside a fit ensemble of 50 KM-SUB fits with a fit correlation MSLE of less than 0.016.

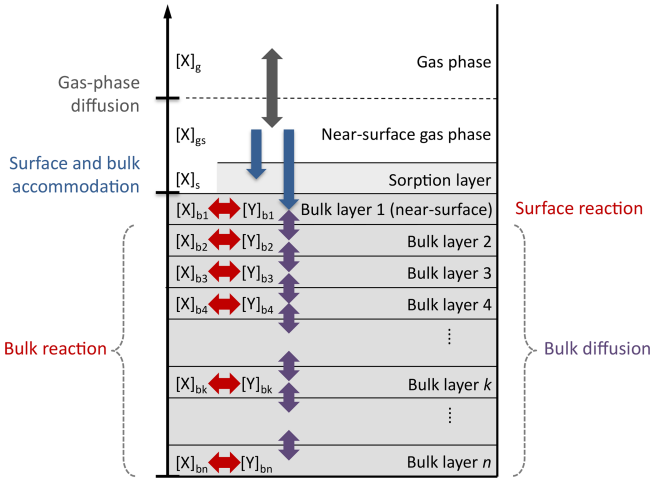

Figure B1. Compartments and processes of the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB).

Table B1. Model parameters for the global optimization of six oleic acid ozonolysis datasets.

Dataset	r_p (cm)	$[X]_{g,0}$ (cm^{-3})	$[Y]_{b,0}$ (cm^{-3})	$t_{10\,\%}$ (s)	$t_{50\,\%}$ (s)	$t_{90\,\%}$ (s)	A	B	τ_e
Berkemeier et al. (2021)	1×10^{-3}	1×10^{13}	1.89×10^{21}	24166	15892	52791	0	1	4.36×10^{-5}
Ziemann (2005)	2×10^{-5}	7×10^{13}	1.2×10^{21}	2.85	18.8	*	0	1	3.69×10^{-2}
Hearn and Smith (2004)	4×10^{-5}	2.5×10^{15}	1.89×10^{21}	0.196	1.29	4.28	0	1	0.538
Gallimore et al. (2017) – A	2.5×10^{-5}	2×10^{14}	1.89×10^{21}	1.91	12.6	41.7	0	1	5.52×10^{-2}
Gallimore et al. (2017) – B	2.5×10^{-5}	3.25×10^{14}	1.89×10^{21}	1.12	7.39	24.6	0	1	9.38×10^{-2}
Gallimore et al. (2017) – C	2.5×10^{-5}	5.51×10^{14}	1.89×10^{21}	11.2	3.37	0.512	0	1	2.06×10^{-1}

* Too far outside data range.

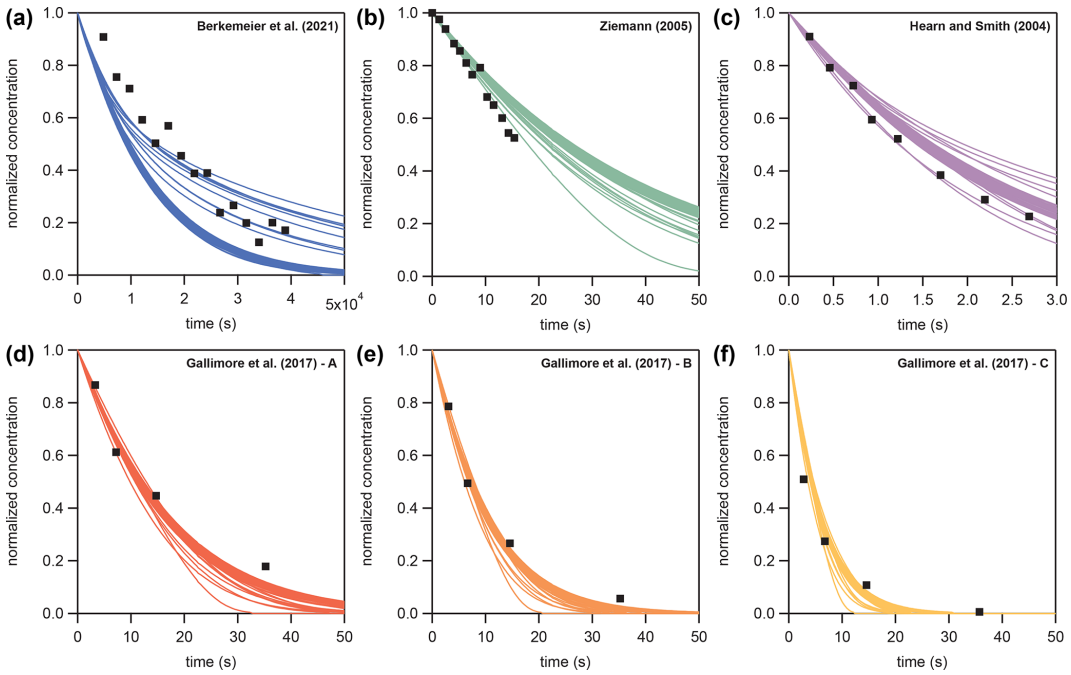

Figure B2. Fit ensembles of KM-SUB ($N = 50$, colored lines) with MSLE < 0.016 to six literature datasets (black square markers) of oleic acid aerosol ozonolysis displayed as normalized oleic acid concentrations ($N_{Y,t}/N_{Y,0}$).

Appendix C: Abbreviations

KM	Kinetic multi-layer model
KM-SUB	Kinetic multi-layer model of aerosol surface and bulk chemistry
MHA	Metropolis–Hastings algorithm
MLP	Multi-layer perceptron
MSE	Mean square error
MSLE	Mean squared (absolute) logarithmic error
NN	Neural network
PCE	Polynomial chaos expansion

Code and data availability. All training data, as well as the source code used for obtaining NN and PCE models, are archived on Zenodo (https://doi.org/10.5281/zenodo.7214880; Berkemeier et al., 2022).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-16-2037-2023-supplement.

Author contributions. TB and UKK conceived the study. All authors designed the research. TB (KM-SUB model), MK (NN model), and AF and MM (PCE model) wrote the code and performed the simulations. All authors discussed and interpreted the calculation results. TB and MK led the writing of the paper and the overall design of graphics and tables. AF and MM co-led the writing and graphics for the sections applying PCE models. All authors contributed to the writing and editing of the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The authors thank Coraline Mattei and Jake Wilson for the helpful discussions. We thank Paul Ziemann, Geoffrey Smith, and Peter Gallimore for providing the published data in tabulated form. The authors gratefully acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (https://hpc.uni-mainz.de/, last access: 11 April 2023) and on the supercomputer Cobra at the Max Planck Computing and Data Facility (https://www.mpcdf.mpg.de/, last access: 11 April 2023).

Financial support. This work was funded by the Max Planck Society (MPG) and supported by ETH Zurich through the ETH Research Grant (grant no. ETH-03 17-2). Matteo Krüger was supported by the Max Planck Graduate Center with the Johannes Gutenberg University (MPGC). Aryeh Feinberg acknowledges financial support from ETH Zurich (grant no. ETH-39 15-2).

The article processing charges for this open-access publication were covered by the Max Planck Society.

Review statement. This paper was edited by Po-Lun Ma and reviewed by two anonymous referees.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, [code], https://www.tensorflow.org/ (last access: 11 April 2023), 2015.

Allotey, J., Butler, K. T., and Thiyagalingam, J.: Entropy-based active learning of graph neural network surrogate models for materials properties, J. Chem. Phys., 155, 174116, https://doi.org/10.1063/5.0065694, 2021.

Almeida, L. B.: Multilayer Perceptrons, in: The Algebraic Mind: Integrating Connectionism and Cognitive Science, The MIT Press, https://doi.org/10.7551/mitpress/1187.003.0004, 2001.

Berkemeier, T., Huisman, A. J., Ammann, M., Shiraiwa, M., Koop, T., and Pöschl, U.: Kinetic regimes and limiting cases of gas uptake and heterogeneous reactions in atmospheric aerosols and clouds: a general classification scheme, Atmos. Chem. Phys., 13, 6663–6686, https://doi.org/10.5194/acp-13-6663-2013, 2013.

Berkemeier, T., Steimer, S. S., Krieger, U. K., Peter, T., Pöschl, U., Ammann, M., and Shiraiwa, M.: Ozone uptake on glassy, semi-solid and liquid organic matter and the role of reactive oxygen intermediates in atmospheric aerosol chemistry, Phys. Chem. Chem. Phys., 18, 12662–12674, https://doi.org/10.1039/C6CP00634E, 2016.

Berkemeier, T., Ammann, M., Krieger, U. K., Peter, T., Spichtinger, P., Pöschl, U., Shiraiwa, M., and Huisman, A. J.: Technical note: Monte Carlo genetic algorithm (MCGA) for model analysis of multiphase chemical kinetics to determine transport and reaction rate coefficients using multiple experimental data sets, Atmos. Chem. Phys., 17, 8021–8029, https://doi.org/10.5194/acp-17-8021-2017, 2017.

Berkemeier, T., Mishra, A., Mattei, C., Huisman, A. J., Krieger, U. K., and Pöschl, U.: Ozonolysis of Oleic Acid Aerosol Revisited: Multiphase Chemical Kinetics and Reaction Mechanisms, ACS Earth Space Chem., 5, 3313–3323, https://doi.org/10.1021/acsearthspacechem.1c00232, 2021.

Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., and Krieger, U.: Generation of surrogate models with artificial neural networks and polynomial chaos expansion (training data and source code), Zenodo [code, data set], https://doi.org/10.5281/zenodo.7214880, 2022.

Bishop, C. M.: Neural networks and their applications, Rev. Sci. Instrum., 65, 1803–1832, 1994.

Blatman, G. and Sudret, B.: Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys., 230, 2345–2367, https://doi.org/10.1016/j.jcp.2010.12.021, 2010.

Booker, A. J., Dennis, J. E., Frank, P. D., Serafini, D. B., Torczon, V., and Trosset, M. W.: A rigorous framework for optimization of expensive functions by surrogates, Struct. Multidiscip. O., 17, 1–13, 1999.

Cavalcanti, F. M., Kozonoe, C. E., Pacheco, K. A., and de Brito Alves, R. M.: Application of artificial neural networks to chemical and process engineering, IntechOpen, https://doi.org/10.5772/intechopen.96641, 2021.

Chib, S. and Greenberg, E.: Understanding the Metropolis-Hastings algorithm, Am. Stat., 49, 327–335, 1995.

Chollet, F. et al.: Keras, [code], https://github.com/fchollet/keras (last access: 11 April 2023), 2015.

Dou, J., Alpert, P. A., Corral Arroyo, P., Luo, B., Schneider, F., Xto, J., Huthwelker, T., Borca, C. N., Henzler, K. D., Raabe, J., Watts, B., Herrmann, H., Peter, T., Ammann, M., and Krieger, U. K.: Photochemical degradation of iron(III) citrate/citric acid aerosol quantified with the combination of three complementary experimental techniques and a kinetic process model, Atmos. Chem. Phys., 21, 315–338, https://doi.org/10.5194/acp-21-315-2021, 2021.

Esche, E., Weigert, J., Rihm, G. B., Göbel, J., and Repke, J.-U.: Architectures for neural networks as surrogates for dynamic systems in chemical engineering, Chem. Eng. Res. Des., 177, 184–199, 2022.

Feinberg, A., Maliki, M., Stenke, A., Sudret, B., Peter, T., and Winkel, L. H. E.: Mapping the drivers of uncertainty in atmospheric selenium deposition with global sensitivity analysis, Atmos. Chem. Phys., 20, 1363–1390, https://doi.org/10.5194/acp-20-1363-2020, 2020.

Feldman, J. A. and Ballard, D. H.: Connectionist Models and Their Applications: Introduction, Cogn. Sci., 6, 205–254, https://doi.org/10.1207/s15516709cog0901_1, 1982.

Galeazzo, T. and Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings, Environ. Sci. Atmos., 2, 362–374, https://doi.org/10.1039/D1EA00090J, 2022.

Gallimore, P., Griffiths, P., Pope, F., Reid, J., and Kalberer, M.: Comprehensive modeling study of ozonolysis of oleic acid aerosol based on real-time, online measurements of aerosol composition, J. Geophys. Res.-Atmos., 122, 4364–4377, 2017.

Gardner, M. W. and Dorling, S. R.: Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences, Atmos. Environ., 32, 2627–2636, https://doi.org/10.1016/S1352-2310(97)00447-0, 1998.

Ghanem, R. G. and Spanos, P. D.: Stochastic finite elements: a spectral approach, Courier Corporation, ISBN 10 0486428184, ISBN 13 9780486428185, 2003.

Gulli, A. and Pal, S.: Deep learning with Keras, Packt Publishing Ltd, ISBN 10 1787128423, ISBN 13 9781787128422, 2017.

Harder, P., Watson-Parris, D., Stier, P., Strassel, D., Gauger, N. R., and Keuper, J.: Physics-informed learning of aerosol microphysics, Environ. Data Sci., 1, e20, https://doi.org/10.1017/eds.2022.22, 2022.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Hearn, J. D. and Smith, G. D.: Kinetics and product studies for ozonolysis reactions of organic particles using aerosol CIMS, J. Phys. Chem. A, 108, 10019–10029, 2004.

Hecht-Nielsen, R.: Theory of the backpropagation neural network, in: Neural networks for perception, 65–93, Elsevier, https://doi.org/10.1016/B978-0-12-741252-8.50010-8, 1992.

Holeňa, M., Linke, D., Rodemerck, U., and Bajer, L.: Neural networks as surrogate models for measurements in optimization algorithms, in: International Conference on Analytical and Stochastic Modeling Techniques and Applications, Cardiff, UK, 14–16 June 2010, 351–366, Springer, https://doi.org/10.1007/978-3-642-13568-2_25, 2010.

Keller, C. A. and Evans, M. J.: Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10, Geosci. Model Dev., 12, 1209–1225, https://doi.org/10.5194/gmd-12-1209-2019, 2019.

Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D., and Tessum, C. W.: Toward Stable, General Machine-Learned Models of the Atmospheric Chemical System, J. Geophys. Res.-Atmos., 125, e2020JD032759, https://doi.org/10.1029/2020JD032759, 2020.

Kelp, M. M., Jacob, D. J., Lin, H., and Sulprizio, M. P.: An online-learned neural network chemical solver for stable long-term global simulations of atmospheric chemistry, J. Adv. Model. Earth Sy., 14, e2021MS002926, https://doi.org/10.1029/2021MS002926, 2022.

Kolb, C. E., Cox, R. A., Abbatt, J. P. D., Ammann, M., Davis, E. J., Donaldson, D. J., Garrett, B. C., George, C., Griffiths, P. T., Hanson, D. R., Kulmala, M., McFiggans, G., Pöschl, U., Riipinen, I., Rossi, M. J., Rudich, Y., Wagner, P. E., Winkler, P. M., Worsnop, D. R., and O' Dowd, C. D.: An overview of current issues in the uptake of atmospheric trace gases by aerosols and clouds, Atmos. Chem. Phys., 10, 10561–10605, https://doi.org/10.5194/acp-10-10561-2010, 2010.

Kröse, B. and van der Smagt, P.: An Introduction to Neural Networks, The University of Amsterdam, https://www.infor.uva.es/~teodoro/neuro-intro.pdf (last access: 11 April 2023), 1996.

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B. A. M., Lammel, G., Schmidt, B., Pöschl, U., and Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, Nat. Sci., 2, e20220016, https://doi.org/10.1002/ntls.20220016, 2022.

Kuwata, M. and Martin, S. T.: Phase of atmospheric secondary organic material affects its reactivity, P. Natl. Acad. Sci. USA, 109, 17354–17359, 2012.

Le Gratiet, L., Marelli, S., and Sudret, B.: Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes, in: Handbook of Uncertainty Quantification, 1289–1325, Springer, https://doi.org/10.1007/978-3-319-12385-1_38, 2017.

Lu, J., Zhang, H., Yu, J., Shan, D., Qi, J., Chen, J., Song, H., and Yang, M.: Predicting rate constants of hydroxyl radical reactions with alkanes using machine learning, J. Chem. Inf. Model., 61, 4259–4265, 2021.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., and Rinke, P.: Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning, Atmos. Chem. Phys., 21, 13227–13246, https://doi.org/10.5194/acp-21-13227-2021, 2021.

Marelli, S. and Sudret, B.: UQLab: A framework for uncertainty quantification in Matlab, in: Vulnerability, uncertainty, and risk: quantification, mitigation, and management, 2554–2563, American Society of Civil Engineers, [code], https://doi.org/10.1061/9780784413609.257, 2014.

McKinney, W. et al.: Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Confer-

ence, Austin, TX, 28 June–3 July 2010, [code], 445, 51–56, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

Milsom, A., Squires, A. M., Ward, A. D., and Pfrang, C.: The impact of molecular self-organisation on the atmospheric fate of a cooking aerosol proxy, Atmos. Chem. Phys., 22, 4895–4907, https://doi.org/10.5194/acp-22-4895-2022, 2022.

O'Gorman, P. A. and Dwyer, J. G.: Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events, J. Adv. Model. Earth Syst., 10, 2548–2563, https://doi.org/10.1029/2018MS001351, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Popescu, M.-C., Balas, V. E., Perescu-Popescu, L., and Mastorakis, N.: Multilayer perceptron and neural networks, WSEAS Trans. Circuits Syst., 8, 579–588, 2009.

Pöschl, U., Rudich, Y., and Ammann, M.: Kinetic model framework for aerosol and cloud surface chemistry and gas-particle interactions – Part 1: General equations, parameters, and terminology, Atmos. Chem. Phys., 7, 5989–6023, https://doi.org/10.5194/acp-7-5989-2007, 2007.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, P. Natl. Acad. Sci. USA, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Robert, C. P. and Casella, G.: The Metropolis-Hastings Algorithm, in: Monte Carlo statistical methods, 231–283, Springer, https://doi.org/10.1007/978-1-4757-3071-5_6, 1999.

Roldin, P., Eriksson, A. C., Nordin, E. Z., Hermansson, E., Mogensen, D., Rusanen, A., Boy, M., Swietlicki, E., Svenningsson, B., Zelenyuk, A., and Pagels, J.: Modelling non-equilibrium secondary organic aerosol formation and evaporation with the aerosol dynamics, gas- and particle-phase chemistry kinetic multilayer model ADCHAM, Atmos. Chem. Phys., 14, 7953–7993, https://doi.org/10.5194/acp-14-7953-2014, 2014.

Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y.: Backpropagation: The basic theory, in: Backpropagation: Theory, architectures and applications, 1–34, Lawrence Erlbaum Hillsdale, NJ, USA, ISBN 0-8058-1259-8, 1995.

Sadeeq, M. A. and Abdulazeez, A. M.: Neural networks architectures design, and applications: A review, in: 2020 International Conference on Advanced Science and Engineering (ICOASE), Duhok, Iraq, 23–24 December 2020, IEEE, 199–204, https://doi.org/10.1109/ICOASE51841.2020.9436582, 2020.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global sensitivity analysis: the primer, John Wiley & Sons, ISBN 978-0-470-05997-5, 2008.

Semeniuk, K. and Dastoor, A.: Current state of atmospheric aerosol thermodynamics and mass transfer modeling: A review, Atmosphere, 11, 156, https://doi.org/10.3390/atmos11020156, 2020.

Shiraiwa, M., Pfrang, C., and Pöschl, U.: Kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB): the influence of interfacial transport and bulk diffusion on the oxidation of oleic acid by ozone, Atmos. Chem. Phys., 10, 3673–3691, https://doi.org/10.5194/acp-10-3673-2010, 2010.

Shiraiwa, M., Ammann, M., Koop, T., and Pöschl, U.: Gas uptake and chemical aging of semisolid organic aerosol particles, P. Natl. Acad. Sci. USA, 108, 11003–11008, 2011.

Shiraiwa, M., Pfrang, C., Koop, T., and Pöschl, U.: Kinetic multilayer model of gas-particle interactions in aerosols and clouds (KM-GAP): linking condensation, evaporation and chemical reactions of organics, oxidants and water, Atmos. Chem. Phys., 12, 2777–2794, https://doi.org/10.5194/acp-12-2777-2012, 2012.

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmos. Chem. Phys., 14, 8323–8341, https://doi.org/10.5194/acp-14-8323-2014, 2014.

Sobol', I. M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simulat., 55, 271–280, https://doi.org/10.1016/S0378-4754(00)00270-6, 2001.

Stone, M.: Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. B, 36, 111–133, 1974.

Sturm, P. O. and Wexler, A. S.: Conservation laws in a neural network architecture: enforcing the atom balance of a Julia-based photochemical model (v0.2.0), Geosci. Model Dev., 15, 3417–3431, https://doi.org/10.5194/gmd-15-3417-2022, 2022.

Sudret, B.: Global sensitivity analysis using polynomial chaos expansions, Reliab. Eng. Syst. Safe., 93, 964–979, https://doi.org/10.1016/j.ress.2007.04.002, 2008.

Thackray, C. P., Friedman, C. L., Zhang, Y., and Selin, N. E.: Quantitative Assessment of Parametric Uncertainty in Northern Hemisphere PAH Concentrations, Environ. Sci. Technol., 49, 9185–9193, https://doi.org/10.1021/acs.est.5b01823, 2015.

Tikkanen, O.-P., Hämäläinen, V., Rovelli, G., Lipponen, A., Shiraiwa, M., Reid, J. P., Lehtinen, K. E. J., and Yli-Juuti, T.: Optimization of process models for determining volatility distribution and viscosity of organic aerosols from isothermal particle evaporation data, Atmos. Chem. Phys., 19, 9333–9350, https://doi.org/10.5194/acp-19-9333-2019, 2019.

Tripathy, R. K. and Bilionis, I.: Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification, J. Comput. Phys., 375, 565–588, 2018.

Vu, K. K., d'Ambrosio, C., Hamadi, Y., and Liberti, L.: Surrogate-based methods for black-box optimization, Int. T. Oper. Res., 24, 393–424, 2017.

Wei, J., Fang, T., Lakey, P. S., and Shiraiwa, M.: Iron-Facilitated Organic Radical Formation from Secondary Organic Aerosols in Surrogate Lung Fluid, Environ. Sci. Technol., 56, 7234–7243, https://doi.org/10.1021/acs.est.1c04334, 2021.

Wong, T.-T. and Yeh, P.-Y.: Reliable accuracy estimates from k-fold cross validation, IEEE T. Knowl. Data En., 32, 1586–1594, https://doi.org/10.1109/TKDE.2019.2912815, 2020.

Xia, D., Chen, J., Fu, Z., Xu, T., Wang, Z., Liu, W., Xie, H.-B., and Peijnenburg, W. J.: Potential application of machine-learning-based quantum chemical methods in environmental chemistry, Environ. Sci. Technol., 56, 2115–2123, 2022.

Xiu, D. and Karniadakis, G. E.: The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput., 24, 619–644, 2002.

Xu, H., Zhang, T., Luo, Y., Huang, X., and Xue, W.: Parameter calibration in global soil carbon models using surrogate-based optimization, Geosci. Model Dev., 11, 3027–3044, https://doi.org/10.5194/gmd-11-3027-2018, 2018.

Ziemann, P. J.: Aerosol products, mechanisms, and kinetics of heterogeneous reactions of ozone with oleic acid in pure and mixed particles, Faraday Discuss., 130, 469–490, 2005.

2.5. A numerical compass for experiment design in chemical kinetics and molecular property estimation

This chapter presents a research article published in the Journal of Cheminformatics. I am the first author and the main contributor to this paper. Following up on the idea and basic feasibility tests carried out by Thomas Berkemeier and Ashmi Mishra, I wrote and refined the NC method in Python, designed and carried out all experiments and included the surrogate models. I prepared all figures, wrote the manuscript together with Thomas Berkemeier and revised the study during peer-review. Finally, I wrote a user-friendly, efficient and versatile Julia package that is openly available for download. More detailed information on the author contributions are provided at the end of the manuscript.

Krüger, M., Mishra, A., Spichtinger, P., Pöschl, U., Berkemeier, T.: A numerical compass for experiment design in chemical kinetics and molecular property estimation, *J. Cheminform.*, doi: 10.1186/s13321-024-00825-0, (2024).

A novel framework, the Numerical Compass (NC) method was developed to integrate computational models, global optimization, ensemble methods, and ML to identify experimental conditions with the greatest potential to constrain model parameters. The approach is based on the quantification of model output variance in an ensemble of solutions that agree with experimental data. The utility of the NC method was demonstrated for the parameters of a multi-layer model describing the heterogeneous ozonolysis of oleic acid aerosols. NN surrogate models of the multiphase chemical reaction system were used to accelerate the application of the NC for a comprehensive mapping and analysis of experimental conditions. The supplement to this work can be found in appendix B4.

Krüger *et al. Journal of Cheminformatics* (2024) 16:34
https://doi.org/10.1186/s13321-024-00825-0

Journal of Cheminformatics

RESEARCH Open Access

A numerical compass for experiment design in chemical kinetics and molecular property estimation

Matteo Krüger[1], Ashmi Mishra[1], Peter Spichtinger[2], Ulrich Pöschl[1] and Thomas Berkemeier[1*]

Abstract

Kinetic process models are widely applied in science and engineering, including atmospheric, physiological and technical chemistry, reactor design, or process optimization. These models rely on numerous kinetic parameters such as reaction rate, diffusion or partitioning coefficients. Determining these properties by experiments can be challenging, especially for multiphase systems, and researchers often face the task of intuitively selecting experimental conditions to obtain insightful results. We developed a numerical compass (NC) method that integrates computational models, global optimization, ensemble methods, and machine learning to identify experimental conditions with the greatest potential to constrain model parameters. The approach is based on the quantification of model output variance in an ensemble of solutions that agree with experimental data. The utility of the NC method is demonstrated for the parameters of a multi-layer model describing the heterogeneous ozonolysis of oleic acid aerosols. We show how neural network surrogate models of the multiphase chemical reaction system can be used to accelerate the application of the NC for a comprehensive mapping and analysis of experimental conditions. The NC can also be applied for uncertainty quantification of quantitative structure–activity relationship (QSAR) models. We show that the uncertainty calculated for molecules that are used to extend training data correlates with the reduction of QSAR model error. The code is openly available as the Julia package *KineticCompass*.

Keywords Chemical kinetics, QSAR, Design of experiments (DOE), Global optimization, Inverse problem, Ensemble methods, Multiphase chemistry, Machine learning

Introduction

In multiphase chemical kinetics, the rate of change in complex systems can be described by resolving mass transport and chemical reactions at the molecular process level [1, 2]. While the underlying physical and chemical principles are well understood, the individual processes are inherently coupled and the chemical and physical parameters, such as reaction, diffusion, or partitioning coefficients, are often unknown or poorly constrained [3, 4]. The integration of these processes occurring in parallel or in sequence often requires computational kinetic models (KM). KM return the concentration time profiles of reactants or products under specified environmental or experimental conditions [5–10]. However, the input parameters for KM may not be known *a priori*, and their determination can be challenging [11–14]. The deduction or constraint of model input parameters using model output is known as solving the inverse problem. In practice, researchers often utilize statistical approaches to solve the inverse problem

*Correspondence:
Thomas Berkemeier
t.berkemeier@mpic.de
[1] Multiphase Chemistry Department, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, Mainz 55128, Rhineland Palatinate, Germany
[2] Institute for Atmospheric Physics, Johannes Gutenberg University, Johann-Joachim-Becher-Weg 21, Mainz 55128, Rhineland Palatinate, Germany

with global optimization techniques [15–18]. Such techniques determine sets of parameter values, so-called fits, that lead to model outputs in agreement with previously acquired experimental data. In ill-posed problems, Berkemeier et al. 2021 [19] proposed the consideration of ensembles of sufficiently well-fitting parameter sets to extract information from the corresponding range of kinetic model solutions in underdetermined optimization problems. This approach is related to approximate Bayesian computation, a method for statistical inference that can be applied if the likelihood function is not known and the posterior distribution cannot be obtained directly [20]. This is often the case for computational or simulation-based models that are evaluated through calculation of a mechanism, like (bio-)chemical kinetic models [21, 22]. In approximate Bayesian computation, a probability density function is replaced by an artificial data set obtained through sampling of an approximate posterior distribution using a distance metric [23]. In this work, the approximate posterior distribution corresponds to the fit ensemble, i.e., kinetic parameter sets that lead to valid solutions matching experimental data within a specified error margin.

Quantitative structure–activity relationship (QSAR) models utilize the concept of molecular similarity to derive properties (e.g., chemical or biological) of new molecules from existing data, often through machine learning [24]. The models are generally trained on data derived from experimental measurements [25] or density functional theory (DFT) calculations [26–28]. Similarly to the acquisition of fit ensembles in global optimization, ensemble learning techniques allow the acquisition and utilization of multitudes of QSAR model predictions [29–31]. Such ensemble predictions have recently been utilized for uncertainty quantification, based on variance in predictions of Siamese neural networks [32].

Surrogate models (SM) are machine learning models that are trained on inputs and outputs of a template model. A SM can be used to substitute the template model in applications that benefit from low computational cost in exchange for slightly increased model uncertainty. Satisfactory model accuracy can be ensured by a sufficient size of the training data set, and therefore depends on the initial investment of computational resources [33]. SM have helped solving the issue of computational cost in many fields of research, such as in geoscientific and atmospheric modelling [34–40], chemical process engineering [41], water resources modelling [42, 43], or optimization in supply chain management [44]. SM can also aid inverse modelling approaches. Berkemeier et al. 2023 [33] showed that SM-supported fit ensemble acquisition greatly outperforms regular sampling with the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) [5] in terms of acquired fits for a given computational effort. However, it remains unclear how SM uncertainty affects the reliability of inverse modelling techniques.

A kinetic model's uncertainty can be based on model form uncertainty, i.e., concerning the underlying physics or chemistry, or model parametric uncertainty, i.e., concerning the knowledge of its input parameters [45]. Parametric uncertainty is often caused by the coupled nature of parameters or by underdetermination of the modelled system. Among model input parameters, we differentiate between kinetic parameters that define the physical and chemical properties of the modelled system (e.g., reaction rate coefficients), and parameters that define the environmental or experimental conditions (e.g., initial concentrations or temperature). When a model is evaluated for experimental conditions that differ from those for which its kinetic parameters were derived, model uncertainty may strongly increase [2]. This situation may arise in particular when the data underlying the model is limited, or when conditions in the laboratory experiment (e.g., a test reactor) deviate from the real-world application of interest (e.g., the atmosphere, an industrial plant, or an engine). Furthermore, when extrapolating a model to conditions outside its calibration range, not all fits in a fit ensemble may behave in the same way. This ensemble variance associated with a fit ensemble can be used to assess the model's parametric uncertainty over a range of experimental conditions [19]. The ensemble variance at a specific set of experimental conditions may also be an indicator for parameter sensitivity, and of the potential to constrain the model if experimental data was available for these conditions. Thus, while data from any additional experiment may decrease the parametric uncertainty of a model, this process can be optimized by selecting experimental conditions associated with high ensemble variance. These conditions are most likely to constrain the underlying model and its physical and chemical parameters.

For experimenters, it is difficult to guess such optimal conditions *a priori*. As quantitative approaches to this problem, a number of methods and frameworks for targeted design of experiments (DOE) for uncertainty minimization have emerged over the past years, mostly in the fields of fuel combustion and computational fluid dynamics [46]. For this purpose, Bayesian experimental design methods have been proposed to maximize a utility function, e.g., through minimization of information entropy, a measure for the degree of disorder, diversity and dispersion [47]. DOE techniques have since then been continuously extended and improved, e.g., through the utilization of polynomial surrogate models [48], sensitivity entropy as a measure of the degree of dispersion

of uncertainty sources of a model output [49], truncated Gaussian probability density functions [50] or surrogate model similarity methods [51]. For example, Lehn et al. successfully applied an iterative model-based experimental design framework based on the criterion of D-optimality [52] as well as polynomial chaos expansion [53] to identify optimal conditions for experimental measurements related to the auto-ignition of dimethyl ether [54]. Through integration of functions for dimension reduction, global sensitivity analysis, forward uncertainty quantification, model-analysis-based experimental design and model optimization, Zhou et al. developed a versatile computational framework (OptEx) to automatically find informative while independent experiments, and refine computational models [55]. Similar methods for so-called calibration experiment design optimization techniques have been developed and are applied in the fields of engineering and materials science [56, 57]. To our knowledge, however, such techniques had not yet been developed and applied to guide laboratory experiments in the fields of atmospheric and environmental multiphase chemistry.

Existing DOE methods are often based on optimality criteria to minimize the trace (A-Optimality), determinant (D-Optimality) or eigenvalue (E-Optimality) of the Fisher information matrix, and require knowledge of a likelihood function, given experimental data and uncertainty [52, 58]. To calculate the Fisher information matrix, derivatives of the likelihood function with respect to the model parameters must be obtained [59]. If automatic differentiation is not applicable to the model [60], the calculation of gradients through, e.g., finite differences [61], requires multiple model evaluations per maximum likelihood estimate and tested experimental condition [62]. In this work, we propose a new approach to the selection of optimal experiments. The numerical compass (NC) method treats experimental uncertainty implicitly through choice of acceptance conditions (e.g., thresholds) to derive a fit ensemble as representation of the underlying solution space. The approach represents a least-squares method for parameter estimation, in contrast to the more common maximum likelihood estimation methods [63]. The optimality criteria for the selection of experiments in our proposed method are formulated as statistical criteria, which we will refer to as *constraint potentials*. These are computationally inexpensive operations that only require one model evaluation per fit and tested experimental condition. The criteria can be specifically tailored to consider additional information associated with the fit ensemble, or specific properties of the model. In the proposed framework, we introduce two constraint potential metrics: one approximates the heterogeneity of models (i.e., posterior distribution samples) at different experimental conditions, and one that further explores the nature of constraint potentials with regards to individual kinetic parameters. The NC is used alongside the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB), and a neural network SM for it, to demonstrate its functionality in experiment design and inverse modelling. In addition to experiment design, we apply the NC to uncertainty quantification of machine learning quantitative structure–activity relationship (QSAR) models. The NC is used to explore molecular structures for which QSAR models exhibit a particularly high uncertainty and test whether this information can be used to suggest new training data that will increase model accuracy.

Method

We present the numerical compass (NC), a method for experiment prioritization and reduction of a model's parametric uncertainty. The method requires a process model, data from previous laboratory experiments, and a set of variable experimental parameters that describe future experiments of interest. The individual steps of the proposed workflow are displayed in Fig. 1.

Inverse modelling solutions and uncertainty

To estimate parametric uncertainty, inverse modelling can be extended to an ensemble of kinetic parameter sets that return sufficient agreement with experimental data [15, 19]. All possible sets of chemical and physical parameter values that lead to a sufficiently low residual between model output and experimental data, so-called fits, form the solution space of a kinetic model. In practice, we use a finite collection of fits, referred to as *fit ensemble*, as representation of the model solution space. Additional experimental data can help to narrow down the fit ensemble and thus decrease model parametric uncertainty.

Operating principle

The NC is a framework to optimize the deduction or constraint of kinetic parameters with experiments. In general, the information gained from new experimental data can be used to reject fits from a fit ensemble. The NC finds experimental conditions with the highest constraint potential, optimizing the reduction of model solution space and hence model parametric uncertainty. For this purpose, the method computes ensemble solutions under experimental conditions that have not been considered previously, and determines the ensemble variance under these conditions. We present two metrics evaluating the ensemble variance, the *ensemble spread* of model solutions (section Ensemble spread) and the *parameter (boundary) constraint potential* (section Parameter

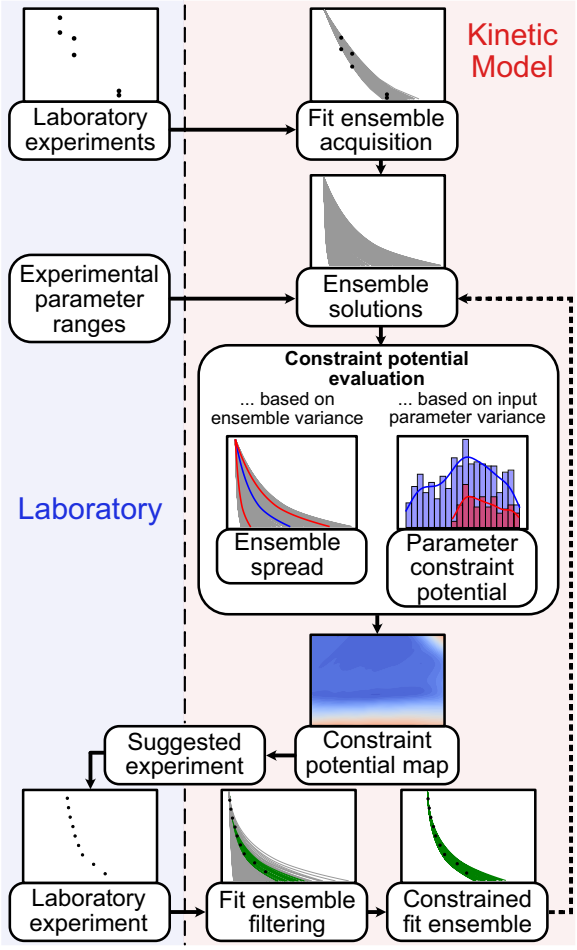

Fig. 1 Workflow of the numerical compass (NC) method presented in this study. The method relies on exchange between laboratory experiments (left) and model calculations (right) to eliminate variance in model output. Data from laboratory experiments are used for the acquisition of a fit ensemble, which are kinetic parameter sets that lead to model outputs in agreement with the experimental measurements. Evaluating the model for the entire fit ensemble and over a defined range of experimental parameters yields sets of ensemble solutions that serve as the basis for all calculations with the NC. The NC offers two metrics for constraint potential evaluation: ensemble spread, and parameter (boundary) constraint potential (section Parameter boundary constraint potential). The metrics are used to build constraint potential maps, which highlight areas with large model output variance in the experimental parameter range. These experimental parameters are suggested as next experiment as they are likely to lead to rejection of a large number of fits during fit ensemble filtering. The NC can be used iteratively (dotted arrow), using the ensemble solutions of the constrained fit ensembles

boundary constraint potential). By sampling the space of feasible experiments, *constraint potential maps* (section Constraint potential maps) of these metrics are obtained. Maxima on these maps represent prospective experiments that are most likely to achieve large constraints of the model. After fit ensemble filtering based on the new experimental data, the NC method can be repeated to suggest the next experiment. In this study, we simulate the suggested laboratory experiments with the model KM-SUB to showcase the alternating application of the NC with laboratory experiments. For more detailed and mathematical definitions of process models, their solution space, as well as fit ensembles and ensemble solutions, see Additional file 1: Note S1.

Ensemble spread

The ensemble spread is a measure for the variance between a multitude of model predictions. Resembling similar concepts in weather and climate forecasting [64], we calculate the ensemble spread (ES) as:

$$\mathrm{ES} = \frac{\int(\overline{Z}(x) + \sigma_Z(x))dx - \int(\overline{Z}(x) - \sigma_Z(x))dx}{\int \overline{Z}(x)dx} \quad (1)$$

where $(x_m)_{m=1,\ldots,n_z}$ is the sequence of independent variables associated with the output sequence $(z_m)_{m=1,\ldots,n_z}$, and $\int \overline{Z}$, $\int \overline{Z} + \sigma$ and $\int \overline{Z} - \sigma$ are integrals of the interpolated sequences $(\overline{Z_m})_{m=1,\ldots,n_z}$, $(\overline{Z_m} + \sigma_m)_{m=1,\ldots,n_z}$ and $(\overline{Z_m} - \sigma_m)_{m=1,\ldots,n_z}$ for n_z model outputs with an ensemble mean $\overline{Z_m}$ and ensemble standard deviation σ_m (Additional file 1: Note S2).

In short, the ensemble spread describes the area enclosed by the curves of the ensemble mean ± its standard deviation, normalized by the area under the ensemble mean curve. Visualizations of the ensemble spread as constraint potential metric are provided in Fig. 2D, E. A large ensemble spread is generally associated with a larger fraction of rejected fits during fit ensemble filtering.

Parameter boundary constraint potential

The parameter (boundary) constraint potential allows an extension of the method to constraint potentials of individual kinetic parameters. The metric quantifies the potential narrowing of an individual parameter's boundaries in the constrained fit ensemble.

In brief, the parameter constraint potential is calculated by iterating over predictions in an ensemble solution. In each iteration, one prediction is considered as the hypothetical result of an experiment. Based on this prediction, we calculate a hypothetical constrained fit ensemble and derive the distribution of the kinetic parameter in the remaining fits. The kinetic parameter's boundaries in this distribution are normalized by its boundaries in the original fit ensemble to compute a numerical value for the parameter's constraint potential.

More specifically, we determine the subset C of the fit ensemble FE. C contains all fits that lead to model

solutions within acceptance threshold θ in comparison to the model solution of fit FE_l that is selected as hypothetical measurement in the iteration l over all predictions in the ensemble solution:

$$\mathrm{C}_l = \{\mathrm{FE}_r : \Delta(\mathrm{ENS}_l, \mathrm{ENS}_r) < \theta\} \tag{2}$$

where ENS_l and ENS_r are the model solutions using fits FE_l and FE_r in the evaluated ensemble solution (ENS). Hence, we obtain one subset C_l in each iteration. If every solution in the ensemble is evaluated as hypothetical experimental result in turn, n_{FE} subsets are generated for every ensemble solution, where n_{FE} is the number of elements in the fit ensemble . The parameter constraint potential (PCP) for a specific parameter λ_p and ensemble solution is then defined as:

$$\mathrm{PCP}_p = \sum_{l=1}^{n_{\mathrm{FE}}} (\mathrm{Q}5_{\lambda_p,l} - \min(\lambda_p)) + (\max(\lambda_p) - \mathrm{Q}95_{\lambda_p,l}) \tag{3}$$

where $\mathrm{Q}5_{\lambda_p,l}$ and $\mathrm{Q}95_{\lambda_p,l}$ are the 5- and 95-percentiles of the distribution of λ_p in subset C_l, respectively. $\min(\lambda_p)$ and $\max(\lambda_p)$ are the global minimum and maximum of the selected kinetic parameter in the entire fit ensemble.

Note that the computational effort associated with this method is large due to the pairwise comparison of all predictions in an ensemble solution. Therefore, we suggest an approximation based on a reduced sample density. A detailed definition of the parameter constraint potential with reduced sample density is presented in Additional file 1: Note S3 and visualized in Additional file 1: Fig. S1.

Note that we can apply the same principle of forming subsets of the fit ensemble based on their behavior under test conditions, to constrain model uncertainty at a specific target condition (Additional file 1: Fig. S2). This can be of high practical relevance for situations where laboratory experiments must be performed outside the typical conditions of the target application, a common problem in the fields of atmospheric chemistry and chemical technology.

Constraint potential maps

The application of a metric for model constraint potential on a range of ensemble solutions (one for each tested experimental condition) can be visualized in a constraint potential map. This map is a n-dimensional hypersurface, where n is the number of varied experimental parameters, and whose maxima represent experimental conditions favorable for constraint of the underlying model. An example for a constraint potential map is presented for two varied experimental parameters and the ensemble spread metric in Fig. 2. For further information on the chemical system (oleic acid ozonolysis) and the variable experimental parameters (particle radius, ozone concentration), as well as a description of the restrictions regarding experimental accessibility applied in this work, see section Kinetic multi-layer model and neural network surrogate model, Additional file 1: Note S4, and Additional file 1: Fig. S3. Note that while we evaluate a full grid of combinations of experimental parameters for the purpose of testing and visualization, the constraint potential metrics can similarly be used as an objective function of an optimization algorithm to reduce the required computational effort.

Kinetic multi-layer model and neural network surrogate model

In this study, we use the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) [5] along with experimental data of the heterogeneous ozonolysis of oleic acid from the literature. However, the NC method can be used with any process model and underlying chemical or physical system. Detailed information about KM-SUB can be found in previous publications [5, 12]. In brief, KM-SUB is a chemical flux model that explicitly describes gas diffusion, accommodation of gas molecules to surfaces, surface-bulk exchange, bulk diffusion, as well as chemical reaction at the surface and in the bulk of a condensed phase. The resulting set of ordinary differential equations is solved numerically. KM-SUB input parameters include initial concentrations, chemical reaction rate coefficients, and mass transport coefficients, and are presented in Table 1. KM-SUB outputs are the concentration profiles over space and time for all chemical species.

For the training of neural network surrogate models, KM-SUB output is simplified to nine points of reaction progress, i.e., the time required to reach 90 %, 80 %, 70 %, 60 %, 50 %, 40 %, 30 %, 20 % and 10 % of the total number of oleic acid (OL) in a single aerosol particle, $N_{\mathrm{OL},0}$. For comparability, we represent the output of the full KM-SUB model in this study in the same way. We train a fully-connected, feed-forward neural network on 1×10^6 KM-SUB outputs as training data. For further information on training of the surrogate model see Berkemeier et al. 2023 [33] and Additional file 1: Note S5.

The NC method requires evaluation of the underlying process model during fit ensemble acquisition and during calculation of ensemble solutions (Fig. 1). In this study, we test and compare three different approaches: using KM-SUB for both steps (KM-only), using an SM of KM-SUB for both steps (SM-only), and a KM/SM-hybrid approach, in which KM-SUB is used for fit ensemble acquisition and the SM to obtain ensemble solutions. Fit ensemble acquisition is achieved by random sampling of kinetic input parameters with the KM or SM within the parameter boundaries

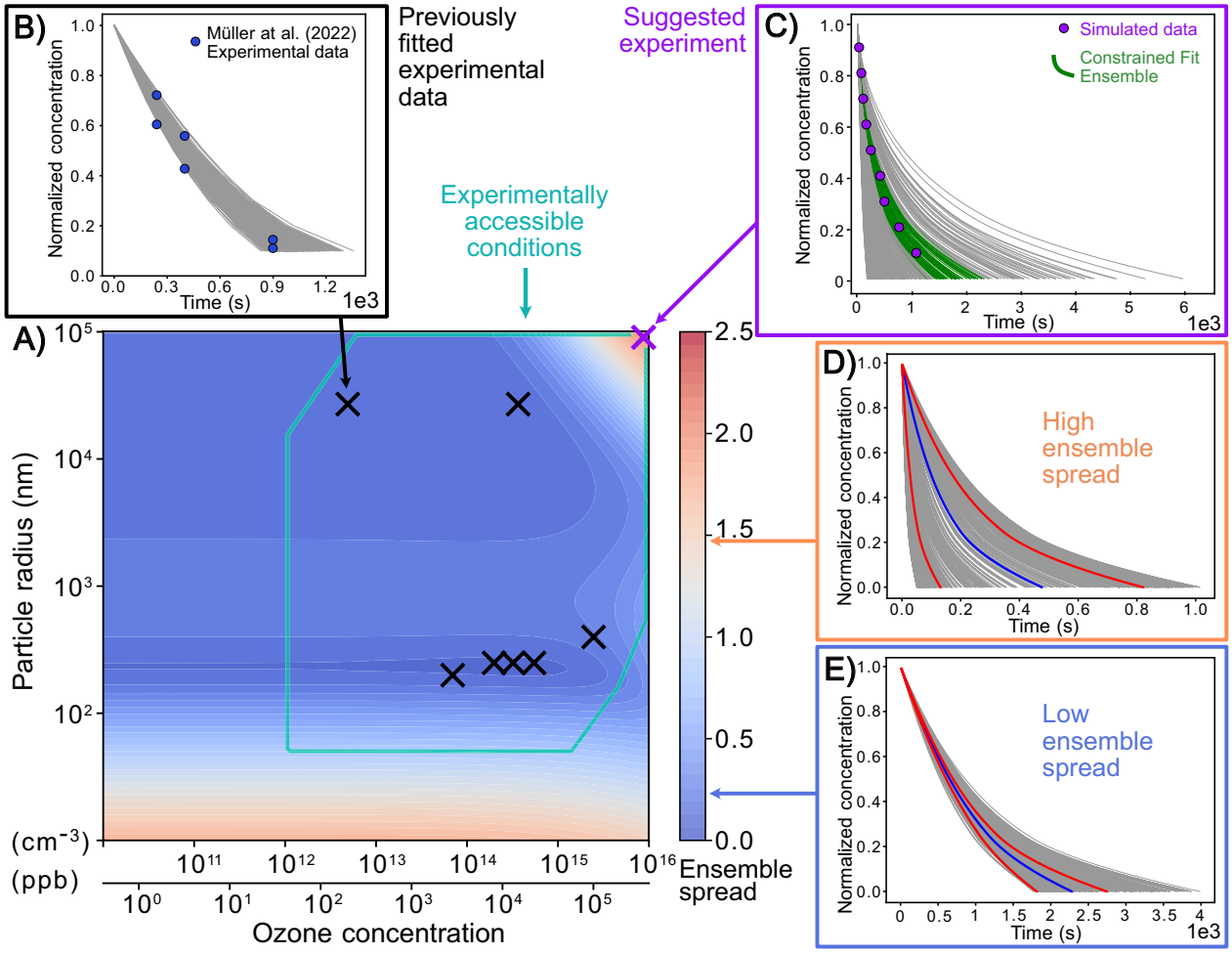

Fig. 2 Constraint potential map obtained with the numerical compass (NC) method. The contour map in **A** shows an exemplary constraint potential map using the ensemble spread metric. Model calculations are obtained with KM-SUB on a 100×100 grid of two experimental parameters, ozone concentration and particle radius, and for a fit ensemble of 500 fits. The teal box frames the area of experimentally accessible conditions with regards to particle radius, ozone concentration and predicted experiment duration (Additional file 1: Note S4). Black crosses in **A** mark the experimental conditions of available experimental data that were used to obtain the fit ensemble (cf. Fig. 3) and **B** shows the ensemble solution (gray lines) in comparison to one of these experimental data sets (blue markers). The purple cross in **A** represents the ensemble spread maximum within experimental accessibility and thus the recommended experiment. **C** Illustrates the ensemble solution at this ensemble spread maximum. New experimental data from the recommended experiment (purple markers) are used to obtain the constrained fit ensemble (green lines) through rejection of fits. **D**, **E** Showcase ensemble solutions with a high ensemble spread of 1.446 and a low ensemble spread of 0.234, respectively. Here, colored lines visualize the mean of the ensemble solution (blue line) and the mean ± 1 standard deviation (red lines)

in Table 1, using a mean square logarithmic error (MSLE) and an acceptance threshold $\theta = 0.0105$ to determine sufficient agreement with experimental data. For the specifications of fit ensemble acquisition and error calculation in this study, see Additional file 1: Note S6.

Quantitative activity structure relationship models and ensemble learning

In addition to experiment design, the NC can be utilized for uncertainty quantification of QSAR models. We use a re-trained version of the CNN_Tabor_nosulf model from Krüger et al. [28], a convolutional neural network model predicting reduction potentials based on SMILES molecular representations of 69,599 quinones from the Tabor et al. [65] data set, excluding quinone structures that contain sulfate functional groups. The models are trained on identical hyper-parameters as in the original study, but using 10-fold instead of 5-fold cross-validation. In this application, the ensemble solution utilized by the NC refers to multiple cross-validation models that are trained on different subsets of the training data. We

Table 1 KM-SUB kinetic and experimental input parameters

Parameter	Lower boundary	Upper boundary	Description
k_{SLR}	1.0×10^{-15}	1.0×10^{-8}	Rate coefficient of OL+O_3 surface reaction ($cm^3\ s^{-1}$)
k_{BR}	1.0×10^{-20}	1.0×10^{-11}	Rate coefficient of OL+O_3 bulk reaction ($cm^3\ s^{-1}$)
$D_{b,O3}$	1.0×10^{-11}	1.0×10^{-5}	Bulk diffusion coefficient of ozone ($cm^2\ s^{-1}$)
$D_{b,OL}$	1.0×10^{-12}	1.0×10^{-6}	Bulk diffusion coefficient of oleic acid ($cm^2\ s^{-1}$)
$H_{cp,O3}$	5.0×10^{-6}	5.0×10^{-3}	Henry's law solubility coefficient of ozone ($mol\ cm^{-3}\ atm^{-1}$)
$\tau_{d,O3}$	1.0×10^{-9}	1.0×10^{-2}	Desorption lifetime of O_3 (s)
$\alpha_{s,0,O3}$	1.0×10^{-4}	1	Surface accommodation coefficient of ozone on an adsorbate-free surface ( )
r_p	2.5×10^{-6}	1.0×10^{-3}	Particle radius (cm)
$[O_3]_{g,0}$	1.0×10^{11}	1.0×10^{15}	Initial gas phase number concentration of ozone (cm^{-3})
$[OL]_{b,0}$	1.0×10^{19}	2.0×10^{21}	Initial bulk number concentration of oleic acid (cm^{-3})

The respective lower and upper boundaries indicate the initial constraints of the fit ensemble and an estimate of experimentally accessible conditions in a laboratory for atmospheric aerosol chemistry

calculate a non-normalized ensemble spread of predicted reduction potentials for a set of autogenerated quinone structures.

Results and discussion

Acquisition of fit ensembles

We demonstrate the applicability of the numerical compass (NC) method for the heterogeneous ozonolysis of oleic acid aerosols using the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB), and a neural network surrogate model (SM) for it. Both models map seven kinetic and three experimental input parameters (Table 1) onto the concentration-time profile of oleic acid. For each model, we obtained fit ensembles (n_{FE}=500) in compliance with seven experimental data sets [8, 66–68] as shown in Fig. 3. Each kinetic parameter set in the fit ensemble is associated with one model output (gray lines) for each experimental condition. Both fit ensembles (of KM-SUB and the SM) have a minimal mean-squared logarithmic error (MSLE) of 0.0085; the median MSLE are 0.0102 for KM-SUB and 0.0099 for the SM.

Ensemble spread

The ensemble spread aims for general minimization of the solution space of a model. Figure 4 displays constraint potential maps for the ensemble spread metric and the variable experimental parameters of particle radius (r_p) and ozone concentration ($[O_3]_{g,0}$). The conditions associated with the experimental data used to obtain the fit ensemble (black crosses) are, naturally, located in areas of low ensemble spread. Maxima of the ensemble spread, i.e., regions associated with large model variance, occur at very low particle radii (< 50 nm), and for the combination of large radii (> 10 μm) with high ozone concentrations (> 100 ppm). The constraint potential maps obtained with the KM-only approach (panel A) and the KM/SM-hybrid approach (panel B) appear similar overall. The absolute ensemble spread maxima are both located at maximal particle radii and ozone concentrations (purple crosses). As main difference, isopleths appear less smooth for the SM. A constraint potential map of the SM-only approach is displayed in Additional file 1: Fig. S7. The computationally less expensive SM-only method leads to slightly larger differences to the KM-SUB constraint potential map. In particular, the ensemble spread maximum at low particle radii is less pronounced.

Parameter boundary constraint potential

In addition to the ensemble spread, we apply the NC using both models with the parameter constraint potential (section Parameter boundary constraint potential). This method aims for a minimization of a chosen kinetic parameter's uncertainty range in the solution space, approximated through its 5-95 percentile range in the fit ensemble. Figure 5A and C display parameter constraint potential maps for the kinetic parameters k_{SLR} and $D_{b,OL}$, respectively. The maximum of the k_{SLR} constraint potential matches the maximum of the ensemble spread at low particle radii in Fig. 4, whereas the maximum of the $D_{b,OL}$ constraint potential matches the maximum of the ensemble spread at large radii and high ozone concentrations. Hence, high ensemble spreads appear to be necessary but not sufficient conditions for high parameter constraint potentials.

We simulate the suggested experiments with KM-SUB, using the best fit in the KM-SUB fit ensemble as simulated truth. Under consideration of the original data and the new synthetic experiment, we filter the fit ensembles using the MSLE threshold of $\theta = 0.0105$. Figure 5B and

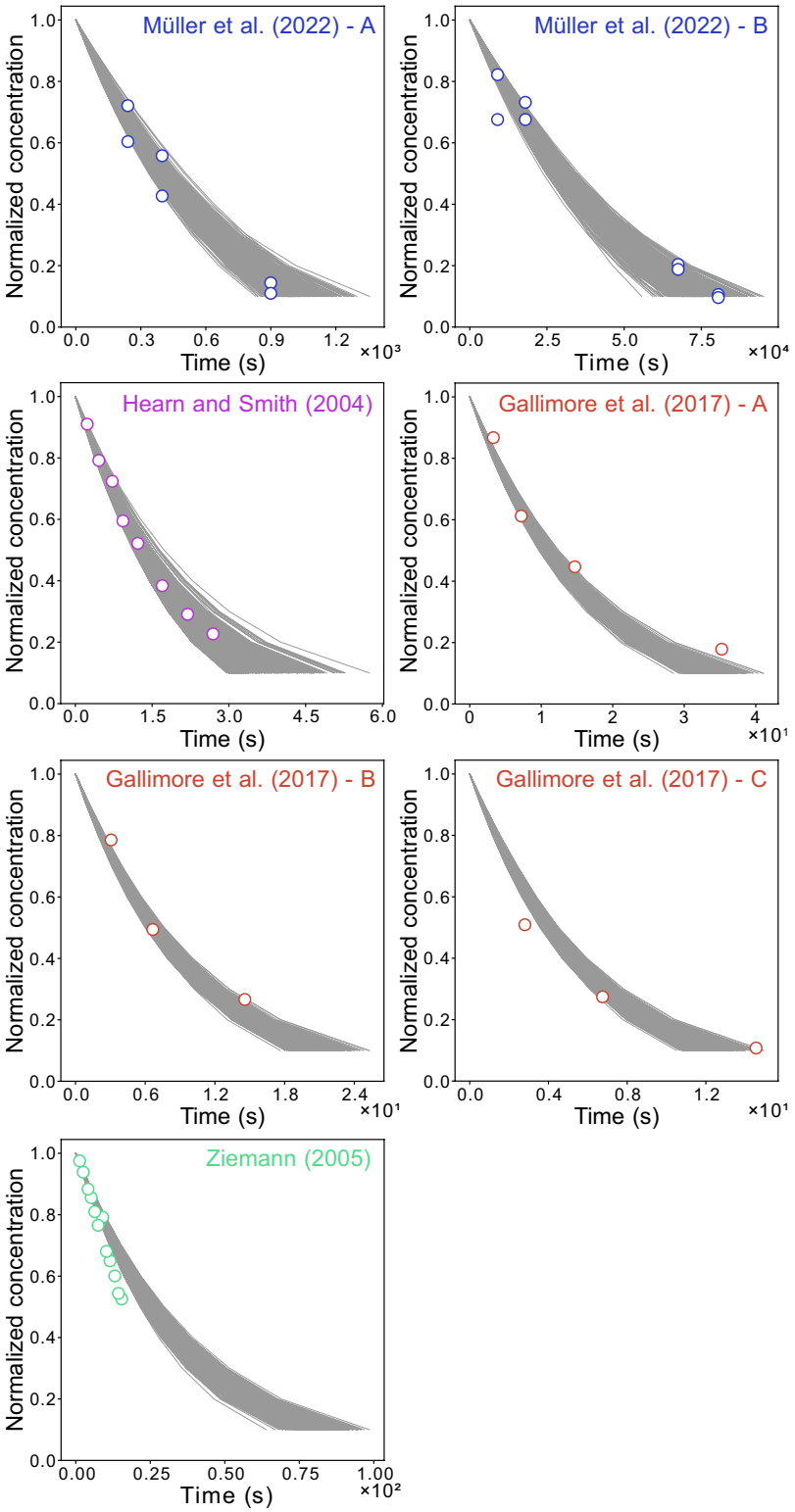

Fig. 3 Ensembles of kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) outputs (n_{FE} = 500, gray lines) with a mean square logarithmic error (MSLE) < 0.0105 in comparison with seven literature data sets (markers) of oleic acid aerosol ozonolysis displayed as normalized oleic acid concentrations ($N_{OL,t}/N_{OL,0}$)

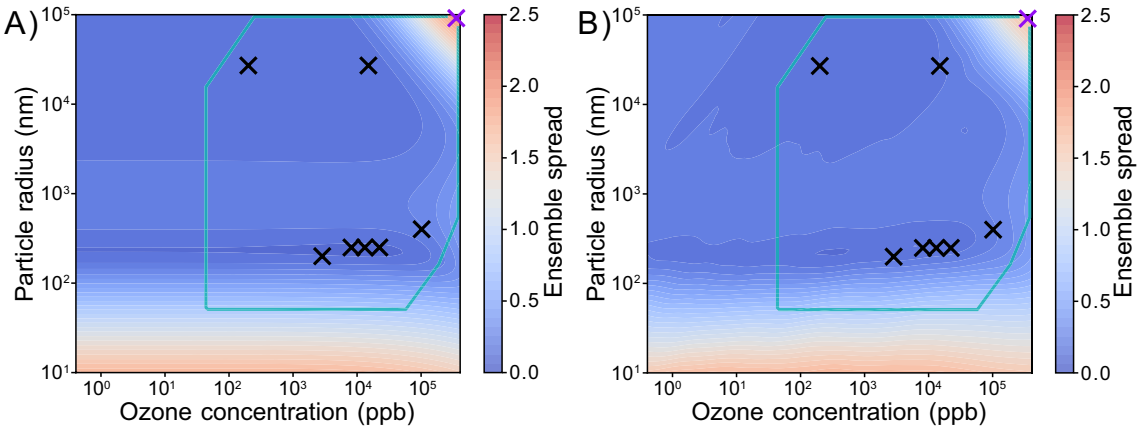

Fig. 4 Constraint potential maps for the ensemble spread, evaluated by **A** KM-SUB (KM-only approach) and **B** SM, based on the KM-SUB fit ensemble (KM/SM-hybrid approach). The teal box outlines conditions for feasible experiments. Black crosses represent the experimental parameters of the seven real experiments that are used for the initial acquisition of the fit ensemble. Purple crosses represent the ensemble spread maximum in each grid with satisfied experimental constraint conditions

D show frequency distributions of five kinetic parameters in the fit ensemble before (blue) and after (red) fit filtering. The experiments suggested by the constraint potential metrics achieve a significant reduction in the 5-95 percentile range for their associated parameters, k_{SLR} and $D_{b,OL}$, respectively. Simultaneously, constraints are achieved for other parameters, e.g., k_{BR} (Fig. 5B), following the similarity between the parameter constraint potential maps (Additional file 1: Fig. S8A, D, G, J). Parameter constraint potential maps and simulated constraints for the SM-only approach (Additional file 1: Fig. S9) are very similar to those using the KM-only approach.

Empirical testing

The NC can be applied repeatedly to narrow down model solutions in iterative fashion. Here, we simulate this procedure using synthetic experimental data, which is obtained by assuming that a single fit from the fit ensemble is the true solution of the modelled system (the *simulated truth*). The simulation is repeated for each fit in the ensemble as simulated truth. Detailed information on the simulation of experimental data is presented in Additional file 1: Note S7.

Figure 6 shows the statistics of a total of 500 of these simulations with three iterations of the NC, and compares the performance of four numerical experiment selection methods: ensemble spread using KM-SUB (blue), ensemble spread using the KM/SM-hybrid approach (orange), random selection (green), and total sensitivities with respect to KM-SUB parameters (red, Additional file 1: Note S8). Figure 6A shows the decreasing number of accepted fits in the fit ensemble. The median numbers of remaining fits after each of the three iterations are (82.5, 43, 38) for the KM-SUB ensemble spread, (82.5, 45.5, 40) for the KM/SM-hybrid ensemble spread, (435, 373, 320.5) for the random selection, and (182, 172.5, 173.5) for the sensitivity-based experiment selection.

Hence, empirically, the NC leads to a significantly larger constraint of the fit ensemble compared to parameter sensitivity maximization or random selection, irrespective of using the full KM or the SM-assisted hybrid approach. Additional file 1: Figures S11–S14 show examples of individual trajectories of the NC, i.e., simulations including numerical experiment selection, synthetic experimental data generation, and fit filtering. We find that in contrast to constraint

(See figure on next page.)

Fig. 5 Constraint potential maps for the kinetic parameters **A** k_{SLR} and **C** $D_{b,OL}$ obtained with KM-SUB. The gray box outlines conditions for feasible experiments. Black crosses represent the experimental parameter sets of the seven real experiments that are used for the initial acquisition of the fit ensemble. The purple crosses represent the parameter constraint potential maxima with satisfied experimental constraint conditions. The suggested experimental conditions are used to obtain synthetic experimental data by evaluating KM-SUB for the best fit in the KM-SUB fit ensemble. Frequency distributions of five kinetic parameters are shown and highlighted for **B** k_{SLR} and **D** $D_{b,OL}$ in the KM-SUB fit ensemble before (blue) and after (red) fit filtering with acceptance threshold $\theta = 0.0105$. Blue and red dotted lines and arrows visualize the 5-95 percentile range of each distribution

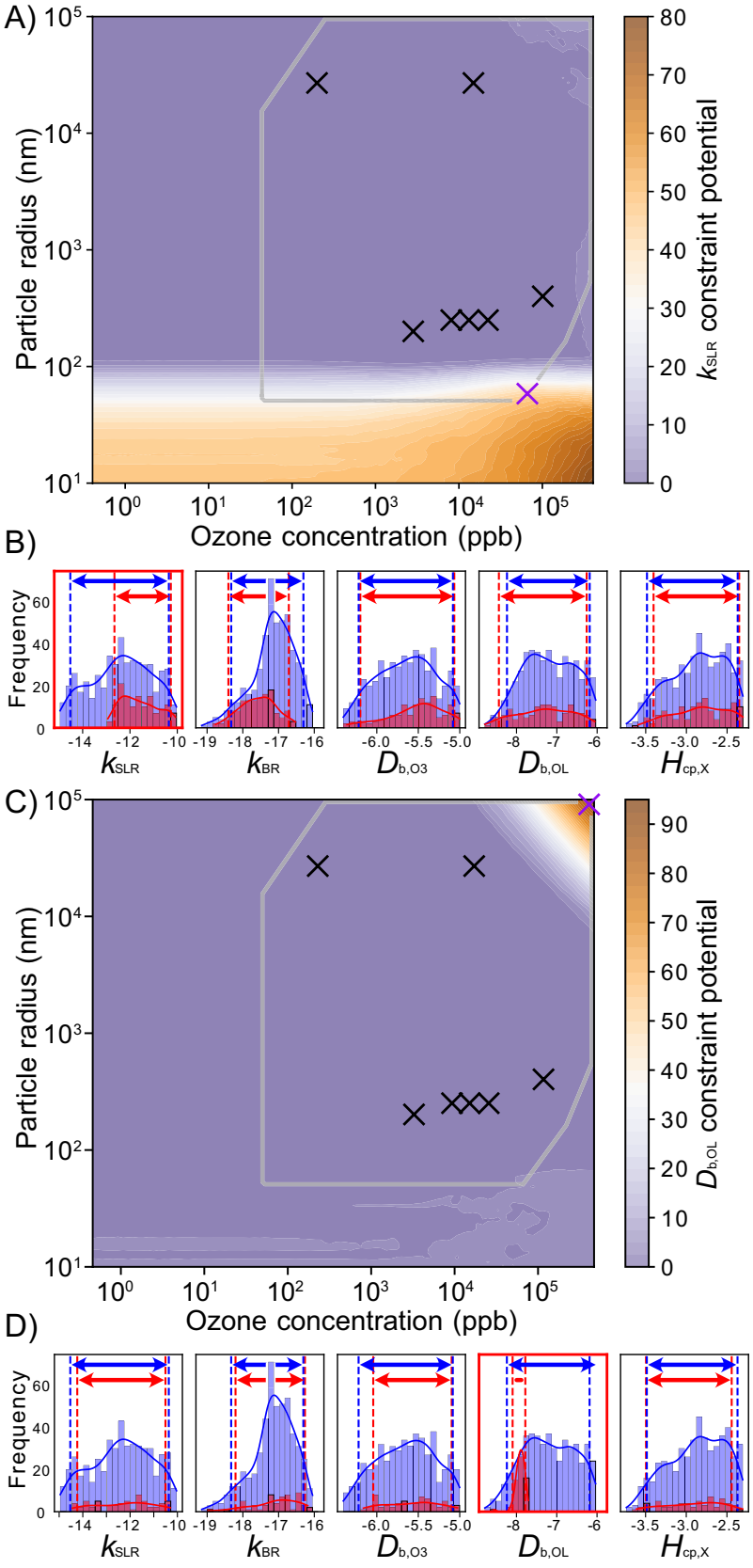

Fig. 5 (See legend on previous page.)

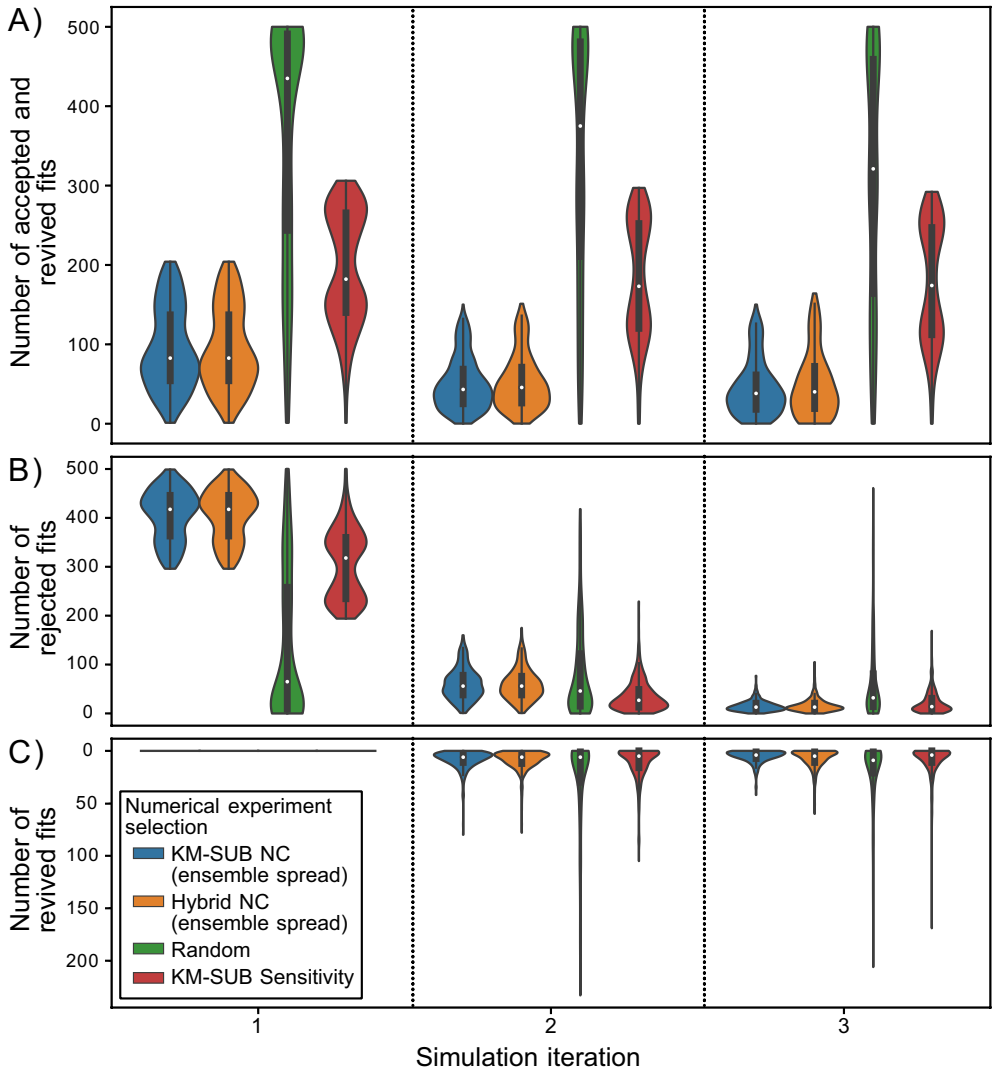

Fig. 6 Number of fits that are **A** accepted, **B** rejected and **C** revived based on synthetic experimental data in three iterations of the numerical compass (NC) method. Numbers are based on statistics for $n = 500$ simulations, where each fit in the KM-SUB fit ensemble is once selected as simulated truth. Medians are shown as white markers, interquartile ranges as vertical wide black lines and 1.5 × interquartile ranges as narrow black lines. While experiment simulation (via KM-SUB) and fit filtering (of the KM-SUB fit ensemble, absolute MSLE threshold, $\theta = 0.0105$) are identical for all approaches, we compare different numerical selection methods of experiments: KM-only NC (blue), KM/SM-hybrid NC (orange), random selection of experiments (green) and parameter sensitivities of the KM (red). The simulation is performed on a reduced 10×10 grid of experimental conditions within the usual ranges. Fit ensemble constraints are significantly larger when experiments are selected using the NC. While the two models utilized for its evaluation lead to very similar fit ensemble constraints, the random and sensitivity-based selection of experiments perform significantly worse

potentials maps, sensitivity maps barely change throughout the iterations of a simulation, and suggested experiments are usually the grid points closest to a persistent sensitivity maximum (Additional file 1: Fig. S15). Consequently, only the first sensitivity-guided experiment leads to a significant constraint of the fit ensemble and, while the performance of the sensitivity-guided method is better than random selection, it performs worse than the ES-guided method of the NC.

Spinning the idea of Fig. 6 further, we can ask: what are the ideal experimental conditions in such a simulation of synthetic experiments? We thus perform a "brute-force" simulation: we repeat the workflow of simulating laboratory experiments for each simulated truth (cf. Fig. 6), but do so for every experimental condition. Instead of the full distribution, we report the median number of rejected fits and plot the results in similar fashion to the constraint potentials into a 2D map (Additional file 1: Fig.

S16B). We find that this map is strongly congruent with the ES map, showing empirically that the experimental conditions associated with the ES maximum are optimal to constrain a fit ensemble. We conducted the same analysis using the PCP metric with similar outcomes, finding major similarities between PCP maps and the maps of reduction of 5-95-percentile ranges for individual parameters in the brute-force simulation, but not to all partial sensitivity maps of individual kinetic parameters (Additional file 1: Fig. S8). Of course, this analysis assumes that there are fits in the fit ensemble that resemble the true solution, which must be ensured when using the compass method by sufficient sampling of the solution space.

Accurate representation of the solution space, especially in the light of experimental error, is contingent on the choice of the acceptance threshold θ. If θ is set too low, a correct solution may be discarded due to incompatibility with a faulty experimental data set. We select a θ in this study so that visual agreement between the scatter in experimental data with the spread of the fit ensemble is achieved. The selection of an appropriate filter threshold is important when quantitative statistical conclusions ought to be drawn for general uncertainty quantification. However, in this context of model optimization or uncertainty minimization through experiments, information is derived by relative comparison of different experimental conditions. This makes the choice of acceptance thresholds for the initial fit acquisition one of practical nature, for example with regards to computational cost [69]. In approximate Bayesian computation, crucial steps like the selection of an acceptance threshold can not be based on general rules, but require testing and evaluation of the performances in the investigated system [70]. Repeating the calculations based on a fit ensemble with an acceptance threshold of $\theta = 0.021$, we found no significant changes in the appearance of constraint potential maps and in the conditions of suggested experiments (Additional file 1: Fig. S17). While absolute values of constraint potential metrics naturally increase with a wider scatter of the ensemble solutions, we find that relative differences between experimental conditions and the locations of constraint potential maxima, denoting suggested experiments, persist across a wide range of acceptance thresholds.

Application to quantitative structure activity relationship (QSAR) model training

Figure 7 shows an exemplary result for quinone structures based on the template 1,2-naphthoquinone which is relevant for atmospheric chemistry and health due to its large reduction potential and ability to undergo redox-cycling. A variety of structures with one or multiple hydroxyl groups is present in the QSAR model's training data, visualized through asterisks in the fields of the heat map. These structures are naturally associated with low ensemble spread values, an indicator for accurate predictions of the QSAR model. Among the newly-generated structures, significant differences in the ensemble spread are observed. In the presented example, structures with a methyl group at position 8, or hydroxyl groups at positions 3 and 4, lead to overall large ensemble spread values of the ensemble predictions. Structures associated with a large ensemble spread may have a larger potential to improve the accuracy of the QSAR models when added to the training data. In basic testing, we find that adding batches of molecules with a high ensemble spread to the model training data generally leads to a much larger improvement of the model compared to adding molecules with a low ensemble spread (Additional file 1: Fig. S18). However, randomly-chosen batches of molecules perform nearly as well, which indicates that more research is needed to optimize the usage of the NC in QSAR applications.

Conclusion

This study demonstrates the application of computational models to guide experiment design and prioritization based on the anticipated reduction of a model's solution space. The method extrapolates current ensemble solutions to conditions of potential future experiments and identifies conditions under which ensemble variance, and thus model parametric uncertainty is largest. In comparison with random selection and selection of experiments associated with maximum sensitivities of kinetic parameters, the reduction of fits in the fit ensemble is much larger for the numerical compass (NC) guided selection of experiments. A disadvantage we find for parameter sensitivities is their lack of variation across the fit ensemble, which makes the sensitivity-guided method mostly agnostic of prior information from experiments.

In contrast to common DOE methods, our proposed statistical approach to experiment design does not require the calculation of Fisher information matrices. This can be advantageous when the model does not permit automatic differentiation or when the computation of numerical gradients is prohibited by computational cost. Furthermore, the novel method is transparent and intuitive: constraints are defined as simple statistical criteria and applied to a tangible fit ensemble, which approximates the solution space. After optimization, the fit ensemble can be used as estimate for the remaining uncertainty of the model solution [19]. The approach can be easily integrated into existing modelling workflows using least-squares parameter estimation and thus offers a low-level entry to experiment design for researchers from various fields.

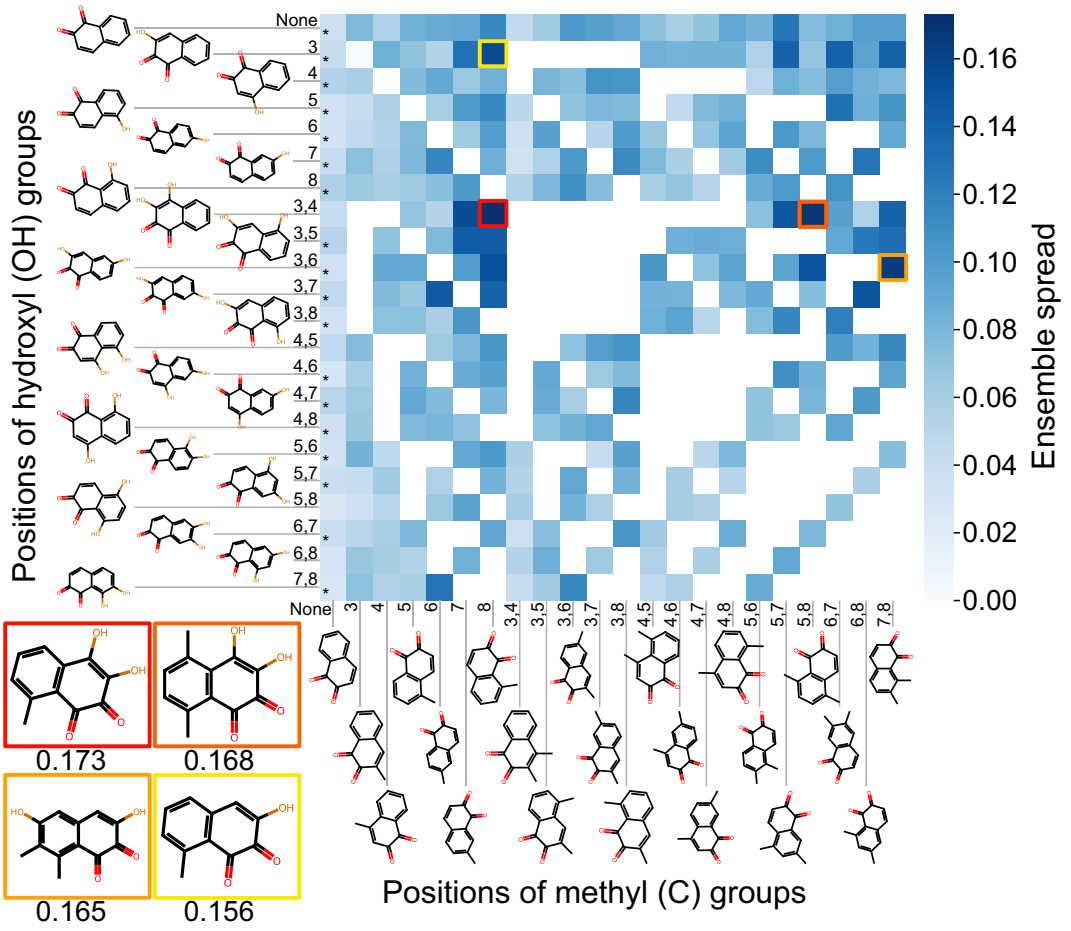

Fig. 7 Heatmap of the non-normalized ensemble spread of QSAR model ensemble predictions for reduction potentials of generated quinone structures based on the template quinone 1,2-naphthoquinone with a maximum of two hydroxyl and methyl groups at varying positions. Ensemble predictions are obtained through 10-fold cross-validation models trained on a data set of 69,599 quinones. Fields marked with '*' are quinones that are present in the training data set. White fields are impossible quinone structures. The molecular structures associated with the four largest ensemble spread values are shown in the bottom left

We find that our method returns near-identical results irrespective of choice of model (KM and SM), fit ensemble (KM and SM fit ensemble) and acceptance threshold for fit ensemble acquisition. This shows the robustness of the method and gives evidence that the properties of the solution space are well-represented by fit ensembles in this study.

Furthermore, the method allows for incorporation of additional information or can be tailored to objectives respective to a specific system, such as chemical kinetic regimes, constraints of specific parameters, or constraints on a specified target condition. We demonstrate this approach by evaluating constraint potentials for individual kinetic parameters (parameter constraint potential; Fig. 5) and by determining optimal experiments for the minimization of model uncertainty under the specific conditions relevant for atmospheric chemistry (target constraint potential; Additional file 1: Fig. S2).

The versatility of the NC is demonstrated through its application on uncertainty quantification of a QSAR model for the prediction of quinone reduction potentials. In analogy to the conditions of kinetic experiments, molecular structures that are associated with high model uncertainty represent potential candidates for future model training. This optimization of training data through uncertainty quantification may be especially useful in organic chemistry, where large quantities of molecules can be generated for computationally-costly density functional theory calculations. In basic tests, we find a correlation of the uncertainty of molecules that are added to the training data and the resulting QSAR model accuracy. However, compared with random selection, only a slight improvement in model accuracy is achieved. Thus, application of the NC for the optimization of QSAR models requires further research and will be the subject of future studies.

The computational effort of the NC can be strongly reduced by training a neural network surrogate model (SM), with nearly identical results. After consideration of the computational effort of SM training, and for the system at hand, we observe an acceleration of the evaluation of the NC by a factor of ~5 using a KM/SM-hybrid approach, and an acceleration by a factor of ~7.5 using only the SM (Additional file 1: Note S9). While SM for multiphase kinetic models have already proven useful in forward modelling applications [33], we here further demonstrate their utility in an inverse modelling approach.

For the kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB) and the heterogeneous ozonolysis of oleic acid, the NC suggests experiments with either very small particles (< 50 nm) or with exceptionally large particles ($\approx 100\ \mu$m) and high ozone concentrations (≈ 1000 ppm) (section Ensemble spread). The first suggestion seems logical: experiments with nano-sized particles of oleic acid have not been conducted and extrapolation to these conditions will be associated with model uncertainty. The method predicts that measurements using nano-sized particles would help especially to constrain the surface reaction rate coefficient k_{SLR}. The second suggestion of the NC may seem counter-intuitive, as these large particle—high ozone conditions are far away from atmospheric relevance. In fact, these experiments likely offer a constraint on the diffusion coefficient of oleic acid, $D_{b,OL}$, a parameter that is rather unimportant under typical atmospheric conditions. Note, however, that the simple model used in this analysis does not consider changes in $D_{b,OL}$ upon formation of oxidation products.

Overall, this analysis of the oleic acid—ozone reaction system shows that additional experiments measuring the loss of oleic acid under conditions typical for the atmosphere will not improve our knowledge of this well-studied system any further. More extreme conditions are needed to narrow down the model solution space, however, this will not come with an improvement of the predictive power of our models for atmospheric conditions (other than small nano-particles). Conversely, any solution in the fit ensemble obtained in this study and in Berkemeier et al. 2021 [19] should perform well under atmospherically-relevant conditions. More knowledge about the system can also be derived by changing the experimental observable. For the heterogeneous ozonolysis of alkenes, for example, product analyses have recently provided additional constraints for kinetic models [68, 71]. Extending the NC from experimental conditions to experimental observables will be a subject of future studies.

Abbreviations

PCP	Parameter boundary constraint potential
DOE	Design of experiments
ENS	Ensemble solution
ES	Ensemble spread
FE	Fit ensemble
KM	Kinetic model
KM-SUB	Kinetic multi-layer model of aerosol surface and bulk chemistry
MSLE	Mean-squared logarithmic error
NC	Numerical compass
OL	Oleic acid
QSAR	Quantitative Structure–Activity Relationship
SM	Surrogate model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00825-0.

Additional file 1: Note S1. Equations for process models, fit ensembles and prediction ensembles. **Note S2.** Equations for ensemble mean and standard deviation. **Note S3.** Parameter boundary constraint potential metric with reduced sample density. **Note S4.** Oleic acid ozonolysis system applied in this study. **Note S5.** Surrogate model training. **Note S6.** Fit ensemble acquisition with KM-SUB and SM. **Note S7.** Uncertainty calibration and simulated experiments. **Note S8.** Sensitivity analysis. **Note S9.** Computational effort. **Figure S1.** Visualization of the parameter constraint potential metric. **Figure S2.** Constraint potential map for the target constraint potential evaluated by KM-SUB. **Figure S3.** Restrictions for constraint potential maps with regards to experimental feasibility. **Figure S4.** Contrariwise cross evaluation of the KM-SUB and SM fit ensembles. **Figure S5.** Scatter plot matrix of the KM-SUB fit ensemble. **Figure S6.** Scatter plot matrix of the SM fit ensemble. **Figure S7.** Constraint potential maps for the ensemble spread, evaluated by KM SUB and SM. **Figure S8.** Comparison of methods to approximate constraints for individual parameters. **Figure S9.** Parameter constraint potential maps evaluated by KM-SUB and the SM. **Figure S10.** Visualization of the uncertainty calibration method. **Figure S11.** Simulated trajectories for iterative NC application. **Figure S12.** Simulated trajectories for iterative NC application. **Figure S13.** Simulated trajectories for iterative NC application. **Figure S14.** Simulated trajectories for iterative NC application. **Figure S15.** Maps of total KM-SUB sensitivity for three iterations of an example simulation for the NC. **Figure S16.** Ensemble spread, median brute-force simulated constraints and total KM-SUB parameter sensitivities. **Figure S17.** Constraint potential map for the ensemble spread evaluated by the SM with a fit ensemble acceptance threshold of 0.021. **Figure S18.** Effect of ensemble spread in additional training data on the QSAR model accuracy of a newly trained model.

Acknowledgements

The authors thank Coraline Mattei for helpful discussions. Parts of this research were conducted using the supercomputer Mogon and/or advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate, www.ahrp.info) and the Gauss Alliance e.V. The authors gratefully acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (hpc.uni-mainz.de). The authors would like to thank two anonymous reviewers for their helpful contributions during peer-review.

Scientific contribution statement

The Numerical Compass method advances the field of computational modelling in the chemical sciences by providing an openly available, versatile tool to determine optimal experimental conditions that are most likely to constrain model parametric uncertainty. In contrast to existing methods, the method does not require maximum likelihood estimation or the optimization of Fisher information matrices. The approach can be easily integrated into existing modelling workflows using least-squares parameter estimation and thus offers a low-level entry to experiment design for researchers in Chemistry and related sciences.

Author contributions
TB conceived the study. MK and TB designed research. TB wrote the kinetic model code and performed simulations. MK wrote the NC, *KineticCompass* package and surrogate model code and performed simulations. All authors discussed and interpreted calculation results. MK and TB wrote the manuscript with contributions from all authors.

Funding
Open Access funding enabled and organized by Projekt DEAL. This work was funded by the Max Planck Society (MPG). AM and MK are supported by the Max Planck Graduate Center with the Johannes Gutenberg University Mainz (MPGC).

Availability of data and materials
The data is openly available at https://doi.org/10.17617/3.D5PCQK.

Code availability
The source code is openly available at https://doi.org/10.17617/3.D5PCQK. The NC is available as package for the programming language Julia (*KineticCompass*) at https://gitlab.mpcdf.mpg.de/mkruege/kineticcompass.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 1 September 2023 Accepted: 10 March 2024
Published online: 22 March 2024

References

1. Worsnop DR, Morris JW, Shi Q, Davidovits P, Kolb CE (2002) A chemical kinetic model for reactive transformations of aerosol particles: reactive transformation of aerosol particles. Geophys Res Lett. 29(20):57–1574. https://doi.org/10.1029/2002GL015542
2. Pöschl U, Rudich Y, Ammann M (2007) Kinetic model framework for aerosol and cloud surface chemistry and gas-particle interactions - Part 1: General equations, parameters, and terminology. Atmos Chem Phys. 7(23):5989–6023. https://doi.org/10.5194/acp-7-5989-2007
3. Kolb CE, Cox RA, Abbatt JPD, Ammann M, Davis EJ, Donaldson DJ, Garrett BC, George C, Griffiths PT, Hanson DR, Kulmala M, McFiggans G, Pöschl U, Riipinen I, Rossi MJ, Rudich Y, Wagner PE, Winkler PM, Worsnop DR, O'Dowd CD (2010) An overview of current issues in the uptake of atmospheric trace gases by aerosols and clouds. Atmos Chem Phys. 10(21):10561–10605. https://doi.org/10.5194/acp-10-10561-2010
4. Abbatt JPD, Lee AKY, Thornton JA (2012) Quantifying trace gas uptake to tropospheric aerosol: recent advances and remaining challenges. Chem Soc Rev. 41:6555–6581. https://doi.org/10.1039/C2CS35052A
5. Shiraiwa M, Pfrang C, Pöschl U (2010) Kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB): the influence of interfacial transport and bulk diffusion on the oxidation of oleic acid by ozone. Atmos Chem Phys. 10(8):3673–3691. https://doi.org/10.5194/acp-10-3673-2010
6. Shiraiwa M, Pfrang C, Koop T, Pöschl U (2012) Kinetic multi-layer model of gas-particle interactions in aerosols and clouds (KM-GAP): linking condensation, evaporation and chemical reactions of organics, oxidants and water. Atmos Chem Phys. 12(5):2777–2794. https://doi.org/10.5194/acp-12-2777-2012
7. Roldin P, Eriksson AC, Nordin EZ, Hermansson E, Mogensen D, Rusanen A, Boy M, Swietlicki E, Svenningsson B, Zelenyuk A, Pagels J (2014) Modelling non-equilibrium secondary organic aerosol formation and evaporation with the aerosol dynamics, gas- and particle-phase chemistry kinetic multilayer model ADCHAM. Atmos Chem Phys. 14(15):7953–7993. https://doi.org/10.5194/acp-14-7953-2014
8. Gallimore PJ, Griffiths PT, Pope FD, Reid JP, Kalberer M (2017) Comprehensive modeling study of ozonolysis of oleic acid aerosol based on real-time, online measurements of aerosol composition: organic aerosol model and measurements. J Geophys Res Atmos. 122(8):4364–4377. https://doi.org/10.1002/2016JD026221
9. Wilson KR, Prophet AM, Willis MD (2022) A kinetic model for predicting trace gas uptake and reaction. J Phys Chem A 126(40):7291–7308. https://doi.org/10.1021/acs.jpca.2c03559
10. Milsom A, Lees A, Squires AM, Pfrang C (2022) MultilayerPy (v1.0): a Python-based framework for building, running and optimising kinetic multi-layer models of aerosols and films. Geosci Model Dev. 15(18):7139–7151. https://doi.org/10.5194/gmd-15-7139-2022
11. Tsuchiya M, Ross J (2001) Application of genetic algorithm to chemical kinetics: systematic determination of reaction mechanism and rate coefficients for a complex reaction network. J Phys Chem A 105(16):4052–4058. https://doi.org/10.1021/jp004439p
12. Berkemeier T, Huisman AJ, Ammann M, Shiraiwa M, Koop T, Pöschl U (2013) Kinetic regimes and limiting cases of gas uptake and heterogeneous reactions in atmospheric aerosols and clouds: a general classification scheme. Atmos Chem Phys. 13(14):6663–6686. https://doi.org/10.5194/acp-13-6663-2013
13. Taylor CJ, Booth M, Manson JA, Willis MJ, Clemens G, Taylor BA, Chamberlain TW, Bourne RA (2021) Rapid, automated determination of reaction models and kinetic parameters. Chem Eng J. 413:127017. https://doi.org/10.1016/j.cej.2020.127017
14. Willis MD, Wilson KR (2022) Coupled interfacial and bulk kinetics govern the timescales of multiphase ozonolysis reactions. J Phys Chem A 126(30):4991–5010. https://doi.org/10.1021/acs.jpca.2c03059
15. Berkemeier T, Ammann M, Krieger UK, Peter T, Spichtinger P, Pöschl U, Shiraiwa M, Huisman AJ (2017) Technical note: Monte Carlo genetic algorithm (MCGA) for model analysis of multiphase chemical kinetics to determine transport and reaction rate coefficients using multiple experimental data sets. Atmos Chem Phys. 17(12):8021–8029. https://doi.org/10.5194/acp-17-8021-2017
16. Tikkanen O-P, Hämäläinen V, Rovelli G, Lipponen A, Shiraiwa M, Reid JP, Lehtinen KEJ, Yli-Juuti T (2019) Optimization of process models for determining volatility distribution and viscosity of organic aerosols from isothermal particle evaporation data. Atmos Chem Phys 19(14):9333–9350. https://doi.org/10.5194/acp-19-9333-2019
17. Wei J, Fang T, Lakey PSJ, Shiraiwa M (2022) Iron-facilitated organic radical formation from secondary organic aerosols in surrogate lung fluid. Environ Sci Technol. 56(11):7234–7243. https://doi.org/10.1021/acs.est.1c04334
18. Milsom A, Squires AM, Ward AD, Pfrang C (2022) The impact of molecular self-organisation on the atmospheric fate of a cooking aerosol proxy. Atmos Chem Phys. 22(7):4895–4907. https://doi.org/10.5194/acp-22-4895-2022
19. Berkemeier T, Mishra A, Mattei C, Huisman AJ, Krieger UK, Pöschl U (2021) Ozonolysis of oleic acid aerosol revisited: multiphase chemical kinetics and reaction mechanisms. ACS Earth Space Chem. 5(12):3313–3323. https://doi.org/10.1021/acsearthspacechem.1c00232
20. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol. 16(12):1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091
21. Nakatani-Webster E, Nath A (2017) Inferring mechanistic parameters from amyloid formation kinetics by approximate Bayesian computation. Biophys J. 112(5):868–880. https://doi.org/10.1016/j.bpj.2017.01.011
22. Tomczak JM, Weglarz-Tomczak E (2019) Estimating kinetic constants in the Michaelis-Menten model from one enzymatic assay using approximate Bayesian computation. FEBS Lett. 593(19):2742–2750. https://doi.org/10.1002/1873-3468.13531
23. Turner BM, Van Zandt T (2012) A tutorial on approximate Bayesian computation. J Math Psychol. 56(2):69–85. https://doi.org/10.1016/j.jmp.2012.02.005
24. Besalú E, Gironés X, Amat L, Carbó-Dorca R (2002) Molecular quantum similarity and the fundamentals of qsar. Acc Chem Res. 35(5):289–295. https://doi.org/10.1021/ar010048x
25. Armeli G, Peters J-H, Koop T (2023) Machine-learning-based prediction of the glass transition temperature of organic compounds using experimental data. ACS Omega 8(13):12298–12309. https://doi.org/10.1021/acsomega.2c08146

26. Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y (2018) Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinform. https://doi.org/10.1186/s12859-018-2523-5
27. Lumiaro E, Todorović M, Kurten T, Vehkamäki H, Rinke P (2021) Predicting gas-particle partitioning coefficients of atmospheric molecules with machine learning. Atmos Chem Phys. 21(17):13227–13246. https://doi.org/10.5194/acp-21-13227-2021
28. Krüger M, Wilson J, Wietzoreck M, Bandowe BAM, Lammel G, Schmidt B, Pöschl U, Berkemeier T (2022) Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects. Nat Sci. 2(4):20220016. https://doi.org/10.1002/ntls.20220016
29. Webb GI, Zheng Z (2004) Multistrategy ensemble learning: reducing error by combining ensemble learning techniques. IEEE Trans Knowl Data Eng. 16(8):980–991. https://doi.org/10.1109/TKDE.2004.29
30. Pradeep P, Povinelli RJ, White S, Merrill SJ (2016) An ensemble model of QSAR tools for regulatory risk assessment. J Cheminform. 8(1):48. https://doi.org/10.1186/s13321-016-0164-0
31. Zhou Z.-H (2021) Ensemble Learning. In: Machine Learning, pp. 181–210. Springer, Singapore . https://doi.org/10.1007/978-981-15-1967-3_8
32. Zhang Y, Menke J, He J, Nittinger E, Tyrchan C, Koch O, Zhao H (2023) Similarity-based pairing improves efficiency of siamese neural networks for regression tasks and uncertainty quantification. J Cheminform. 15(1):75. https://doi.org/10.1186/s13321-023-00744-6
33. Berkemeier T, Krüger M, Feinberg A, Müller M, Pöschl U, Krieger UK (2023) Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks. Geosci Model Dev. 16(7):2037–2054. https://doi.org/10.5194/gmd-16-2037-2023
34. O'Gorman PA, Dwyer JG (2018) Using machine learning to parameterize moist convection: potential for modeling of climate, climate change, and extreme events. J Adv Model Earth Syst. 10(10):2548–2563. https://doi.org/10.1029/2018MS001351
35. Rasp S, Pritchard MS, Gentine P (2018) Deep learning to represent subgrid processes in climate models. Proc Natl Acad Sci USA 115(39):9684–9689. https://doi.org/10.1073/pnas.1810286115
36. Keller CA, Evans MJ (2019) Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10. Geosci Model Dev. 12(3):1209–1225. https://doi.org/10.5194/gmd-12-1209-2019
37. Lu D, Ricciuto D (2019) Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques. Geosci Model Dev. 12(5):1791–1807. https://doi.org/10.5194/gmd-12-1791-2019
38. Kelp M.M, Jacob D.J, Kutz J.N, Marshall J.D, Tessum C.W (2020) Toward stable, general machine-learned models of the atmospheric chemical system. J Geophys Res Atmos. https://doi.org/10.1029/2020JD032759
39. Harder P, Watson-Parris D, Stier P, Strassel D, Gauger NR, Keuper J (2022) Physics-informed learning of aerosol microphysics. Environ Data Sci 1:20. https://doi.org/10.1017/eds.2022.22
40. Sturm PO, Wexler AS (2022) Conservation laws in a neural network architecture: enforcing the atom balance of a Julia-based photochemical model (v0.2.0). Geosci Model Dev. 15(8):3417–3431. https://doi.org/10.5194/gmd-15-3417-2022
41. McBride K, Sundmacher K (2019) Overview of surrogate modeling in chemical process engineering. Chem Ing Tech. 91(3):228–239. https://doi.org/10.1002/cite.201800091
42. Yan S, Minsker B (2011) Applying dynamic surrogate models in noisy genetic algorithms to optimize groundwater remediation designs. J Water Resour Plann Manage. 137(3):284–292. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000106
43. Razavi S, Tolson BA, Burn DH (2012) Review of surrogate modeling in water resources. Water Resour Res. https://doi.org/10.1029/2011WR011527
44. Wan X, Pekny JF, Reklaitis GV (2005) Simulation-based optimization with surrogate models-application to supply chain management. Comput Chem Eng. 29(6):1317–1328. https://doi.org/10.1016/j.compchemeng.2005.02.018
45. Sullivan TJ (2015) Introduction to uncertainty quantification, vol 63. Springer, Cham Heidelberg New York Dordrecht London
46. Weissman SA, Anderson NG (2015) Design of Experiments (DoE) and process optimization. A review of recent publications. Org Process Res Dev. 19(11):1605–1633. https://doi.org/10.1021/op500169m
47. Chaloner K, Verdinelli I (1995) Bayesian experimental design: a review. Statist Sci. https://doi.org/10.1214/ss/1177009939
48. Huan X, Marzouk YM (2013) Simulation-based optimal Bayesian experimental design for nonlinear systems. J Comput Phys. 232(1):288–317. https://doi.org/10.1016/j.jcp.2012.08.013
49. Li S, Tao T, Wang J, Yang B, Law CK, Qi F (2017) Using sensitivity entropy in experimental design for uncertainty minimization of combustion kinetic models. Proc Combust Inst. 36(1):709–716. https://doi.org/10.1016/j.proci.2016.07.102
50. Bisetti F, Kim D, Knio O, Long Q, Tempone R (2016) Optimal Bayesian experimental design for priors of compact support with application to shock-tube experiments for combustion kinetics. Int J Numer Methods Eng 108(2):136–155. https://doi.org/10.1002/nme.5211
51. Wang J, Li S, Yang B (2018) Combustion kinetic model development using surrogate model similarity method. Combust Theory Model. 22(4):777–794. https://doi.org/10.1080/13647830.2018.1454607
52. Franceschini G, Macchietto S (2008) Model-based design of experiments for parameter precision: state of the art. Chem Eng Sci. 63(19):4846–4872. https://doi.org/10.1016/j.ces.2007.11.034
53. Sheen DA, Manion JA (2014) Kinetics of the reactions of H and CH_3 Radicals with n- Butane: an experimental design study using reaction network analysis. J Phys Chem A 118(27):4929–4941. https://doi.org/10.1021/jp5041844
54. Lehn FV, Cai L, Pitsch H, (2021) Iterative model-based experimental design for efficient uncertainty minimization of chemical mechanisms. Proc Combust Inst. 38(1):1033–1042. https://doi.org/10.1016/j.proci.2020.06.188
55. Zhou Z, Lin K, Wang Y, Wang J, Law CK, Yang B (2022) OptEx: an integrated framework for experimental design and combustion kinetic model optimization. Combust Flame 245:112298. https://doi.org/10.1016/j.combustflame.2022.112298
56. Hu Z, Ao D, Mahadevan S (2017) Calibration experimental design considering field response and model uncertainty. Comput Methods Appl Mech Eng. 318:92–119. https://doi.org/10.1016/j.cma.2017.01.007
57. Jung Y, Lee I (2021) Optimal design of experiments for optimization-based model calibration using Fisher information matrix. Reliab Eng Syst Saf. 216:107968. https://doi.org/10.1016/j.ress.2021.107968
58. Atkinson A, Donev A, Tobias R (2007) Optimum experimental designs, with SAS, vol 34. OUP Oxford, Oxford
59. Spall JC (2005) Monte Carlo computation of the fisher information matrix in nonstandard settings. J Comput Graph Stat 14(4):889–909. https://doi.org/10.1198/106186005X78800
60. Griesse R, Walther A (2004) Evaluating gradients in optimal control: continuous adjoints versus automatic differentiation. J Optim Theory Appl 122:63–86. https://doi.org/10.1023/B:JOTA.0000041731.71309.f1
61. Spall JC (2005) Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley, Hoboken
62. Das S, Spall J.C, Ghanem R (2007) Efficient Monte Carlo computation of Fisher information matrix using prior information, 242–249 . https://doi.org/10.1145/1660877.1660912
63. Myung IJ (2003) Tutorial on maximum likelihood estimation. J Math Psychol 47(1):90–100. https://doi.org/10.1016/S0022-2496(02)00028-7
64. Whitaker JS, Loughe AF (1998) The relationship between ensemble spread and ensemble mean skill. Mon Weather Rev. 126(12):3292–3302. https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2
65. Tabor DP, Gómez-Bombarelli R, Tong L, Gordon RG, Aziz MJ, Aspuru-Guzik A (2019) Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries. J Mater Chem A 7(20):12833–12841. https://doi.org/10.1039/c9ta03219c
66. Hearn JD, Smith GD (2004) Kinetics and product studies for ozonolysis reactions of organic particles using aerosol CIMS. J Phys Chem A 108(45):10019–10029. https://doi.org/10.1021/jp0404145
67. Ziemann PJ (2005) Aerosol products, mechanisms, and kinetics of heterogeneous reactions of ozone with oleic acid in pure and mixed particles. Faraday Discuss. 130:469. https://doi.org/10.1039/b417502f
68. Müller M, Mishra A, Berkemeier T, Hausammann E, Peter T, Krieger UK (2022) Electrodynamic balance-mass spectrometry reveals impact of oxidant concentration on product composition in the ozonolysis of oleic

acid. Phys Chem Chem Phys. 24(44):27086–27104. https://doi.org/10.1039/D2CP03289A
69. Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J (2016) Fundamentals and recent developments in approximate Bayesian computation. Syst Biol. https://doi.org/10.1093/sysbio/syw077
70. Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Mol Ecol. 19(13):2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x
71. Reynolds R, Ahmed M, Wilson KR (2023) Constraining the reaction rate of criegee intermediates with carboxylic acids during the multiphase ozonolysis of aerosolized alkenes. ACS Earth Space Chem. 7(4):901–911. https://doi.org/10.1021/acsearthspacechem.3c00026

Publisher's Note

2.6. Towards an annual carbon balance of biological soil crusts: parametric equations and neural networks to model gas exchange and net primary productivity

This chapter contains the final manuscript draft that is currently under peer-review in the journal Functional Ecology. I am the first author and the main contributor to this manuscript. I wrote all code related to the NN model, prepared all figures and wrote the manuscript together with Bettina Weber and Thomas Berkemeier. More detailed information on the author contributions are provided at the end of the manuscript.

Krüger, M., Alves R.P., Kratz, A., Weber, B., Berkemeier, T.: Towards an annual carbon balance of biological soil crusts: parametric equations and neural networks to model gas exchange and net primary productivity, ***under review.***

Two methods were developed and applied to model the physiological response, specifically CO_2 gas exchange rates, of biological soil crusts as a function of soil moisture, temperature and light intensity. The models are a parametric equation with optimized fitting parameters, and an NN model. Both methods were applied to two types of biocrusts, a cyanobacteria- and a lichen-dominated biocrust, using laboratory measurements of CO_2 gas exchange rates as training data. The models achieved very good agreement with independent test data and permitted detailed insights into the physiological response of biocrusts to environmental conditions. It was also demonstrated how such models can be used alongside field measurements of micrometeorological conditions in order to calculate the net primary productivity of biocrusts in specific locations. The supplement to this work can be found in appendix B5.

Abstract

Biological soil crusts (biocrusts) are communities of photoautotrophic and heterotrophic organisms forming a common ecological feature in dryland areas across the globe. Their ability to fix atmospheric carbon may coin them as an important factor for carbon balance and cycling, yet, the quantification of the net primary production of various biocrust types in natural environments remains largely uncertain. Therefore, the physiological response of biocrusts, as related to CO_2 gas exchange is a key area of investigation using both laboratory and modelling approaches. We present two methods to model the physiological response, specifically CO_2 gas exchange rates of biocrusts as a function of soil moisture, temperature and light intensity. The models are a parametric equation with optimized fitting parameters, and an artificial neural network model. Both methods are applied to two types of biocrusts, a cyanobacteria- and a lichen-dominated biocrust, using laboratory measurements of CO_2 gas exchange rates as training data. Our models achieve very good agreement with independent test data and permit detailed insights into the physiological response of biocrusts to environmental conditions. As the models are not mechanistic, they can easily be applied to other organisms or environmental parameters in a similar fashion. We also demonstrate how such models can be used alongside field measurements of micrometeorological conditions in order to calculate the net primary productivity of biocrusts in specific locations.

Introduction

Biological soil crusts (biocrusts) occur globally, forming a regular feature of dryland soils or wherever dry microclimatic conditions occur (Pointing and Belnap, 2012; Belnap et al., 2016). They comprise communities of differing proportions of photoautotrophic cyanobacteria, algae, lichens, and bryophytes, which grow together with heterotrophic bacteria, archaea, and fungi, as well as microfauna feeding on the organisms and their metabolic compounds (Darby and Neher, 2016; Dumack et al., 2016; Maier et al., 2018). They live in, or immediately on top of, the uppermost millimeters of soil (Weber et al., 2022). Biocrusts are known to fulfill various ecosystem services, as they stabilize surface soils, thus reducing erosion by wind and water, fix atmospheric nitrogen, and influence soil water cycling as well as the germination and growth of vascular plants (Belnap, 2006; Brankatschk et al., 2013; Barger et al., 2016; Belnap and Büdel, 2016; Chamizo et al., 2016; Li et al., 2016; Zhang et al., 2016; Rodríguez-Caballero et al., 2018).

Biocrusts also fix atmospheric carbon (C), mostly assessed by means of short-term CO_2 gas exchange measurements investigating the C uptake and respiration rates of biocrusts (Lange et al., 1997; Zaady et al., 2000; Housman et al., 2006; Lange et al., 2006; Büdel et al., 2013; Gypser et al., 2016). These studies often assess the maximum

photosynthetic rates of biocrusts under specific environmental conditions or examine their response to environmental parameters such as light, temperature, water content, and ambient CO_2 concentration (Jeffries et al., 1993a,b; Lange et al., 1997, 1998, 1999; Brostoff et al., 2002; Wilske et al., 2008). While such short-term measurements provide important insights into the physiological functioning and environmental adaptation of biocrusts, they are not well suited to assess their long-term net primary production (NPP). Many of these measurements also analyze only the photoautotrophic component, excluding the heterotrophic fraction, which was recently shown to contribute considerably to overall biocrust respiration and thus reduce its NPP (Weber et al., 2012). Moreover, potential abiotic sources of soil CO_2 efflux, such as inorganic carbon sources (including carbonate weathering) and photodegradation, need to be considered and excluded to avoid overestimating biological respiration (Ma et al., 2013; Rey, 2015; Kim et al., 2024). Measurements of the growth rate as a proxy of for C uptake can yield better estimates of long-term carbon balances (Bisbee et al., 2001; Sancho and Pintado, 2004; Armstrong and Bradwell, 2010), but often fail to capture short-term responses and additional processes like erosion and leaching (Coxson et al., 1992). Long-term measurements, as conducted by Lange et al. (Lange, 2002, 2003) are very time-consuming. They pose a viable basis for mathematical modelling approaches (Bader et al., 2010), particularly due to the high-frequency data collection over sufficiently long periods of time.

Thus, despite several attempts to quantify the NPP of biocrusts, there are still major uncertainties. Long-term CO_2 gas exchange measurements in the field can be conducted by means of automated gas exchange systems, but in cuvettes only single samples detached from the surrounding soil can be measured, and by means of field chambers no fully satisfying results could be obtained, yet. Extrapolations from exemplary field measurements, even when combined with long-term climate data, still carry considerable uncertainty. In addition, most mathematical modelling approaches have not sufficiently addressed the complex interdependence of environmental variables that influence CO_2 exchange. To obtain reliable long-term estimates of carbon dynamics in biocrusts, both methodological improvements and integrative approaches are still required.

Mathematical models describing photosynthesis and respiration of photoautotrophs are subject to fundamental research in plant physiology and for applied developments in agriculture and ecology (O. Rauff and Bello, 2015; Sukhova et al., 2021). In plant physiology, models are often used to evaluate responses to stressors, or to identify potential targets for improvements, e.g., with regard to the resistance to such stressors (Serôdio and Lavaud, 2011; Von Caemmerer, 2013; Bennett et al., 2019). To understand the principles that give rise to the responses of interest, mechanistic models are commonly applied (Farquhar et al., 1980; Rubio et al., 2003). These permit the assessment of the importance and relation of individual sub-processes and can be based on detailed knowledge of the principles governing elementary processes (Serôdio and Lavaud, 2011). They are often limited to specific sub-processes, plant species or organs, but can

be linked to address broader research questions (Zhu et al., 2013; Matuszyńska et al., 2019).

In agriculture and ecology, modelling of photosynthesis and respiration often aims for quantitative yield, or NPP predictions in large-scale systems (Adams et al., 2004). Early empirical models have been based on temperature and precipitation (Lieth, 1973), more recent models furthermore include parameters like growing degree-days, soil moisture stress index, or photosynthetically active radiation (Zaks et al., 2007). In recent years, empirical models for the gas exchange of specific biocrusts have been developed, based on both laboratory and on-site measurements of specific biocrust types (Ma et al., 2023; Nikolić et al., 2024). Ma et al. (2023) derived several abiotic and biotic parameters to apply the Farquhar photosynthesis scheme (Farquhar et al., 1980). Nikolić et al. (2024) developed the model PoiCarb by coupling a carbon-dynamics module with a water-dynamics module.

In this study, we present two approaches to model the physiological response of biocrusts, specifically their CO_2 gas exchange, as a function of environmental conditions such as light intensity, soil moisture (i.e. precipitation equivalent), and temperature. We parameterize the models based on the results of factorial lab analyses along the full ecological amplitude of a cyanobacteria- and lichen-dominated biocrust (Weber et al., 2012; Tamm et al., 2018; Weber et al., 2018). The first method is a parametric equation containing selected mathematical terms that describe the effect of each environmental parameter alone or in combination with others (cross-terms). The second method is a shallow artificial neural network (NN) that is trained to predict CO_2 gas exchange from the environmental parameters. In a second step, we employ these models on microclimate data logged at 10-minute intervals to obtain the potential physiological activity data of the biocrust types. Thus, once the model has been established, microclimate data are sufficient to calculate detailed long-term NPP.

Materials and methods

In this work, we present two empirical models for the quantitative description of CO_2 gas exchange in biocrusts as a function of three independent parameters: light intensity, soil moisture (as precipitation equivalent) and temperature. The empirical models are a parametric equation (Sec.) and an artificial neural network (NN; Sec.). The trained models are evaluated and compared in terms of accuracy, training time, and ease of use.

Data set and data pre-processing

The training data used in this study originate from the data set published in Tamm et al. (2018) and consist of laboratory CO_2-gas exchange measurements across a wide range of values for the independent parameters light intensity, measured as photosynthetically active radiation [μmol PAR m^{-2} s^{-1}], soil moisture, acquired as precipitation

equivalent [mm precipitation], and temperature [°C]. The study encompasses three types of biocrusts: one dominated by cyanobacteria and cyanolichens such as *Chroococcidiopsis*, *Pseudanabaena*, *Phormidium*, *Leptolyngbya*, *Microcoleus*, *Nostoc*, and *Collema coccophorum*, from now on referred to as cyano biocrust type. The second biocrust type is dominated by chlorolichens, in particular *Psora crenata* and *Psora decipiens*, referred to as lichen biocrust. The third, moss-dominated biocrust in Tamm et al. (2018) is not used in this study due to significant measurement variability between the three replicates. For details on the sampling site, located in Soebatsfontein, Succulent Karoo, South Africa, and CO_2 gas exchange measurement procedures, please see Tamm et al. (2018). All available data points are indiscriminately mixed without distinction between replicates to account for biological variability in the models. We then apply 5-fold cross validation with 80% training and 20% test data, a statistical technique where data are partitioned into subsets, and multiple models are trained and validated on different combinations of these subsets to ensure robust model evaluation (James et al., 2013). Between the parametric equation and NN model, we use identical data splits to ensure comparability and reproducibility. For the NN models, data are normalized to [0, 1] intervals using Z-standardization. All models are fitted and optimized using the same criterion, a mean square error (MSE) metric for an independent test set that is identical for all models of the same crust type. The field data used for application of the models are part of meso- and microclimate measurements conducted in Soebatsfontein, Succulent Karoo, South Africa, from 17 October 2008 through 16 October 2009. For details on the measurement setup and biocrust wetness probes (BWP), see Weber et al. (2016).

Parametric equation model

Mathematical formula The net photosynthetis NP of biological soil crust samples is parameterized according to Eq. 3 as a function of three independent parameters: light intensity L (in µE), precipitation equivalents M (in mm) and temperature T (in C°). The formula is divided into an independent respiration (R) (Eq. 1), and a gross photosynthesis (P) term (Eq. 2).

$$R = \Big(m_{\mathrm{R},1} \times (M - M_0) + t_{\mathrm{R},1} \times (T + T_0)^{t_{\mathrm{R},2}} \times \exp(-t_{\mathrm{R},3} \times (T + T_0)^{t_{\mathrm{R},4}}) \times \big(1 - \exp(-m_{\mathrm{R},2} \times (M - M_0))\big)\Big) \times \exp\big(-m_{\mathrm{R},3} \times (M - M_0)^{m_{\mathrm{R},4}}\big) \tag{1}$$

$$P = p \times (M - M_0)^{m_{\mathrm{P},1}} \times T^{t_{\mathrm{P},1}} \times (L + 1)^{l_{\mathrm{P},1}} \times \exp(-t_{\mathrm{P},2} \times T^{t_{\mathrm{P},3}}) \times \exp(-m_{\mathrm{P},2} \times (M - M_0)^{m_{\mathrm{P},3} \times \mathrm{T}^{-t_{\mathrm{P},4}}}) \times \Big(1 - \exp\big(-m_{p4} \times (M - M_0)\big)\Big) \times (1 - \exp(-\mathrm{l}_{p2} \times \mathrm{L})) \tag{2}$$

$$NP = P - R \tag{3}$$

The model is designed to reproduce the characteristic shape of net photosynthetic activity of biological soil crusts, which often operate best at a certain soil moisture, temperature and light intensity. Hence, polynomial functions of the form $a * x^y$ are combined with exponential decay functions of the form e^{-a*x^y}, where a and y are fit parameters and x is an independent variable. Since the exponential term dominates the polynomial term for large values of x, this combination of terms is able to describe the characteristic shape of the gas exchange curves: a quick increase followed by a plateau of maximal NP and an asymptotic decrease of the gas exchange rate towards zero. Furthermore, saturation effects are represented by exponential functions of the form 1-e^{-a*x^y}. This accounts for the saturation of photosynthetic activity with increasing light intensity and sometimes also soil moisture. A special term describes the interaction of moisture and temperature dependence by using a temperature-dependent exponent ($y(T)$) inside the exponential decay function for moisture-dependence of the production term of the form $e^{-a*x^{y(T)}}$. The model has a total of 21 soil crust type-dependent fitting parameters. The model parameters are labeled according to the process they describe (R, P) and independent variable (L, M, T) they act upon. For example, $m_{\mathrm{R},1}$ describes moisture-dependence in the respiration term. p is a general scaling factor for the production term. A precipitation offset M_0 is introduced to account for the minimum amount of humidity that is needed to wet the soil in order for the soil crust to take up water. A temperature offset T_0 is introduced in the respiration term to account for activity that begins above 0 °C.

Global optimization of model parameters The model in Eq. 3 interpolates laboratory data of net photosynthetic activity of biological soil crusts as a function of three independent parameters (L, M, T) and 21 sample- and soil crust type-dependent fitting parameters. The fitting parameters are determined using the Monte-Carlo Genetic Algorithm (Berkemeier et al., 2017), which is a two-step global optimization method that combines a random search over the entire optimization hypersurface with a genetic algorithm that uses concepts known from natural evolution to find the global minimum of the optimization problem. The method has previously been applied to inverse modelling problems in atmospheric chemistry (Arangio et al., 2015; Berkemeier et al., 2016; Lakey et al., 2016; Tikkanen et al., 2019; Berkemeier et al., 2020).

The optimization is conducted in a three-step process. The respiration term in Eq. 1 is first optimized to all data measured in the absence of light irradiation. The parameters obtained in this first step are fixed in a second optimization step of the full model to all data that are used to find the remaining model parameters. In a third step, the simplex and golden section search optimization methods (Kiefer, 1953; Morgan and Deming, 1974) are used iteratively on triplets of model parameters to improve overall model-experiment correlation, followed by repeated execution of golden section search for local, one-at-a-time parameter optimization. Numerical values for all fit parameters of each biocrust type are given in Tables 1 and 2.

Table 1: *Fit parameters for the net productivity parameterization (Eqs.* (1)-(3)*), optimized for the cyanobacteria/cyanolichen-dominated biocrust samples in five-fold cross validation.*

Parameter name	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
$m_{R,1}$	1.53e-05	2.43e-01	5.52e-02	1.10	1.97
$m_{R,2}$	2.90e-02	9.14e-02	5.83e-03	1.72e-02	7.55e-03
$m_{R,3}$	2.76	2.25	2.10	2.75	2.19
$m_{R,4}$	7.48e-01	9.26e-01	9.73e-01	7.20e-01	9.84e-01
$t_{R,1}$	3.89e-01	4.34e-01	1.57e-01	6.18e-02	1.72e-01
$t_{R,2}$	1.81	1.49	2.23	2.36	2.27
$t_{R,3}$	3.45e-13	5.87e-16	2.34e-18	7.48e-14	1.45e-01
$t_{R,4}$	1.09	5.63e-01	2.80	1.19	2.33e-10
$m_{P,1}$	2.88e-16	1.17e-01	2.09	2.58e-16	2.07
$m_{P,2}$	3.20	3.58	7.98	3.43	1.12e+01
$m_{P,3}$	9.06e-01	9.42e-01	4.84e-01	8.88e-01	3.96e-01
$m_{P,4}$	9.16	1.50	1.16e+11	2.60	5.47e+01
$t_{P,1}$	1.19	1.15	8.97e-01	1.03	9.23e-01
$t_{P,2}$	4.43e-04	8.48e-05	1.41e-25	2.24e-04	1.58e-25
$t_{P,3}$	1.89	2.25	2.29	1.87	2.28
$t_{P,4}$	6.42e-15	2.33e-16	3.45e-14	4.36e-17	1.44e-13
$l_{P,1}$	3.96e-01	1.74e-01	4.71e-01	2.62e-01	4.50e-01
$l_{P,2}$	1.03	1.01	1.05	1.01	1.04
p	5.62e-02	1.66	6.29	4.96e-01	1.90e+02
M_0	1.42e-01	1.23e-01	8.70e-02	1.38e-01	9.07e-02
T_0	3.36e+01	1.16e+01	3.48e+01	2.73e+01	2.22e+01

Table 2: *Fit parameters for the net productivity parameterization (Eqs.* (1)-(3)*), optimized for the lichen-dominated biocrust samples in five-fold cross validation.*

Parameter name	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
$m_{R,1}$	2.79e-02	3.08e-01	1.67e+01	7.66	3.34
$m_{R,2}$	1.69e-01	3.76	2.43e-01	3.91e-03	5.96e-02
$m_{R,3}$	7.67e-01	2.47e-01	1.39	1.13	1.85
$m_{R,4}$	1.18	5.04e-18	7.06e-01	8.51e-01	5.05e-01
$t_{R,1}$	2.17e-01	6.93e-02	4.80e-01	9.80e-03	2.92
$t_{R,2}$	1.43	1.18	1.36	3.22	1.36
$t_{R,3}$	1.84e-18	6.43e-19	2.76e-22	4.02e-06	3.60e-13
$t_{R,4}$	7.80e-01	1.60	2.26e-01	2.84	1.02
$m_{P,1}$	8.22e-01	7.98e-20	1.54e-16	4.47e-01	2.06e-01
$m_{P,2}$	3.20	2.25	2.10	2.92	2.82
$m_{P,3}$	1.03	3.02	1.40	1.44	1.93
$m_{P,4}$	1.62e+01	3.39	3.59	3.38	2.63
$t_{P,1}$	9.16e-01	1.14	7.15e-01	6.92e-01	9.27e-01
$t_{P,2}$	1.05e-03	3.62e-03	1.29e-04	1.46e-04	1.95e-04
$t_{P,3}$	1.93	1.60	2.40	2.30	2.28
$t_{P,4}$	3.11e-16	2.95e-01	1.00e-15	9.84e-02	2.15e-01
$l_{P,1}$	2.39e-01	6.05e-02	2.19e-01	1.02e-01	1.23e-01
$l_{P,2}$	1.03	1.01	1.02	1.01	1.02
p	1.63	1.82	1.40	8.04	3.72
M_0	8.01e-02	1.45e-01	1.51e-01	1.10e-01	1.40e-01
T_0	1.52e+01	3.76	6.85e-11	1.91e+01	1.41

Table 3: *Neural network hyper-parameter description and tested ranges.*

Hyper-parameter	Description	Tested range
num_layers	Number of layers	[1, 3]
num_nodes	Number of nodes in each layer	[16, 4096]
learning_rate	Learning rate during training	$[1 \times 10^{-5}, 1 \times 10^{-1}]$
acivations	Activation of each layer	'ReLU', 'LeakyReLU'
dropout_rates	Dropout rate for each layer	[0, 0.7]
batch_size	Number data points in each training batch	[2, 32]
num_epochs	Number of training epochs	[20, 60]

Neural network model

As a fully data-centric machine learning approach, we employ shallow, fully-connected, feedforward multilayer-perceptrons with a maximum of three hidden layers. In addition to hidden layers, NN models encompass an in- and output layer with nodes referring to the model's three inputs (L, M, T) and one output (NP), respectively. Each node in the network is connected with all nodes in the previous and following layer, and each connection is associated with a weight that is optimized during training. Activation functions, defining the mathematical operation performed in each node, network architecture, loss function and optimization algorithm are examples for hyper-parameters that can be optimized for a specific training data set. A description of relevant hyper-parameters and tested ranges in this study are summarized in Tab. 3. We use the Python package *optuna* to perform efficient hyper-parameter tuning with a total of 40 trials for each model (Akiba et al., 2019).

Results

Model evaluation and comparison

We trained and evaluated the parametric equation and neural network model to predict CO_2 gas exchange, a key indicator for net photosynthesis (NP) and respiration (R), as a function of light intensity, temperature and soil moisture for two different biocrust types: cyano and lichen biocrusts. For each method, we employed a prediction ensemble consisting of five cross-validation models. Each model was validated on a different data subset with the remaining subsets used as training data. This allows for individual evaluation of each cross-validation model, while the ensemble provides insights into the effects of biological variability and potential biases arising from training data selection. In the following, each environmental parameter is discussed in terms of overall effect on CO_2 gas exchange in both data and model predictions.

Figure 1 shows model prediction ensembles as lines and colored bands and experimental measurements for CO_2 gas exchange (including training and test data) as markers as a function of soil moisture for three temperatures and two light intensities. Both biocrust types exhibit a distinct CO_2 gas exchange maximum under optimum soil moisture in the light. While the position of the maximum as related to soil moisture is constant (cyano: $\sim$0.4 mm, lichen: $\sim$0.5 mm), its height also depends on the temperature. For the cyano biocrust, the maximum CO_2 uptake values are reached at the highest experimental temperature of 32°C, whereas the lichen biocrust exhibits the highest photosynthetic activity at an intermediate temperature of 22°C. In the absence of light (respiration only), a distinct peak in CO_2 emissions can be observed for the cyano crust type at $\sim$0.7 mm precipitation equivalent, whereas the lichen biocrusts maintain a high respiratory activity at high soil moisture.

Both models capture the biocrusts' behavior well, with the majority of outliers resulting from scattering due to biological variability or measurement uncertainty. Ensemble prediction bands (colored areas) are mostly narrow, especially for the parametric equation model, indicating robust models independent of the training data selection. For the cyano biocrust parametric equation model under dark conditions, a large fraction of measurements, especially for the highest temperature, fall below the prediction ensemble, indicating a slight systematic error of the parametric equation model. This is likely due to the parametric equation's design to eventually approach a NP of zero at high humidity for both light and dark conditions. Here, the inherent functional form of the parametric equation imposes limitations on its flexibility to accurately represent experimental data. Since the NN is not constrained in this regard, the NN models predict a monotonic decrease in CO_2 uptake at high temperature. While the NN's behavior is associated with a smaller error, it is likely unphysical when extrapolated to high humidity.

The dependency of CO_2 gas exchange on light intensity is displayed in Fig. 2 in a similar fashion and for a fixed soil moisture roughly corresponding to the bioactivity maximum observed in Fig. 1 (i.e., 0.5 ± 0.05 mm precipitation equivalent). The relationship between light intensity and CO_2 gas exchange can be described as a saturation curve. With increasing irradiation, the lichen biocrusts reach a CO_2 gas exchange of up to 3 μmol CO_2 m^{-2} s^{-1} at a photosynthetic photon flux density (PPFD) of roughly 800 μmol m^{-2} s^{-1}, and level off quickly. For the cyano biocrust, the overall productivity is lower, and saturation does not set in as quickly, especially at high temperatures.

All models fit the data well with minor, unbiased scattering, given the selected precipitation equivalent tolerance of 0.05 mm for displayed experimental measurements. The cyano data exhibit larger scatter at 22 °C with one measurement series falling significantly below the remaining measurements at the same temperature. Interestingly, while the parametric equation falls in between the discordant measurement values, the neural network seems to disregard the measurements with lower gas exchange rate. Larger prediction ensemble spreads can be observed for the NN, especially for the lichen biocrust.

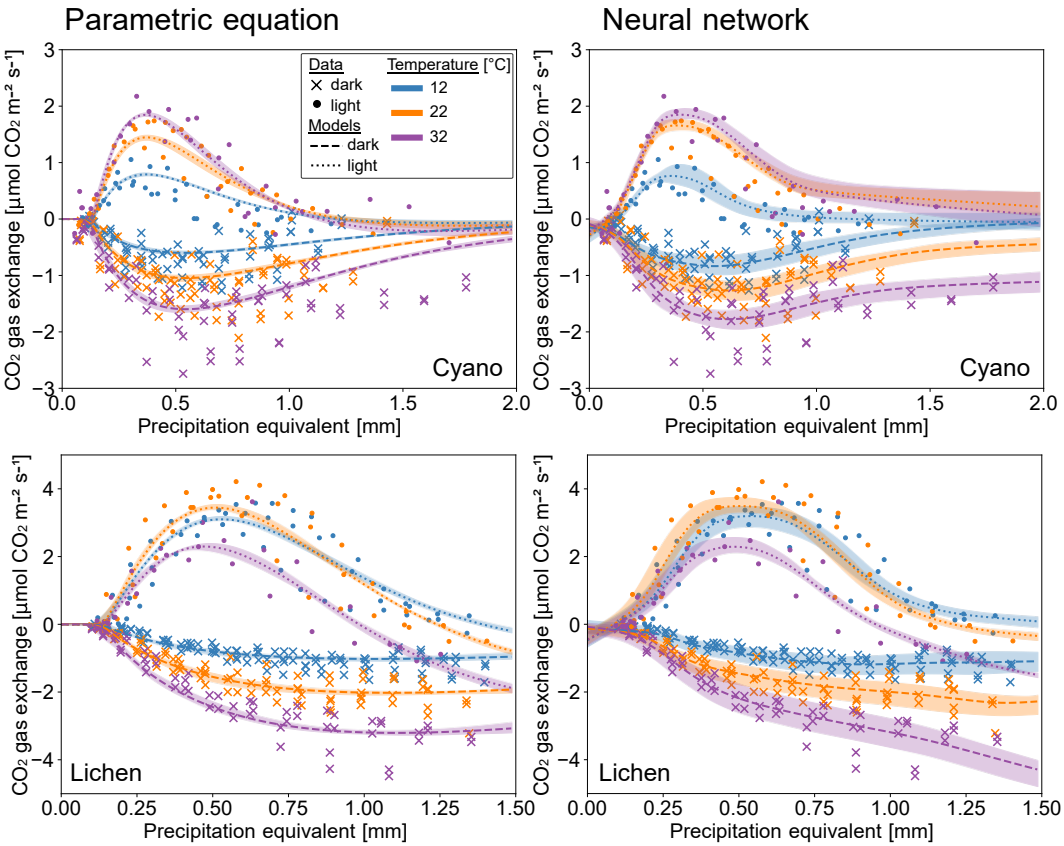

Figure 1: *Respiration and net photosynthesis CO_2 gas exchange measurements for the cyano (top) and lichen (bottom) biocrust, from Tamm et al. (2018), and predictions of the parametric equation (left) and NN (right) models at three temperatures and as a function of soil moisture. Colored areas encompass all predictions of the five cross-validation models trained on different subsets of the full data, while dotted or dashed lines are average predictions. "Light" refers to a light intensity of 1000, "dark" to 0 µmol m^{-2} s^{-1}.*

Figure 3 displays biocrust CO_2 gas exchange as a function of temperature under light and dark conditions for the soil moisture optimum at 0.5 mm precipitation equivalent. In the dark, both biocrust types show a direct, positive correlation between temperature and respiratory activity. At a PPFD of 1000 μmol m^{-2} s^{-1}, the cyano biocrust exhibits a similar positive correlation between temperature and NP within the measurement range, while the lichen biocrust shows a distinct NP maximum at roughly 20°C.

While the neural network accurately aligns with the data at low temperatures, this may to some extent be attributed to over-fitting, as the apparent steep increase of NP between 7 and 12°C observed for the cyano biocrust may simply be due to random scatter in the laboratory measurements. The large number of fitting parameters enables the NN model to follow this behavior and achieve a lower error, however, it is unclear, if this behavior can be generalized to all samples of this biocrust type. In contrast, the

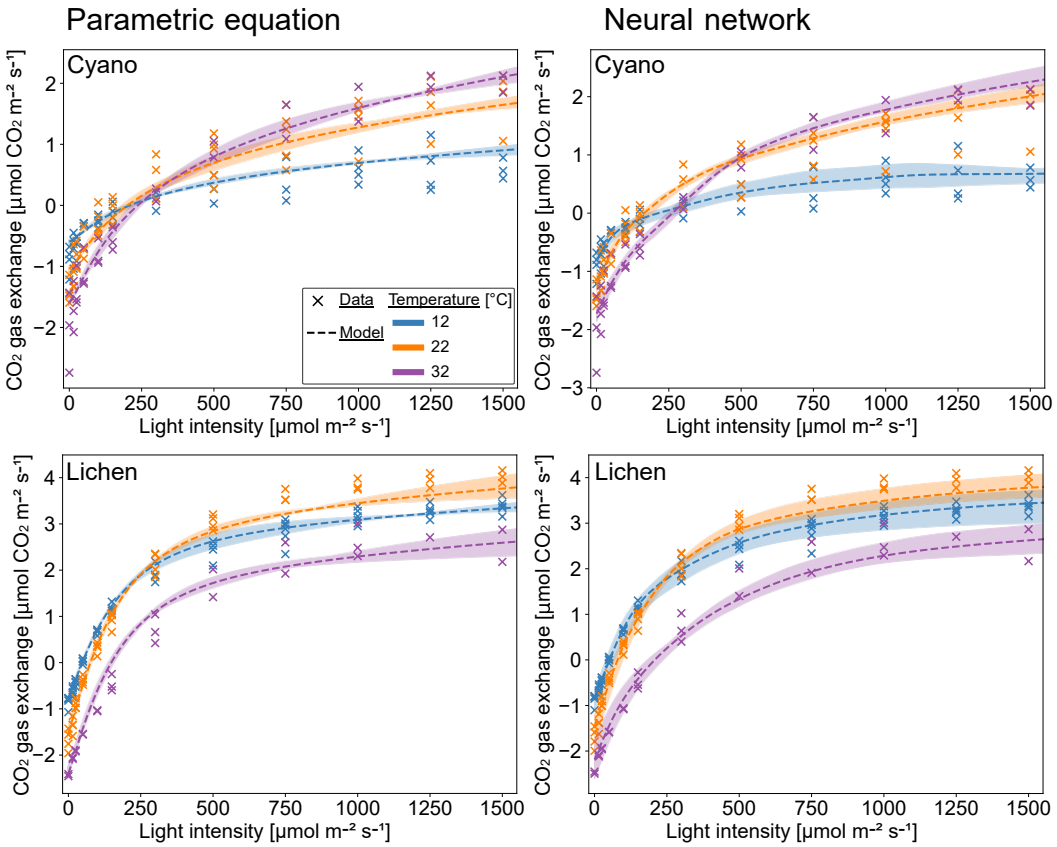

Figure 2: *Respiration and net photosynthesis CO_2 gas exchange measurements for the cyano (top) and lichen (bottom) biocrust type from Tamm et al. (2018), and predictions of the parameterization (left) and NN (right) models across various temperatures as a function of light intensity. Colored areas encompass all predictions of the five cross-validation models trained on different subsets of the full data, dashed lines are average predictions. Humidity is fixed at 0.5 mm precipitation equivalent with a tolerance of 0.05 mm for the displayed experimental measurements.*

functional form of the parametric equation and the much smaller number of fitting parameters limit the model to a possibly more generalizable response function.

In Fig. 4, measured and estimated CO_2 gas exchange are compared with each other for the full data set. Each model prediction originates from the cross-validation model which had the respective data point in the test set, thus, only predictions for unseen data points are shown. Overall, the neural network achieves a slightly better accuracy with a mean square error (MSE) of 0.10 and 0.20 for the cyano and lichen biocrust, respectively, in contrast to 0.11 and 0.20 for the parametric equation. Both models predict smaller absolute measured gas exchange values for sample 3 of the cyano biocrust as a result of variability between samples. For the lichen biocrust, sample one shows a distinct behavior during positive NP, which is not well captured by either model.

The full optimization time of the parametric equation was roughly 50 CPU-hours for five cross-validation models. The training time of NN models was much lower with

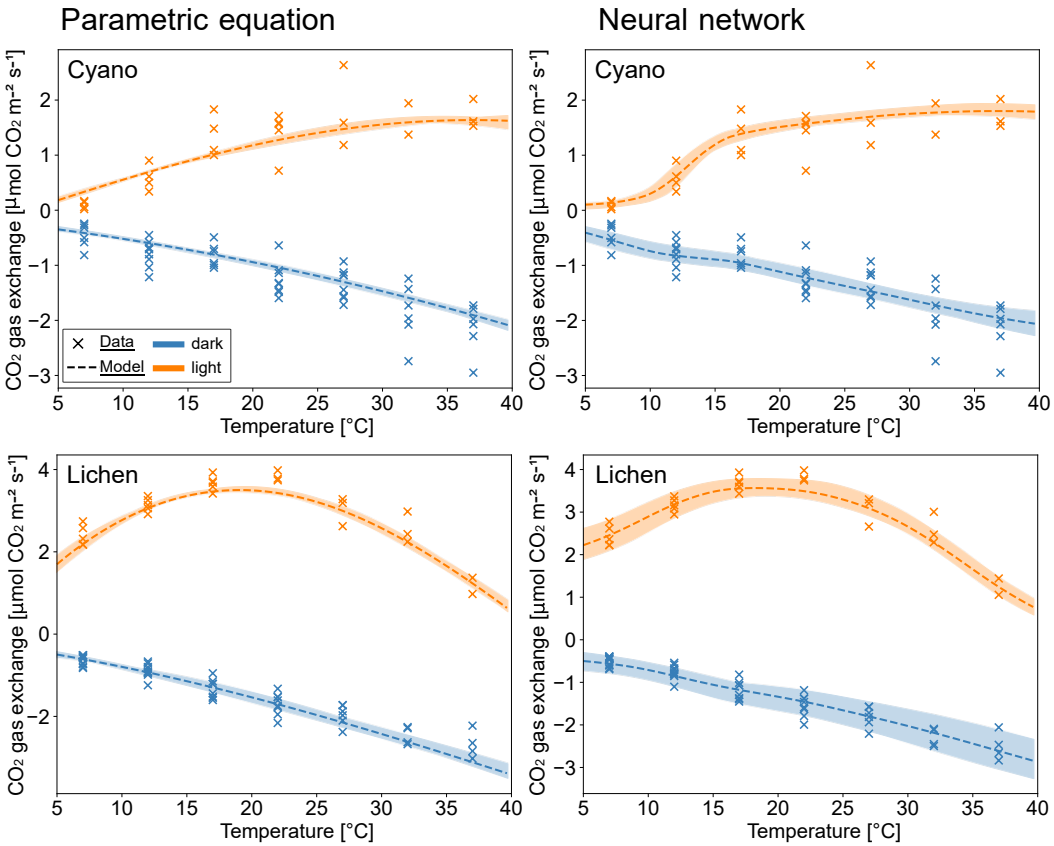

Figure 3: *Respiration and net photosynthesis CO_2 gas exchange measurements for the cyano (top) and lichen (bottom) biocrust from Tamm et al. (2018), and predictions of the parameterization (left) and NN (right) models under light and dark conditions as a function of temparature. Colored areas encompass all predictions of the five cross-validation models trained on different subsets of the full data, dashed lines are average predictions. Humidity is fixed at 0.5 mm precipitation equivalent with a tolerance of 0.05 mm for the displayed experimental measurements. "Light" refers to a light intensity of 1000, "dark" to 0 $\mu mol\ m^{-2}\ s^{-1}$.*

1-2 CPU-hours, depending on the model architecture defined by hyper-parameters. If full hyper-parameter tuning was conducted with 40 trials, the computational effort roughly matched the optimization cost of the parametric equation. NN training time can be drastically reduced through utilization of GPUs and employing minimal hyper-parameter tuning, provided that robust NN model architectures have been previously tested.

Model application on field data

The trained models can be used to estimate the NPP of biocrusts in their natural habitat. We used micrometeorological data from Soebatsfontein, Succulent Karoo, South Africa, encompassing temperature, light intensity, precipitation and soil moisture measurements every 10 minutes. Figure 5 shows an exemplary application of

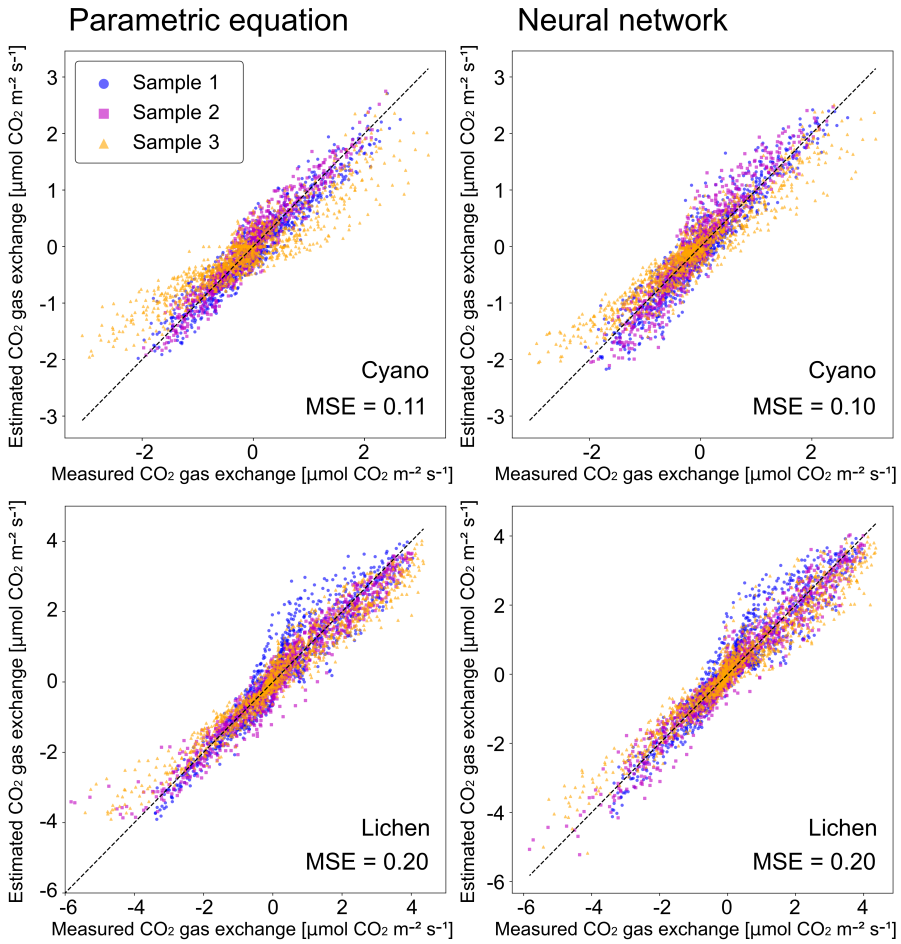

Figure 4: *Correlation scatter plots of measured and predicted CO_2 gas exchange for the cyano (top) and lichen (bottom) biocrust from Tamm et al. (2018). Each cross-validation model of the parameterization (left) and NN (right) is used to predict its respective test set, and all five test set folds are merged in the plots. Mean square errors are averaged across all cross-validation test sets. Plots for the individual cross-validation folds are shown in Fig. [SI].*

the parametric equation and NN models on a subset of 432 time steps (3 days) from June 14 to June 16, 2009. The time period encompasses two rain events leading to high soil moisture in the morning of June 14 and June 16. With the following increase in light intensity and temperature, both models exhibit CO_2 uptake peaks at noon of June 14 and June 16. On June 15, the photosynthetic activity is reduced due to low soil moisture and reduced light intensity. The CO_2 gas exchange amplitude of the lichen biocrust (Fig. S1) is larger in comparison with the cyano biocrust (Fig. 5), which matches the observations made in Fig. 1, 2 and 3. The parametric equation and NN yield overall similar results. While gas exchange predictions are nearly identical during daytime on June 15, the NN predicts higher gas exchange rates for the cyano biocrust on June 16 (Fig. 5). In contrast, for the lichen biocrust, the NN predicts lower gas exchange rates on the same day (Fig. S1). However, we do not find a systematic

Table 4: *Integrated CO_2 gas exchange rate predictions in $\mu mol\ CO_2\ m^{-2}$ based on micrometeorological data from Soebatsfontein, Succulent Karoo, South Africa over 3 days from June 14 to June 16, 2009.*

Model	Biocrust type	June 14	June 15	June 16
Parametric equation	Cyano	2.37	0.86	-5.36
Neural network	Cyano	2.85	-1.88	-3.41
Parametric equation	Lichen	12.47	10.96	0.45
Neural network	Lichen	12.05	10.00	0.33

bias of either estimation method. Furthermore, while the parametric equation model predicts zero NP at low soil moisture, e.g., in the night from June 15 to June 16, for the cyano biocrusts, the NN model shows minor respiratory activity. Here, it may be questioned if the moisture contents are indeed still high enough to support this physiological activity. CO_2 gas exchange rate predictions were integrated over each day using the trapezoidal rule and reported in Tab. 4. For the lichen biocrust, we find positive values for the cumulative gas exchange, indicating net growth during the chosen interval. For the cyano biocrust, the high respiratory activity on June 16 implies net respiration in the same interval, despite very similar micrometeorological conditions. The NN and parametric equation give similar results, however, the conditional fixation of the parametric equation output to 0 at very low soil moisture leads to an overall higher (and thus positive) balance of the cyano biocrust on June 15. This methodology can be used to estimate long-term NPP of soil biocrusts.

Discussion

Our findings suggest that parametric equations and artificial neural network (NN) models are capable of accurately describing the physiological response of biological soil crusts based on multiple environmental parameters. Prediction errors are generally low and mostly a result of biological variability, leading to data scattering and systematic differences between replicates.

When comparing both models, we find the NN models to be more flexible, i.e., more drastic shapes of the response function can be realized due to a larger number of fitting parameters. In contrast, the parameterization model is a bit more stiff, as the underlying mathematical terms limit the function's complexity. Consequently, the NN models achieve equal or slightly lower test set errors than the parametric equation models. However, NN models have a much larger number of fitting parameters and are more likely to overfit, i.e., they may perform well on training data but worse on unseen data due to capturing noise as patterns. Furthermore, the equation model allows for a better investigation and comparison of meaningful parameters, as these relate to relevant sub-processes and how independent variables affect these. It can therefore be regarded as a semi-mechanistic model, not simulating specific metabolic

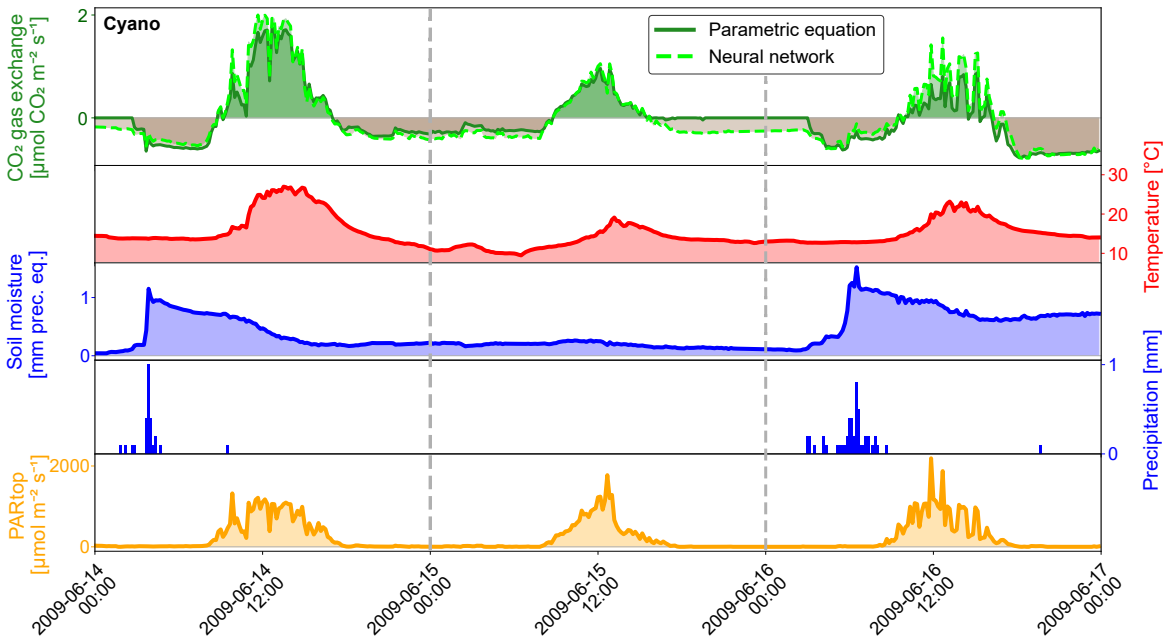

Figure 5: *Application of the parametric equation and neural network models for cyano on micrometeorological data from Soebatsfontein, Succulent Karoo, South Africa over 3 days from June 14 to June 16, 2009. Red, yellow and blue lines show temperature, light intensity and soil moisture profiles, blue bars represent precipitation, the green lines display CO_2 gas exchange of cyano biocrusts predicted by the respective models. Brown areas in the top panel indicate periods where respiration surpasses photosynthetic activity, while green areas show times of net photosynthetic productivity.*

processes, but mathematically describing the effect and interaction of environmental parameters in separate, meaningful terms. For the NN models, an interpretation of underlying principles is near impossible, as their operating principle relies on a very large number of statistical and unphysical weights, i.e. trained parameters. Only sensitivity studies can be used to investigate underlying principles of a trained NN model. However, a beneficial aspect of this complexity is that NN models can be applied in an 'uninformed' manner, even when mathematical principles are complex or unknown. Another notable difference between the two presented models relates to the extrapolation beyond experimental ranges of environmental parameters. In the parametric equation model, the mathematical terms define and constrain model behavior beyond the independent parameter range. This is not the case for the NN, where behavior for extreme values may be unexpected and unphysical. However, NN models can in principle be confined, e.g., through addition of synthetic data points.

Using these trained models, CO_2 gas exchange predictions can be integrated over large time periods to evaluate long-term NPP at specific locations. However, the extrapolation of laboratory measurements to natural habitats and conditions can be impeded by multiple factors. Environmental parameters under laboratory conditions may not fully represent the complex interactions at play in nature. While soil moisture was assessed gravimetrically in the laboratory, a newly-developed biocrust wetness

probe was applied in the field, which potentially induces deviations in the measured soil moisture (Weber et al., 2016). These different methods may cause some measurement deviations. As another example, acclimation likely plays a significant role in the natural physiological response of biological organisms to environmental conditions. Consequently, samples collected at a specific time point likely do not reflect variations in response occurring over the course of the year. Laboratory experiments are often designed to explore the steady-state effect of environmental conditions systematically, altering only one parameter at a time and measuring the response of sample organisms. Such an approach is necessary to understand the effect of each environmental parameter individually, as well as their interaction in an equilibrium state. In nature, however, environmental conditions may change in varying rates and combinations. To enhance future predictions of physiological responses such as CO_2 gas exchange, experiments could be structured to explore not only the combinations of environmental parameters but also the sequence and rate in which changes occur, as both the order and combination of environmental variations over time may significantly impact the outcomes.

Abbreviations

biocrust - biological soil crust
C - carbon
CO_2 - carbon dioxide
L - light intensity
M - precipitation equivalent
MSE - mean square error
NN - Neural Network
NP - net photosynthesis
NPP - net primary productivity
PAR - photosynthetically active radiation
PPFD - photosynthetic photon flux density
R - respiration
T - temperature

References

Adams, B., White, A., and Lenton, T.: An analysis of some diverse approaches to modelling terrestrial net primary productivity, Ecol. Model., 177, 353–391, 2004.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, ACM, Anchorage AK USA, ISBN 978-1-4503-6201-6, 2019.

Arangio, A. M., Slade, J. H., Berkemeier, T., Pöschl, U., Knopf, D. A., and Shiraiwa, M.: Multiphase Chemical Kinetics of OH Radical Uptake by Molecular Organic Markers of Biomass Burning Aerosols: Humidity and Temperature Dependence, Surface Reaction, and Bulk Diffusion, J. Phys. Chem. A, 119, 4533–4544, 2015.

Armstrong, R. and Bradwell, T.: Growth of crustose lichens: a review, Geogr. Ann. A: Phys. Geogr., 92, 3–17, 2010.

Bader, M. Y., Zotz, G., and Lange, O. L.: How to minimize the sampling effort for obtaining reliable estimates of diel and annual CO_2 budgets in lichens, Lichenologist, 42, 97–111, 2010.

Barger, N. N., Weber, B., Garcia-Pichel, F., Zaady, E., and Belnap, J.: Patterns and Controls on Nitrogen Cycling of Biological Soil Crusts, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 257–285, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Belnap, J.: The potential roles of biological soil crusts in dryland hydrologic cycles, Hydrol. Process., 20, 3159–3178, 2006.

Belnap, J. and Büdel, B.: Biological Soil Crusts as Soil Stabilizers, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 305–320, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Belnap, J., Weber, B., and Büdel, B.: Biological Soil Crusts as an Organizing Principle in Drylands, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 3–13, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Bennett, D. I. G., Amarnath, K., Park, S., Steen, C. J., Morris, J. M., and Fleming, G. R.: Models and mechanisms of the rapidly reversible regulation of photosynthetic light harvesting, Open Biol., 9, 190 043, 2019.

Berkemeier, T., Steimer, S. S., Krieger, U. K., Peter, T., Pöschl, U., Ammann, M., and Shiraiwa, M.: Ozone uptake on glassy, semi-solid and liquid organic matter and the role of reactive oxygen intermediates in atmospheric aerosol chemistry, Phys. Chem. Chem. Phys., 18, 12 662–12 674, 2016.

Berkemeier, T., Ammann, M., Krieger, U. K., Peter, T., Spichtinger, P., Pöschl, U., Shiraiwa, M., and Huisman, A. J.: Technical note: Monte Carlo genetic algorithm (MCGA) for model analysis of multiphase chemical kinetics to determine transport and reaction rate coefficients using multiple experimental data sets, Atmos. Chem. Phys., 17, 8021–8029, 2017.

Berkemeier, T., Takeuchi, M., Eris, G., and Ng, N. L.: Kinetic modeling of formation and evaporation of secondary organic aerosol from NO3; oxidation of pure and mixed monoterpenes, Atmos. Chem. Phys., 20, 15 513–15 535, 2020.

Bisbee, K. E., Gower, S. T., Norman, J. M., and Nordheim, E. V.: Environmental controls on ground cover species composition and productivity in a boreal black spruce forest, Oecologia, 129, 261–270, 2001.

Brankatschk, R., Fischer, T., Veste, M., and Zeyer, J.: Succession of N cycling processes in biological soil crusts on a Central European inland dune, FEMS Microbiol. Ecol., 83, 149–160, 2013.

Brostoff, W. N., Sharifi, M. R., and Rundel, P. W.: Photosynthesis of cryptobiotic crusts in a seasonally inundated system of pans and dunes at Edwards Air Force Base, western Mojave Desert, California: Laboratory studies, Flora: Morphol. Distrib. Funct. Ecol. Plants, 197, 143–151, 2002.

Büdel, B., Vivas, M., and Lange, O. L.: Lichen species dominance and the resulting photosynthetic behavior of Sonoran Desert soil crust types (Baja California, Mexico), Ecol. Process., 2, 6, 2013.

Chamizo, S., Belnap, J., Eldridge, D. J., Cantón, Y., and Malam Issa, O.: The Role of Biocrusts in Arid Land Hydrology, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 321–346, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Coxson, D. S., McIntyre, D. D., and Vogel, H. J.: Pulse Release of Sugars and Polyols from Canopy Bryophytes in Tropical Montane Rain Forest (Guadeloupe, French West Indies), Biotropica, 24, 121, 1992.

Darby, B. J. and Neher, D. A.: Microfauna Within Biological Soil Crusts, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 139–157, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Dumack, K., Koller, R., Weber, B., and Bonkowsk, M.: Estimated abundance and diversity of heterotrophic protists in South African biocrusts, S. Afr. J. Sci., 112, 5, 2016.

Farquhar, G. D., von Caemmerer, S., and Berry, J. A.: A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species, Planta, 149, 78–90, 1980.

Gypser, S., Herppich, W. B., Fischer, T., Lange, P., and Veste, M.: Photosynthetic characteristics and their spatial variance on biological soil crusts covering initial soils of post-mining sites in Lower Lusatia, NE Germany, Flora: Morphol. Distrib. Funct. Ecol. Plants, 220, 103–116, 2016.

Housman, D., Powers, H., Collins, A., and Belnap, J.: Carbon and nitrogen fixation differ between successional stages of biological soil crusts in the Colorado Plateau and Chihuahuan Desert, J. Arid Environ., 66, 620–634, 2006.

James, G., Witten, D., Hastie, T., Tibshirani, R., et al.: An introduction to statistical learning, vol. 112, Springer, 2013.

Jeffries, D. L., Link, S. O., and Klopatek, J. M.: CO2 fluxes of cryptogamic crusts: I. Response to resaturation, New Phytologist, 125, 163–173, 1993a.

Jeffries, D. L., Link, S. O., and Klopatek, J. M.: CO2 fluxes of cryptogamic crusts: II. Response to dehydration, New Phytologist, 125, 391–396, 1993b.

Kiefer, J.: Sequential minimax search for a maximum, Proc. Amer. Math. Soc., 4, 502–506, 1953.

Kim, M., Lopez-Canfin, C., Lázaro, R., Sánchez-Cañete, E. P., and Weber, B.: Unravelling the main mechanism responsible for nocturnal CO2 uptake by dryland soils, Sci. Total Environ., 926, 171 751, publisher: Elsevier BV, 2024.

Lakey, P. S. J., Berkemeier, T., Tong, H., Arangio, A. M., Lucas, K., Pöschl, U., and Shiraiwa, M.: Chemical exposure-response relationship between air pollutants and reactive oxygen species in the human respiratory tract, Sci. Rep., 6, 32 916, 2016.

Lange, O. L.: Photosynthetic productivity of the epilithic lichen Lecanora muralis: Long-term field monitoring of CO2 exchange and its physiological interpretation. I. Dependence of photosynthesis on water content, light, temperature, and CO2 concentration from laboratory measurements, Flora: Morphol. Distrib. Funct. Ecol. Plants., 197, 233–249, 2002.

Lange, O. L.: Photosynthetic productivity of the epilithic lichen Lecanora muralis: long-term field monitoring of CO2 exchange and its physiological interpretation: II. Diel and seasonal patterns of net photosynthesis and respiration, Flora: Morphol. Distrib. Funct. Ecol. Plants., 198, 55–70, 2003.

Lange, O. L., Belnap, J., Reichenberger, H., and Meyer, A.: Photosynthesis of green algal soil crust lichens from arid lands in southern Utah, USA: role of water content on light and temperature responses of CO2 exchange, Flora: Morphol. Distrib. Funct. Ecol. Plants., 192, 1–15, 1997.

Lange, O. L., Belnap, J., and Reichenberger, H.: Photosynthesis of the cyanobacterial soil-crust lichen *Collema tenax* from arid lands in southern Utah, USA: role of water content on light and temperature responses of CO_2 exchange, Funct. Ecol., 12, 195–202, 1998.

Lange, O. L., Green, T. A., and Reichenberger, H.: The Response of Lichen Photosynthesis to External CO2 Concentration and its Interaction with Thallus Water-status, J. Plant Physiol., 154, 157–166, 1999.

Lange, O. L., Allan Green, T., Melzer, B., Meyer, A., and Zellner, H.: Water relations and CO2 exchange of the terrestrial lichen Teloschistes capensis in the Namib fog desert: Measurements during two seasons in the field and under controlled conditions, Flora: Morphol. Distrib. Funct. Ecol. Plants., 201, 268–280, 2006.

Li, B., Gao, J., Wang, X., Ma, L., Cui, Q., and Vest, M.: Effects of biological soil crusts on water infiltration and evaporation Yanchi Ningxia, Maowusu Desert, China, Int. J. Sediment Res., 31, 311–323, 2016.

Lieth, H.: Primary production: Terrestrial ecosystems, Hum. Ecol., 1, 303–332, 1973.

Ma, J., Wang, Z.-Y., Stevenson, B. A., Zheng, X.-J., and Li, Y.: An inorganic CO2 diffusion and dissolution process explains negative CO2 fluxes in saline/alkaline soils, Sci. Rep., 3, 2025, 2013.

Ma, Y., Weber, B., Kratz, A., Raggio, J., Colesie, C., Veste, M., Bader, M. Y., and Porada, P.: Exploring environmental and physiological drivers of the annual carbon budget of biocrusts from various climatic zones with a mechanistic data-driven model, Biogeosciences, 20, 2553–2572, 2023.

Maier, S., Tamm, A., Wu, D., Caesar, J., Grube, M., and Weber, B.: Photoautotrophic organisms control microbial abundance, diversity, and physiology in different types of biological soil crusts, The ISME Journal, 12, 1032–1046, 2018.

Matuszyńska, A., Saadat, N. P., and Ebenhöh, O.: Balancing energy supply during photosynthesis – a theoretical perspective, Physiol. Plant., 166, 392–402, 2019.

Morgan, S. L. and Deming, S. N.: Simplex optimization of analytical chemical methods, Analytical chemistry, 46, 1170–1181, 1974.

Nikolić, N., Zotz, G., and Bader, M. Y.: Modelling the carbon balance in bryophytes and lichens: Presentation of PoiCarb 1.0, a new model for explaining distribution

patterns and predicting climate-change effects, American J. of Botany, 111, e16 266, 2024.

O. Rauff, K. and Bello, R.: A Review of Crop Growth Simulation Models as Tools for Agricultural Meteorology, AS, 06, 1098–1105, 2015.

Pointing, S. B. and Belnap, J.: Microbial colonization and controls in dryland systems, Nat. Rev. Microbiol., 10, 551–562, 2012.

Rey, A.: Mind the gap: non-biological processes contributing to soil CO2 efflux, Glob. Change Biol., 21, 1752–1761, 2015.

Rodríguez-Caballero, E., Castro, A. J., Chamizo, S., Quintas-Soriano, C., Garcia-Llorente, M., Cantón, Y., and Weber, B.: Ecosystem services provided by biocrusts: From ecosystem functions to social values, J. Arid Environ., 159, 45–53, publisher: Elsevier BV, 2018.

Rubio, F. C., Camacho, F. G., Sevilla, J. M. F., Chisti, Y., and Grima, E. M.: A mechanistic model of photosynthesis in microalgae, Biotechnol. Bioeng., 81, 459–473, 2003.

Sancho, L. G. and Pintado, A.: Evidence of high annual growth rate for lichens in the maritime Antarctic, Polar Biology, 27, 312–319, 2004.

Serôdio, J. and Lavaud, J.: A model for describing the light response of the nonphotochemical quenching of chlorophyll fluorescence, Photosynth. Res., 108, 61–76, 2011.

Sukhova, E. M., Vodeneev, V. A., and Sukhov, V. S.: Mathematical Modeling of Photosynthesis and Analysis of Plant Productivity, Biochem. Moscow Suppl. Ser. A, 15, 52–72, 2021.

Tamm, A., Caesar, J., Kunz, N., Colesie, C., Reichenberger, H., and Weber, B.: Ecophysiological properties of three biological soil crust types and their photoautotrophs from the Succulent Karoo, South Africa, Plant Soil, 429, 127–146, 2018.

Tikkanen, O.-P., Hämäläinen, V., Rovelli, G., Lipponen, A., Shiraiwa, M., Reid, J. P., Lehtinen, K. E. J., and Yli-Juuti, T.: Optimization of process models for determining volatility distribution and viscosity of organic aerosols from isothermal particle evaporation data, Atmos. Chem. Phys., 19, 9333–9350, 2019.

Von Caemmerer, S.: Steady-state models of photosynthesis, Plant Cell Environ., 36, 1617–1630, 2013.

Weber, B., Graf, T., and Bass, M.: Ecophysiological analysis of moss-dominated biological soil crusts and their separate components from the Succulent Karoo, South Africa, Planta, 236, 129–139, 2012.

Weber, B., Berkemeier, T., Ruckteschler, N., Caesar, J., Heintz, H., Ritter, H., and Braß, H.: Development and calibration of a novel sensor to quantify the water content of surface soils and biological soil crusts, Methods Ecol. Evol., 7, 14–22, 2016.

Weber, B., Tamm, A., Maier, S., and Rodríguez-Caballero, E.: Biological soil crusts of the Succulent Karoo: a review, Afr. J. Range Forage Sci., 35, 335–350, publisher: National Inquiry Services Center (NISC), 2018.

Weber, B., Belnap, J., Büdel, B., Antoninka, A. J., Barger, N. N., Chaudhary, V. B., Darrouzet-Nardi, A., Eldridge, D. J., Faist, A. M., Ferrenberg, S., Havrilla, C. A., Huber-Sannwald, E., Malam Issa, O., Maestre, F. T., Reed, S. C., Rodriguez-Caballero, E., Tucker, C., Young, K. E., Zhang, Y., Zhao, Y., Zhou, X., and Bowker, M. A.: What is a biocrust? A refined, contemporary definition for a broadening research community, Biological Reviews, 97, 1768–1785, 2022.

Wilske, B., Burgheimer, J., Karnieli, A., Zaady, E., Andreae, M. O., Yakir, D., and Kesselmeier, J.: The CO2 exchange of biological soil crusts in a semiarid grass-shrubland at the northern transition zone of the Negev desert, Israel, Biogeosciences, 5, 1411–1423, 2008.

Zaady, E., Kuhn, U., Wilske, B., Sandoval-Soto, L., and Kesselmeier, J.: Patterns of CO2 exchange in biological soil crusts of successional age, Soil Biol. Biochem., 32, 959–966, 2000.

Zaks, D. P. M., Ramankutty, N., Barford, C. C., and Foley, J. A.: From Miami to Madison: Investigating the relationship between climate and terrestrial net primary production, Glob. Biogeochem. Cycles, 21, 2006GB002 705, 2007.

Zhang, Y., Aradottir, A. L., Serpe, M., and Boeken, B.: Interactions of Biological Soil Crusts with Vascular Plants, in: Biological Soil Crusts: An Organizing Principle in Drylands, edited by Weber, B., Büdel, B., and Belnap, J., vol. 226, pp. 385–406, Springer International Publishing, Cham, ISBN 978-3-319-30212-6 978-3-319-30214-0, series Title: Ecological Studies, 2016.

Zhu, X., Wang, Y., Ort, D. R., and Long, S. P.: *e* -photosynthesis: a comprehensive dynamic mechanistic model of C3 photosynthesis: from light capture to sucrose synthesis, Plant Cell Environ., 36, 1711–1727, 2013.

3. Conclusions and Outlook

This PhD project advanced the utilization of machine learning (ML) and other data-centric methods in the fields of multiphase and atmospheric chemistry. Using advanced architectures of artificial neural networks, quantitative structure-activity relationship (QSAR) models were developed and published that outperformed previous group contribution-based prediction models commonly used by atmospheric scientists (Pankow, 1987; Compernolle et al., 2011). A convolutional neural network (CNN) for the prediction of reduction potentials of quinones and a group contribution-assisted graph convolutional neural network (GC^2NN) for the prediction of saturation vapor pressures of atmospheric compounds both showed excellent agreement with experimental data, underlying their potential application in broader atmospheric modelling approaches (Krüger et al., 2022, 2025).

Aside from the direct utilization of trained QSAR models, their newly developed model architectures could be applied to other data sets and physicochemical properties in a similar fashion, e.g., for the prediction of Henry's law coefficients (Sander, 2015). Further advancements in QSAR model architectures may encompass the inclusion of heuristic rules or physics-informed modules (Bilde et al., 2015), transfer learning (Lansford et al., 2023) or pre-trained foundation models that can be fine-tuned using small data sets (Burns et al., 2025). Depending on data availability, more generalized models that predict multiple physicochemical properties simultaneously may be advantageous over highly specified ones. In a similar fashion, siamese neural networks can be used for predictions that require multiple codependent inputs (Jeon et al., 2019), e.g., reaction rate coefficients between multiple molecules.

Multiphase chemical kinetics models were advanced with methods in the fields of inverse modelling, uncertainty quantification and experiment design through the development of fast and accurate surrogate models and the Numerical Compass (NC) method for uncertainty quantification and experiment design (Berkemeier et al., 2023; Krüger et al., 2024). It was shown that fit ensembles of plausible model solutions consistent with experimental data could be generated much faster with the support of neural network surrogate models. Based on these fit ensembles, the NC framework was successfully applied to assess the parametric uncertainty of multiphase chemical kinetics models. Beyond multiphase chemical kinetics models, the method can be applied to other process models. Using this method, laboratory experiments can be designed not only to constrain overall model uncertainty but also to target specific parameters or conditions (e.g., those prevailing in the atmosphere). Both methods - the surrogate modelling and NC - were thoroughly tested in simulations, independently and in conjunction (Krüger et al., 2024). The versatility of the NC was furthermore

demonstrated through its application to a QSAR model to identify potential candidates for future model training. While targeted selection of QSAR training candidates slightly outperformed random selection in preliminary tests, applying the NC to optimize QSAR model training requires further research, e.g., via the development of an iterative cycle that interleaves model training with training data selection.

In addition to inverse modelling and uncertainty quantification, fast and accurate surrogate models could be used to bridge the gap between multiphase chemical kinetics models that simulate aerosol processes at small scales and large-scale atmospheric models. Such chemical transport models often require large numbers of module evaluations because they may incorporate many grid cells and long time scales, making the direct application of complex differential equation models as modules impractical or infeasible. Accelerating such models by multiple orders of magnitude using surrogate modelling may enable the replacement of less accurate parameterizations currently in use. Another multiscale modelling approach demonstrated in this work applied artificial neural networks (NN) trained on laboratory experiments to quantify gas-exchange rates of biological soil crusts (biocrusts). These NN could be used to estimate the net primary production of biocrusts in dryland ecosystems from micrometeorological data, thereby improving understanding of biosphere–atmosphere carbon exchange.

4. Bibliography

Al-Ghussain, L.: Global warming: review on driving forces and mitigation, Env. Prog. and Sustain. Energy, 38, 13–21, 2019.

Armeli, G., Peters, J.-H., and Koop, T.: Machine-Learning-Based Prediction of the Glass Transition Temperature of Organic Compounds Using Experimental Data, ACS Omega, 8, 12 298–12 309, 2023.

Benedetti, A., Reid, J. S., Knippertz, P., Marsham, J. H., Di Giuseppe, F., Rémy, S., Basart, S., Boucher, O., Brooks, I. M., Menut, L., Mona, L., Laj, P., Pappalardo, G., Wiedensohler, A., Baklanov, A., Brooks, M., Colarco, P. R., Cuevas, E., Da Silva, A., Escribano, J., Flemming, J., Huneeus, N., Jorba, O., Kazadzis, S., Kinne, S., Popp, T., Quinn, P. K., Sekiyama, T. T., Tanaka, T., and Terradellas, E.: Status and future of numerical atmospheric aerosol prediction with a focus on data requirements, Atmos. Chem. Phys., 18, 10 615–10 643, 2018.

Berkemeier, T., Shiraiwa, M., Pöschl, U., and Koop, T.: Competition between water uptake and ice nucleation by glassy organic aerosol particles, Atmos. Chem. Phys., 14, 12 513–12 531, 2014.

Berkemeier, T., Ammann, M., Krieger, U. K., Peter, T., Spichtinger, P., Pöschl, U., Shiraiwa, M., and Huisman, A. J.: Technical note: Monte Carlo genetic algorithm (MCGA) for model analysis of multiphase chemical kinetics to determine transport and reaction rate coefficients using multiple experimental data sets, Atmos. Chem. Phys., 17, 8021–8029, 2017.

Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., and Krieger, U. K.: Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks, Geosci. Model Dev., 16, 2037–2054, 2023.

Besel, V., Todorović, M., Kurtén, T., Rinke, P., and Vehkamäki, H.: Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules, Sci. Data, 10, 450, 2023.

Bilde, M. and Pandis, S. N.: Evaporation Rates and Vapor Pressures of Individual Aerosol Species Formed in the Atmospheric Oxidation of α- and β-Pinene, Environ. Sci. Technol., 35, 3344–3349, 2001.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., Topping, D., Ziemann, P., Barley, M., Clegg, S., Dennis-Smither, B., Hallquist, M., Hallquist, A. M., Khlystov, A., Kulmala, M., Mogensen, D., Percival, C. J., Pope, F., Reid, J. P., Ribeiro da Silva, M. A. V., Rosenoern, T., Salo, K., Soonsin, V. P., Yli-Juuti, T., Prisle, N. L., Pagels, J., Rarey, J., Zardini, A. A., and Riipinen, I.: Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures, Chem. Rev., 115, 4115–4156, 2015.

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., Apte, J. S., Brauer, M., Cohen, A., Weichenthal, S., Coggins, J., Di, Q., Brunekreef, B., Frostad, J., Lim, S. S., Kan, H., Walker, K. D., Thurston, G. D., Hayes, R. B., Lim, C. C., Turner, M. C., Jerrett, M., Krewski, D., Gapstur, S. M., Diver, W. R., Ostro, B., Goldberg, D., Crouse, D. L., Martin, R. V., Peters, P., Pinault, L., Tjepkema, M., Van Donkelaar, A., Villeneuve, P. J., Miller, A. B., Yin, P., Zhou, M., Wang, L., Janssen, N. A. H., Marra, M., Atkinson, R. W., Tsang, H., Quoc Thach, T., Cannon, J. B., Allen, R. T., Hart, J. E., Laden, F., Cesaroni, G., Forastiere, F., Weinmayr, G., Jaensch, A., Nagel, G., Concin, H., and Spadaro, J. V.: Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter, Proc. Natl. Acad. Sci. U.S.A., 115, 9592–9597, 2018.

Burnett, R. T., Pope, C. A., Ezzati, M., Olives, C., Lim, S. S., Mehta, S., Shin, H. H., Singh, G., Hubbell, B., Brauer, M., Anderson, H. R., Smith, K. R., Balmes, J. R., Bruce, N. G., Kan, H., Laden, F., Prüss-Ustün, A., Turner, M. C., Gapstur, S. M., Diver, W. R., and Cohen, A.: An Integrated Risk Function for Estimating the Global Burden of Disease Attributable to Ambient Fine Particulate Matter Exposure, Environ. Health Perspect., 122, 397–403, 2014.

Burns, J., Zalte, A., and Green, W.: Descriptor-based Foundation Models for Molecular Property Prediction, version Number: 1, 2025.

Charrier, J. G., McFall, A. S., Richards-Henderson, N. K., and Anastasio, C.: Hydrogen Peroxide Formation in a Surrogate Lung Fluid by Transition Metals and Quinones Present in Particulate Matter, Environ. Sci. Technol., 48, 7010–7017, 2014.

Christensen, H. and Zanna, L.: Parametrization in Weather and Climate Models, in: Oxford Research Encyclopedia of Climate Science, Oxford University Press, ISBN 978-0-19-022862-0, 2022.

Christopoulos, C. D., Garimella, S., Zawadowicz, M. A., Möhler, O., and Cziczo, D. J.: A machine learning approach to aerosol classification for single-particle mass spectrometry, Atmos. Meas. Tech., 11, 5687–5699, 2018.

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C. A., Shin,

H., Straif, K., Shaddick, G., Thomas, M., Van Dingenen, R., Van Donkelaar, A., Vos, T., Murray, C. J. L., and Forouzanfar, M. H.: Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015, The Lancet, 389, 1907–1918, 2017.

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmos. Chem. Phys., 11, 9431–9450, 2011.

DeCarlo, P. F., Kimmel, J. R., Trimborn, A., Northway, M. J., Jayne, J. T., Aiken, A. C., Gonin, M., Fuhrer, K., Horvath, T., Docherty, K. S., Worsnop, D. R., and Jimenez, J. L.: Field-Deployable, High-Resolution, Time-of-Flight Aerosol Mass Spectrometer, Anal. Chem., 78, 8281–8289, 2006.

Donahue, N. M., Robinson, A. L., and Pandis, S. N.: Atmospheric organic particulate matter: From smoke to secondary organic aerosol, Atmos. Environ., 43, 94–106, 2009.

Dovrou, E., Lelieveld, S., Mishra, A., Pöschl, U., and Berkemeier, T.: Influence of ambient and endogenous H_2 O_2 on reactive oxygen species concentrations and OH radical production in the respiratory tract, Environ. Sci.: Atmos., 3, 1066–1074, 2023.

Galeazzo, T. and Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds *via* machine learning and molecular embeddings, Environ. Sci.: Atmos., 2, 362–374, 2022.

George, C., Ammann, M., D'Anna, B., Donaldson, D. J., and Nizkorodov, S. A.: Heterogeneous Photochemistry in the Atmosphere, Chem. Rev., 115, 4218–4258, 2015.

Gianquintieri, L., Oxoli, D., Caiani, E. G., and Brovelli, M. A.: State-of-art in modelling particulate matter (PM) concentration: a scoping review of aims and methods, Environ. Dev. Sustain., 2024.

Giles, D. M., Sinyuk, A., Sorokin, M. G., Schafer, J. S., Smirnov, A., Slutsker, I., Eck, T. F., Holben, B. N., Lewis, J. R., Campbell, J. R., Welton, E. J., Korkin, S. V., and Lyapustin, A. I.: Advancements in the Aerosol Robotic Network (AERONET) Version 3 database – automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements, Atmos. Meas. Tech., 12, 169–209, 2019.

Haywood, J. and Boucher, O.: Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review, Rev. Geophys., 38, 513–543, 2000.

Hoose, C., Kristjánsson, J. E., Iversen, T., Kirkevåg, A., Seland, O., and Gettelman, A.: Constraining cloud droplet number concentration in GCMs suppresses the aerosol indirect effect, Geophys. Res. Lett., 36, 2009GL038 568, 2009.

Hopke, P. K., Dai, Q., Li, L., and Feng, Y.: Global review of recent source apportionments for airborne particulate matter, Sci. Total Environ., 740, 140 091, 2020.

Housman, D., Powers, H., Collins, A., and Belnap, J.: Carbon and nitrogen fixation differ between successional stages of biological soil crusts in the Colorado Plateau and Chihuahuan Desert, J. Arid Environ., 66, 620–634, 2006.

Intergovernmental Panel On Climate Change (IPCC): Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1 edn., ISBN 978-1-009-15789-6, 2023.

Jeon, M., Park, D., Lee, J., Jeon, H., Ko, M., Kim, S., Choi, Y., Tan, A.-C., and Kang, J.: ReSimNet: drug response similarity prediction using Siamese neural networks, Bioinformatics, 35, 5249–5256, 2019.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., E., Dunlea, J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525–1529, 2009.

Kanakidou, M., Seinfeld, J. H., Pandis, S. N., Barnes, I., Dentener, F. J., Facchini, M. C., Van Dingenen, R., Ervens, B., Nenes, A., Nielsen, C. J., Swietlicki, E., Putaud, J. P., Balkanski, Y., Fuzzi, S., Horth, J., Moortgat, G. K., Winterhalter, R., Myhre, C. E. L., Tsigaridis, K., Vignati, E., Stephanou, E. G., and Wilson, J.: Organic aerosol and global climate modelling: a review, Atmos. Chem. Phys., 5, 1053–1123, 2005.

Kok, J. F., Storelvmo, T., Karydis, V. A., Adebiyi, A. A., Mahowald, N. M., Evan, A. T., He, C., and Leung, D. M.: Mineral dust aerosol impacts on global climate and climate change, Nat. Rev. Earth Environ., 4, 71–86, 2023.

Kolb, C. E., Cox, R. A., Abbatt, J. P. D., Ammann, M., Davis, E. J., Donaldson, D. J., Garrett, B. C., George, C., Griffiths, P. T., Hanson, D. R., Kulmala, M., McFiggans, G., Pöschl, U., Riipinen, I., Rossi, M. J., Rudich, Y., Wagner, P. E., Winkler, P. M., Worsnop, D. R., and O' Dowd, C. D.: An overview of current issues in the uptake of atmospheric trace gases by aerosols and clouds, Atmos. Chem. Phys., 10, 10 561–10 605, 2010.

Kröse, B. and van der Smagt, P.: Introduction to Neural Networks, 1996.

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B. A. M., Lammel, G., Schmidt, B., Pöschl, U., and Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, Nat. Sci., 2, e20220 016, 2022.

Krüger, M., Mishra, A., Spichtinger, P., Pöschl, U., and Berkemeier, T.: A numerical compass for experiment design in chemical kinetics and molecular property estimation, J. Cheminform., 16, 34, 2024.

Krüger, M., Galeazzo, T., Eremets, I., Schmidt, B., Pöschl, U., Shiraiwa, M., and Berkemeier, T.: Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^2 NN), Geosci. Model Dev., 18, 7357–7371, 2025.

Kumar, V., Sahu, M., and Biswas, P.: Source Apportionment of Particulate Matter by Application of Machine Learning Clustering Algorithms, Aerosol Air Qual. Res., 22, 210 240, 2022.

Lam, Y. F., Fu, J. S., Wu, S., and Mickley, L. J.: Impacts of future climate change and effects of biogenic emissions on surface ozone and particulate matter concentrations in the United States, Atmos. Chem. Phys., 11, 4789–4806, 2011.

Lange, O. L.: Photosynthetic productivity of the epilithic lichen Lecanora muralis: Long-term field monitoring of CO2 exchange and its physiological interpretation. I. Dependence of photosynthesis on water content, light, temperature, and CO2 concentration from laboratory measurements, Flora: Morphol. Distrib. Funct. Ecol. Plants., 197, 233–249, 2002.

Lansford, J. L., Jensen, K. F., and Barnes, B. C.: Physics-informed Transfer Learning for Out-of-sample Vapor Pressure Predictions, Propellants Explo. Pyrotec., 48, e202200 265, 2023.

Lelieveld, J., Pozzer, A., Pöschl, U., Fnais, M., Haines, A., and Münzel, T.: Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective, Cardiovasc. Res., 116, 1910–1917, 2020.

Lelieveld, S., Wilson, J., Dovrou, E., Mishra, A., Lakey, P. S. J., Shiraiwa, M., Pöschl, U., and Berkemeier, T.: Hydroxyl Radical Production by Air Pollutants in Epithelial Lining Fluid Governed by Interconversion and Scavenging of Reactive Oxygen Species, Environ. Sci. Technol., 55, 14 069–14 079, 2021.

Li, M., Su, H., Li, G., Ma, N., Pöschl, U., and Cheng, Y.: Relative importance of gas uptake on aerosol and ground surfaces characterized by equivalent uptake coefficients, Atmos. Chem. Phys., 19, 10 981–11 011, 2019.

Li, N., Hao, M., Phalen, R. F., Hinds, W. C., and Nel, A. E.: Particulate air pollutants and asthma, Clin. Immunol., 109, 250–265, 2003.

Li, Y., Pöschl, U., and Shiraiwa, M.: Molecular corridors and parameterizations of volatility in the chemical evolution of organic aerosols, Atmos. Chem. Phys., 16, 3327–3344, 2016.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., and Rinke, P.: Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning, Atmos. Chem. Phys., 21, 13 227–13 246, 2021.

Masmoudi, S., Elghazel, H., Taieb, D., Yazar, O., and Kallel, A.: A machine-learning framework for predicting multiple air pollutants' concentrations via multi-target regression and feature selection, Sci. Total Environ., 715, 136 991, 2020.

Mellouki, A., Wallington, T. J., and Chen, J.: Atmospheric Chemistry of Oxygenated Volatile Organic Compounds: Impacts on Air Quality and Climate, Chem. Rev., 115, 3984–4014, 2015.

Mishra, A., Lelieveld, S., Pöschl, U., and Berkemeier, T.: Multiphase Kinetic Modeling of Air Pollutant Effects on Protein Modification and Nitrotyrosine Formation in Epithelial Lining Fluid, Environ. Sci. Technol., 57, 12 642–12 653, 2023.

Monks, T. J., Hanzlik, R. P., Cohen, G. M., Ross, D., and Graham, D. G.: Quinone chemistry and toxicity, Toxicol. Appl. Pharmacol., 112, 2–16, 1992.

Mukherjee, A. and Agrawal, M.: World air particulate matter: sources, distribution and health effects, Environ. Chem. Lett., 15, 283–309, 2017.

Muthukumar, P., Cocom, E., Nagrecha, K., Comer, D., Burga, I., Taub, J., Calvert, C. F., Holm, J., and Pourhomayoun, M.: Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data, Air Qual. Atmos. Health, 15, 1221–1234, 2022.

Nair, A. A., Yu, F., Campuzano-Jost, P., DeMott, P. J., Levin, E. J. T., Jimenez, J. L., Peischl, J., Pollack, I. B., Fredrickson, C. D., Beyersdorf, A. J., Nault, B. A., Park, M., Yum, S. S., Palm, B. B., Xu, L., Bourgeois, I., Anderson, B. E., Nenes, A., Ziemba, L. D., Moore, R. H., Lee, T., Park, T., Thompson, C. R., Flocke, F., Huey, L. G., Kim, M. J., and Peng, Q.: Machine Learning Uncovers Aerosol Size Information From Chemistry and Meteorology to Quantify Potential Cloud-Forming Particles, Geophys. Res. Lett., 48, e2021GL094 133, 2021.

Pankow, J. F.: Review and comparative analysis of the theories on partitioning between the gas and aerosol particulate phases in the atmosphere, Atmos. Environ., 21, 2275–2283, 1987.

Pointing, S. B. and Belnap, J.: Microbial colonization and controls in dryland systems, Nat. Rev. Microbiol., 10, 551–562, 2012.

Pozzer, A., Tsimpidi, A. P., Karydis, V. A., De Meij, A., and Lelieveld, J.: Impact of agricultural emission reductions on fine-particulate matter and public health, Atmos. Chem. Phys., 17, 12 813–12 826, 2017.

Pöschl, U., Rudich, Y., and Ammann, M.: Kinetic model framework for aerosol and cloud surface chemistry and gas-particle interactions – Part 1: General equations, parameters, and terminology, Atmos. Chem. Phys., 7, 5989–6023, 2007.

Roginsky, V. A., Barsukova, T. K., and Stegmann, H. B.: Kinetics of redox interaction between substituted quinones and ascorbate under aerobic conditions, Chem.-Biol. Interact., 121, 177–197, 1999.

Roldin, P., Eriksson, A. C., Nordin, E. Z., Hermansson, E., Mogensen, D., Rusanen, A., Boy, M., Swietlicki, E., Svenningsson, B., Zelenyuk, A., and Pagels, J.: Modelling non-equilibrium secondary organic aerosol formation and evaporation with the aerosol dynamics, gas- and particle-phase chemistry kinetic multilayer model ADCHAM, Atmos. Chem. Phys., 14, 7953–7993, 2014.

Rudich, Y., Donahue, N. M., and Mentel, T. F.: Aging of Organic Aerosol: Bridging the Gap Between Laboratory and Field Studies, Annu. Rev. Phys. Chem., 58, 321–352, 2007.

Ruske, S., Topping, D. O., Foot, V. E., Morse, A. P., and Gallagher, M. W.: Machine learning for improved data analysis of biological aerosol using the WIBS, Atmos. Meas. Tech., 11, 6203–6230, 2018.

Sander, R.: Compilation of Henry's law constants (version 4.0) for water as solvent, Atmos. Chem. Phys., 15, 4399–4981, 2015.

Sandström, H., Rissanen, M., Rousu, J., and Rinke, P.: Towards data-driven mass spectrometry in atmospheric science, publisher: arXiv Version Number: 2, 2023.

Semeniuk, K. and Dastoor, A.: Current State of Atmospheric Aerosol Thermodynamics and Mass Transfer Modeling: A Review, Atmosphere, 11, 156, 2020.

Shiraiwa, M., Pfrang, C., and Pöschl, U.: Kinetic multi-layer model of aerosol surface and bulk chemistry (KM-SUB): the influence of interfacial transport and bulk diffusion on the oxidation of oleic acid by ozone, Atmos. Chem. Phys., 10, 3673–3691, 2010.

Shiraiwa, M., Pfrang, C., Koop, T., and Pöschl, U.: Kinetic multi-layer model of gas-particle interactions in aerosols and clouds (KM-GAP): linking condensation, evaporation and chemical reactions of organics, oxidants and water, Atmos. Chem. Phys., 12, 2777–2794, 2012a.

Shiraiwa, M., Selzle, K., and Pöschl, U.: Hazardous components and health effects of atmospheric aerosol particles: reactive oxygen species, soot, polycyclic aromatic compounds and allergenic proteins, Free Radic. Res., 46, 927–939, 2012b.

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmos. Chem. Phys., 14, 8323–8341, 2014.

Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing, Rev. Geophys., 55, 509–559, 2017.

Steiner, A. L.: Role of the Terrestrial Biosphere in Atmospheric Chemistry and Climate, Acc. Chem. Res., 53, 1260–1268, 2020.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., and Aspuru-Guzik, A.: Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries, J. Mater. Chem. A, 7, 12 833–12 841, 2019.

Tang, D., Zhan, Y., and Yang, F.: A review of machine learning for modeling air quality: Overlooked but important issues, Atmos. Res., 300, 107 261, 2024.

Tao, W., Chen, J., Li, Z., Wang, C., and Zhang, C.: Impact of aerosols on convective clouds and precipitation, Rev. Geophys., 50, 2011RG000 369, 2012.

Tolocka, M. P., Jang, M., Ginter, J. M., Cox, F. J., Kamens, R. M., and Johnston, M. V.: Formation of Oligomers in Secondary Organic Aerosol, Environ. Sci. Technol., 38, 1428–1434, 2004.

Vance, T. C., Huang, T., and Butler, K. A.: Big data in Earth science: Emerging practice and promise, Science, 383, eadh9607, 2024.

Vignati, E., Facchini, M., Rinaldi, M., Scannell, C., Ceburnis, D., Sciare, J., Kanakidou, M., Myriokefalitakis, S., Dentener, F., and O'Dowd, C.: Global scale emission and distribution of sea-spray aerosol: Sea-salt and organic enrichment, Atmos. Environ., 44, 670–677, 2010.

Wang, C., Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q., and Wania, F.: Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products, Atmos. Chem. Phys., 17, 7529–7540, 2017.

Weber, T., Corotan, A., Hutchinson, B., Kravitz, B., and Link, R.: Technical note: Deep learning for creating surrogate models of precipitation in Earth system models, Atmos. Chem. Phys., 20, 2303–2317, 2020.

Wilson, J., Pöschl, U., Shiraiwa, M., and Berkemeier, T.: Non-equilibrium interplay between gas–particle partitioning and multiphase chemical reactions of semi-volatile

compounds: mechanistic insights and practical implications for atmospheric modeling of polycyclic aromatic hydrocarbons, Atmos. Chem. Phys., 21, 6175–6198, 2021.

Worsnop, D. R., Morris, J. W., Shi, Q., Davidovits, P., and Kolb, C. E.: A chemical kinetic model for reactive transformations of aerosol particles, Geophys. Res. Lett., 29, 2002.

Xu, L. and Penner, J. E.: Global simulations of nitrate and ammonium aerosols and their radiative effects, Atmos. Chem. Phys., 12, 9479–9504, 2012.

Yang, X., Guo, L., Zheng, Z., Riemer, N., and Tessum, C. W.: Atmospheric Chemistry Surrogate Modeling With Sparse Identification of Nonlinear Dynamics, J. Geophys. Res., 1, e2024JH000 132, 2024.

Yorks, J. E., Selmer, P. A., Kupchock, A., Nowottnick, E. P., Christian, K. E., Rusinek, D., Dacic, N., and McGill, M. J.: Aerosol and Cloud Detection Using Machine Learning Algorithms and Space-Based Lidar Data, Atmosphere, 12, 606, 2021.

A. Personal List of Publications

Journal Articles

1. Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B.A.M., Lammel, G., Schmidt, B., Pöschl, U., Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, *Nat. Sci.*, doi: 10.1002/ntls.20220016, (2022).

2. Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., Krieger, U.K.: Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks, *Geosci. Model Dev.*, doi: 10.5194/gmd-16-2037-2023, (2023).

3. Krüger, M., Mishra, A., Spichtinger, P., Pöschl, U., Berkemeier, T.: A numerical compass for experiment design in chemical kinetics and molecular property estimation, *J. Cheminform.*, doi: 10.1186/s13321-024-00825-0, (2024).

4. Krüger, M., Galeazzo, T., Eremets, I., Schmidt, B., Pöschl, U., Shiraiwa, M., Berkemeier, T.: Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC2NN), *Geosci. Model Dev.*, doi: 10.5194/gmd-18-7357-2025, (2025).

5. Krüger, M., Alves R.P., Kratz, A., Weber, B., Berkemeier, T.: Towards an annual carbon balance of biological soil crusts: parametric equations and neural networks to model gas exchange and net primary productivity, *in peer-review (Funct. Ecol.)*, (2025).

6. Krüger, M., Klingmüller, K., Rosanka, S., Mishra, A., Pöschl, U., Lelieveld, J., Pozzer, A., Berkemeier, T.: Global Health Map: Coupling EMAC with KM-SUB-ELF to estimate air pollution health effects using accurate iron soluble fractions, *in preparation*.

7. Alves, R.P., Kim, M., Krüger, M., Kratz, A., Berg, G., Berkemeier, T., Weber, B.: Impact of elevated CO_2 levels on biocrusts from South Africa, *in preparation*.

8. Mishra, A., Lelieveld S., Krüger M., Campbell C. J., Srivastava, D., Lanzafame, G. M., Tomaz, S., Favez, O., Bonnaire, N., Lucarelli, F., Alleman, L., Uzu, G., Jaffrezo, J-L., Green D. C., Priestman, M., Tremper, A. H., Barth, A., Kalberer, M., Bandowe, B. A. M., Lammel, G., Pöschl, U., Shahpoury, P., Albinet, A., Berkemeier,

T.: Chemical kinetics and reaction mechanisms of reactive species production and antioxidant depletion in different assays measuring aerosol oxidative potential, *in prepraration*.

9. Backes, A.T., Krüger, M., Mishra, A., Weller, M. G., Fröhlich-Nowoisky, J., Pöschl, U., Berkemeier, T.: Chemical kinetics of protein modification (nitration/oligomerization) in bubble reactors, *in preparation*.

10. Backes, A.T., Mishra, A., Krüger, M., Kocakanat, S., Weller, M.G., Fröhlich-Nowoisky, J., Pöschl, U., Berkemeier, T.: Reaction mechanism and kinetics of protein modification (nitration/oligomerization) by air pollutants (O_3/NO_2) in aqueous solution and its suppression by antioxidants (ascorbic acid), *in preparation*.

11. Berkemeier, T., Mishra, A., Kang, H.G., Radecka, M., Eremets, I., Backes, A.T., Krüger, M., Polscheit, F., Zhang, Z., Pöschl, U.: Kinetic multi-layer model of multiphase chemistry (KM3C) within the kinetic multi-layer meta model (KM-MEMO): automated model generation from atmospheric aerosol chemistry to air pollution health effects, *in preparation*.

12. Berkemeier, T., Mishra, A., Lelieveld, S., Krüger, M., Backes, A.T., Sies, H., Pöschl, U.: Reactive oxygen species production by endogenous sources and air pollutants in the respiratory tract, *in preparation*.

13. Berkemeier, T., Eremets, I., Polscheit, F., Krüger, M., Kang, H.G., Pöschl, U.: KM-GAP-SOOT CO_2/H_2O modelling, *in preparation*.

14. Krüger, M., Eremets, I., Pöschl, U., Berkemeier, T.: Runtime comparison between MATLAB, Python and Julia for differential equation-based models, *in preparation*.

15. Eremets, I., Krüger, M., Pöschl, U., Berkemeier, T.: Application of gradient boosting machine learning algorithms for the prediction of vapor pressures of atmospherically relevant compounds, *in preparation*.

16. Eremets, I., Krüger, M., Pöschl, U., Berkemeier, T.: Application of gradient boosting and graph convolutional neural networks for the prediction of Henry's law coefficients of atmospherically relevant compounds, *in preparation*.

17. Raj, S.S., de Angelis, I.H., Basic, S., Aardema, H.M., Slagter, H.A., Weber, J., Calleja, M.L., Krüger, M., Andreae, M.O., Dragoneas, A., Nillius, B., Walter, D., Berkemeier, T., Haug, G.H., Pöschl, U., Schiebel, R., Pöhlker, C.: Exploring aerosol size distributions from polar to tropical zones of the Atlantic Ocean, *in preparation*.

Oral presentations

1. Krüger, M., Mishra, A., Pöschl, U., Berkemeier, T.: The Kinetic Laboratory Compass: Using kinetic models and machine learning for experiment design, European Aerosol Conference, Malaga, Spain, Sept. 3-8, 2023.

2. Krüger, M., Berkemeier, T.: Machine learning for quantitative structure-activity relationship (QSAR) modelling, MC^3 4 Earth series, online, WiSe 2024/2025.

Poster presentations

1. Krüger, M., Mishra, A., Spichtinger, P., Pöschl, U., Berkemeier, T.: The Kinetic Laboratory Compass: Using kinetic models and machine learning for experiment design, workshop "Aerosols, Health and Climate: Gigacity and Future", Bad Honnef, Germany, Mar. 20-24 2023.

2. Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., Krieger, U.K.: Accelerating Models for Multiphase Chemical Kinetics through Machine Learning, SCALES conference, Mainz, Germany, Jun. 27-30 2023.

3. Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., Krieger, U.K.: Accelerating Models for Multiphase Chemical Kinetics through Machine Learning, European Aerosol Conference, Malaga, Spain, Sept. 3-8, 2023.

4. Krüger, M., Klingmüller, K., Rosanka, S., Mishra, A., Pöschl, U., Lelieveld, J., Pozzer, A., Berkemeier, T.: Global Health Map: Coupling EMAC and KM-SUB-ELF to estimate air pollution health effects using accurate iron soluble fractions, Earth and Solar System Research Partnership Meeting, Mainz, Germany, Jun. 2-4 2025.

5. Krüger, M., Klingmüller, K., Rosanka, S., Mishra, A., Pöschl, U., Lelieveld, J., Pozzer, A., Berkemeier, T.: Global Health Map: Coupling EMAC and KM-SUB-ELF to estimate air pollution health effects using accurate iron soluble fractions, European Aerosol Conference, Lecce, Spain, Aug. 31 - Sep. 5 2025.

B. Supplementary

Supplementary Material: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects

Matteo Krüger[1], Jake Wilson[1], Marco Wietzoreck[1], Benjamin A. Musa Bandowe[1], Gerhard Lammel[1], Bertil Schmidt[2], Ulrich Pöschl[1], and Thomas Berkemeier[1]

[1]Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany
[2]Johannes Gutenberg University, Institute of Computer Science , Staudingerweg 9, 55128 Mainz, Germany

Correspondence: Thomas Berkemeier (t.berkemeier@mpic.de)

Supplementary Note 1 - Data Sets

Data used in this work originates from two studies: Tabor et al. (2019) and Kristensen et al. (2020). The data obtained from Tabor et al. contains structural information of molecules represented in canonical SMILES encoding. Kristensen et al. however supply a pdf file with the names of quinones in chemical nomenclature and images of their structural formula. In order to obtain canonical SMILES representations, a Python script is developed that automatically collects the names of quinones from the file and requests SMILES representations of the molecules from chemspider or pubchem (Pence and Williams, 2010; Kim et al., 2021). To ensure unity of representation and eliminate multiple occurrences of the same quinone, the merged data is transformed to rdkit.Chem.Mol molecule data structure and then reversed to canonical SMILES, using the MolFromSmiles() and MolToSmiles() functions provided by the python library *RDKit* (Landrum et al., 2010). Also, an algorithm is developed to eliminate any structure that does not contain two keto-groups directly associated to a conjugated carbon ring system.

Supplementary Note 2 - Data augmentation

As both data augmentation methods, ShiftSMILES and RandomSMILES are applied on single molecules and yield various numbers of augmentations depending on the molecule size (more ShiftSMILES for smaller molecules due to more empty characters in frame, fewer RandomSMILES for smaller molecules due to fewer variations), augmentation data is obtained in two steps for each of the methods to ensure similar numbers of both kinds of augmentation for each molecule. In the first step, a specified number of augmentations (m) are randomly sampled from all possible augmentations for one molecule and added to a pool. From this pool of alternations from all molecules, another specified number of augmentations (n) are then randomly sampled and added to the training data set. Regarding that RandomSMILES applies on the semantics of SMILES representation and ShiftSMILES on the data structure, RandomSMILES augmentations are obtained first (following the two steps), then ShiftSMILES augmentations are obtained for all molecules in the training data set, including the augmentations from RandomSMILES and again following the two-step procedure. This results in four parameters which define the extent of the performed data augmentation (k and n for both methods), referred to as m_r, n_r, m_s and n_s (Fig. S1).

To further ensure that a similar number of shifted augmentations for each molecule can be added to the first pool, including the largest one, which defines the size of the reading frame, within which ShiftSMILES is applied, this reading frame is extended by a specified number of empty characters, which is another parameter for the model training and will be referred to as ef (extend frame).

In the application of the augmentation methods, selected hyper parameter values can exceed the number of possible augmentations obtained from a molecule or from a pool. The following scenarios can be thought of:

- m_r: Small molecules often only allow a small number of RandomSMILES augmentations due to the lack of atoms in the molecule that allow different numbering. To avoid a strong bias in the training data set due to larger numbers of augmentations of larger molecules, m_r should not exceed a certain value. This value depends on the fraction of small molecules in the data and can be determined under consideration of the distribution of the maximum number of randomSMILES augmentations per molecule.

- m_s: Unlike m_r, the maximum number of m_s for one molecule can be limited for larger molecules, as it is restricted to

$$m_{Smax} = l_{max} + ef - l, \tag{S.1}$$

  where l_{max} is the number of characters of the largest molecule in data set and l is the number of characters in the molecule under consideration. Similarly to m_r, $(k_S - ef)$ should not exceed the maximum number of augmentations for a large fraction of molecules in the data set.

- n_s, n_r: The upper limits for augmentations of each method added to the training set are:

$$n_{Rmax} = m_R \cdot size(training\_data) \tag{S.2}$$

$$n_{Smax} = m_S \cdot (size(training\_data) + n_R) \tag{S.3}$$

  This limit can further decrease, if m_r and m_s are not fully obtained for all molecules. m_s and m_r are thus always set in a way, that on one hand the specified number of n_s and n_r can surely be satisfied, but on the other hand no strong bias results from a large fraction of molecules providing fewer than m augmentations.

In the augmentation experiment in this study, we set m_r to 10 if $n_\mathrm{r} < 5000$, 30 if $n_\mathrm{r} = 5000$ and 70 else. m_s is 25 if $n_\mathrm{r} = 0$ and $n_\mathrm{s} = 5000$ or if $n_\mathrm{r} = 500$ and $n_\mathrm{s} = 10000$, 50 if $n_\mathrm{r} = 0$ and $n_\mathrm{s} = 10000$ or 10 in any other case. Not all fields in the grid are tested, as the maximum number of augmentations is exceeded in one case ($n_\mathrm{s} = 10000$ if $n_\mathrm{r} = 0$).

Supplementary Note 3 - Convolutional Neural Network model and input data specifications

Convolutional layers, a specific type of layers in CNNs, perform convolution operations, allowing the detection of substructures and an association of their occurrence with the model output. The discrete convolution operation is described in Heaton (2018) with the following equations:

$$s(i) = \sum_{a=-\infty}^{\infty} x(a)w(i-a) \tag{S.4}$$

where x is an input, w is a weighting function, often called *kernel*, i the input and $s(i)$ the output, or *feature map*. In most applications of CNNs, the input as well as the kernel are multidimensional arrays. As an example, a two-dimensional kernel K and image I applies in the following way:

$$S(i,j) = (I * K)(i,j) = \sum_{m}\sum_{n} I(m,n)K(i-m,j-n) \tag{S.5}$$

A schematic representation of the CNN model architecture of this work is displayed in Fig. S2. The shape of the two-dimensional input layer is determined by the number of features (30) and the maximum number of characters of one SMILES-encoded molecule in the full data set (166), added to extend_frame and the number of empty character-indicating lines appended at the top of the two-dimensional molecule representation. The list of all features, ordered the same way, features are ordered on the corresponding axis of the one-hot encoded data structure, is the following: *5, +, F, l, s, i, Br, 6, c, 1, 2, O, =, (, C,), P, H, -, none, 3, S, 4, o, N, #, 7, [, n,]*. Most frequent features, especially the ones often occurring in small relative distance in the one-dimensional SMILES-string, for instance 'O' and '=', representing a keto group when coherent, are placed in the center of the associated axis. Shifting the most frequent features to the center lead to a significant decrease in validation and test error of the CNN model (in association with the selection of larger filter sizes as CNN hyper-parameter, see Tab. S1), compared to features being encoded in an order resembling their overall frequency in the data, resulting in the most frequent ones positioned close to the left edge. This simple finding shows, that broad, systematic testing of feature order applied for one-hot encoding represents a logical extension of this work and an initial step in future applications of CNNs for molecular structure detection based on SMILES.

Additionally to the feature order, performance boosts for models that are not trained on shiftSMILES-augmented data have been achieved in basic testing, when a set number of characters indicating an empty section of the reading frame is shifted to the front of the molecule. This *enclosing* of the structure-encoding section of the molecule representation by five shifted characters and ef set to five is applied continuously in this work, when shiftSMILES is not applied in pre-processing.

As most applications of CNNs address images, videos or natural language processing, only few information about successful CNN architectures and hyper-parameters for the application on SMILES-encoded molecules, following a very different logic, are available. When training a classification CNN on the prediction of chemical motifs, based on SMILES representations of molecules, Hirohara et al. (2018) went through relatively broad ranges for all hyper-parameters resulting from their basic

architecture of two convolutional layers followed by a pooling- and fully connected layer (Hirohara et al., 2018). For this study, this basic architecture is adapted to a mutable number of convolutional layers, each followed by a pooling layer. Regarding the hyper-parameters defining those layers, along with the other hyper-parameters, the ranges from Hirohara et al. are taken into account and adapted to smaller ranges in order to compensate the larger hyper-parameter space resulting from the variable number of layers. As the predictions of the CNN in this study are not classifications, but regressions, the activation function of the output layer is set to *linear*. An overview of all hyper-parameters and tested ranges is displayed in Tab. S1.

Supplementary Note 4 - Computational execution and Software

All training of CNN models was conducted on GPU-nodes of the supercomputer Mogon II, operated by Johannes Gutenberg University Mainz. The cluster contains 30 GPU-nodes that could be used for this work, each containing six NVIDIA GeForce GTX 1080 Ti GPUs. One model training requests one GPU on one of the nodes and is fully trained within 0.5 to 100 hours, depending on CNN hyper-parameters and size of the input data set. Model parameters and test data are saved for each model, well performing trained models can also be saved and transferred to a local machine for further analysis. Data pre-processing is parallelized using the Python Multiprocessing library and distributed on single nodes containing 40 or 64 CPU-cores. A program called "Simple Linux Utility for Resource Management" (SLURM) is used on the Mogon II system to manage the allocation of jobs or so called *job-arrays* on the cluster, as well as simple "highest-level paralellization", for instance by submission of such job arrays (Feitelson et al., 2003).

All data pre-processing-, training- and evaluation methods in this work are developed in Python 3.7. The following list contains all relevant python libraries as well as the versions used in this work: keras (2.4.3), keras-preprocessing (1.1.2), matplotlib (3.3.3), numpy (1.18.5), pandas (1.1.4), pubchempy (1.0.4), rdkit (2020.9.2), scikit-learn (0.23.2), scipy (1.5.4), seaborn (0.11.1) and tensorflow (2.3.1). On Mogon II, the easybuild toolchain 2.1.0-fosscuda-2019b-Python-3.7.4 was loaded, which contained all packages required for the model training. A few of the libraries used in this toolchain differ in version from the libraries used on the local machine, in detail Keras (2.3.1), scikit-learn (0.21.3) and scipy (1.4.1). This however was considered when transferring data in any direction and was not associated with any errors. Job allocation on Mogon II was achieved using the SLURM workload manager, as pre-installed on the system (20.11).

Supplementary Note 5 - Feature map activations

The data relevant for this analysis is represented by the *feature maps*, as introduced in equation S.4. As one feature map is obtained for each filter in a layer, and each of the convolution layers in the CNN models evaluated in this work contains at least 128 filters, all feature maps of one convolution layer are condensed to one *sum feature map* by summing up the values of all feature maps field by field. Under consideration of the kernel and stride size in the corresponding layers, each position of the original input string can be associated to a certain area in each feature map, which is effected by this sting position, as schematically shown in Fig. S3.

By obtaining the maximum or minimum value of all fields in this area in the sum feature map, the three dimensional intermediate data structure that is passed from one convolution layer to the following pooling layer, can be reduced in dimensionality and mapped to each character of the input sting. Area positions can also be effected by padding or pooling operations possibly following a convolution layer. This mapping of feature map values to input sub-structures however merely represents some form of "activation scheme" with regards to a specific input. Its interpretation and relevance in terms of underlying principles regarding chemistry or model performance could be subject of further research.

A set of synthetically generated molecular structures of quinones, based on a set of template quinones and functional groups is displayed in Fig. S5. The coloring of the structures is according to the activation scheme of the third, or last convolution layer of the trained CNN. The identical positions of the different functional groups for one template quinone allow a direct comparison of their effects on the activation schemes. Some similarities across the different functional groups or the different template quinones indicate a possible association of underlying chemical principles with the activation schemes. For instance, amino groups (fourth column) appear to be associated with the absence of local maxima, indicated by the blue color of the group, as well as the associated conjugated carbon structure. In contrast, hydroxy-groups at the same positions of the template molecules are mostly associated with larger local maxima in the associated sum feature maps, indicated by the red color.

In our opinion, a clear, straightforward interpretation of the activation schemes with regards to model predictions or underlying chemical principles is not feasible, based on the data available. However, we speculate, that an in-depth investigation of such activation schemes for large amounts of data may allow the identification of certain principles for a specific model. It is important to note, that various types of activation schemes can be obtained from a deep neural network. Feature maps of different layers of the CNN, for instance, are more or less closely associated with in- or outputs. Additionally, instead of indicating local minima or maxima in the relevant areas of feature maps, other mathematical operations like sums could be applied to obtain the associated activation schemes. Feature maps generally contain all information associated to the previous layers' interpretation of the original input data, as well as all information needed by the following layers to make a prediction. Therefore, an association with underlying principles of model functionality or chemistry is likely for well-performing models.

Table S1. Hyper-parameters determining CNN model architecture as well as its training process are presented in this table with short descriptions. Elaborated explanations of the parameters as well as their possible values can be found in the Keras API (Gulli and Pal, 2017). Hyper-parameters displayed in **bold style** represent layer-specific hyper-parameters. They must be specified individually for each of the layers of the concerning type.

Hyper-parameter	Description	Tested range
activation_function	Activation function of conv.layers	["relu", "sigmoid" or "softmax"]
filters	Number of filters in conv. layers	[16, 1024]
window_size_c	Window frame size of conv. layers	[1, 12]
sitride_size_c	Stride size of conv. layers	[1, 6]
padding_c	Which kind of padding is performed after conv. layer	["valid" or "same"]
pooling_type	MaxPooling or AveragePooling	["max" or "average"]
window_size_p	Window frame size of pooling layer	[1, 12]
stride_size_p	Stride size of pooling layer	[1, 6]
padding_p	Which kind of padding is performed after pooling layer	["valid" or "same"]
global_pooling	If global pooling is performed after all conv. and pooling layers	[0 or 1]
batch_size	Size of batches used for each training step	[2, 128]
batch_norm	If batch normalization is applied after conv- layers	[0 or 1]
epochs	Number of training epochs	[8, 60]
learning_rate	Learning rate in NN training	[0.0001, 0.1]
decay	If decay is applied on the learning rate during training	[0 or 1]
optimizer	Optimizer function for training	[Adam or nAdam]
dropout	Dropout rate applied to fully connected hidden layer	[0, 1]

Table S2. CNN predictions for reduction potentials of quinones that can be found in atmospheric samples, to estimate the impact of these substances on atmospheric chemistry or public health. The model CNN_Tabor_Kristensen is used to predict their reduction potential. By application of the equation in Roginsky et al. (1999) that associates reduction potentials of quinones with their $\log(k_{\text{eff}})$ in an ascorbate assay, predicted reduction potentials are translated to OP proxies. The test root mean square error (RMSE) of the CNN model is 116 mV. Propagation of errors results in an average error of 1.67 for the OP proxies of the individual species.

Quinone	Reference	CNN Prediction (H_2O/NHE System, offset: -575.78 mV) [mV]	OP Proxy: log(k_{eff} / [$M^{-1}s^{-1}$])
1,2-Aceanthrenequinone	Walgraeve et al. (2010)	-333.61	0
1,2-Acenaphthenequinone	Walgraeve et al. (2010)	-398.15	0
1,2-Chrysenequinone	Toriba et al. (2016)	82.53	0
1,2-Naphthoquinone	Walgraeve et al. (2010)	-32.06	3.45
1,4-Anthraquinone	Walgraeve et al. (2010)	-158.14	1.63
1,4-Benzoquinone	Delgado-Saborit et al. (2013)	64.42	0
1,4-Chrysenequinone	Walgraeve et al. (2010)	-27.97	3.51
1,4-Naphthoquinone	Walgraeve et al. (2010)	-139	1.91
1,4-Phenanthrenequinone	Walgraeve et al. (2010)	-9.48	3.77
1,6-Pyrenequinone	Toriba et al. (2016)	-23.25	3.58
1,8-Pyrenequinone	Toriba et al. (2016)	121.2	0
1-Hydroxy-9,10-anthraquinone	Walgraeve et al. (2010)	-478.69	0
2,3-Dimethyl-9,10-anthraquinone	Walgraeve et al. (2010)	-615.6	0
2,5-Dimethyl-1,4-benzoquinone	Toriba et al. (2016)	-35.8	3.39
2,6-Dimethyl-1,4-benzoquinone	Toriba et al. (2016)	-53.83	3.14
2-Ethyl-9,10-anthraquinone	Walgraeve et al. (2010)	-470.56	0
2-Hydroxy-1,4-naphthoquinone	Walgraeve et al. (2010)	-175.58	1.38
2-Methyl-1,4-benzoquinone	Toriba et al. (2016)	102.05	0
2-Methyl-1,4-naphthoquinone	Walgraeve et al. (2010)	-240.27	0.45
2-Methyl-9,10-anthraquinone	Walgraeve et al. (2010)	-478.82	0
4,5-Pyrenequinone	Walgraeve et al. (2010)	-194.46	1.11
5,12-Naphthacenequinone	Walgraeve et al. (2010)	-567.44	0

5,6-Chrysenequinone	Walgraeve et al. (2010)	-128.89	2.06
5,6-Dimethoxy-1,4-naphthoquinone	Walgraeve et al. (2010)	-137.65	1.93
5-Hydroxy-1,4-naphthoquinone	Walgraeve et al. (2010)	-125.67	2.1
Benz[a]anthracene-7,12-dione	Walgraeve et al. (2010)	-402.9	0
9,10-Anthraquinone	Walgraeve et al. (2010)	-501.06	0
9,10-Phenanthrenequinone	Walgraeve et al. (2010)	-189.54	1.18
Benzo[a]pyrene-1,6-dione	Walgraeve et al. (2010)	-253.94	0
Benzo[a]pyrene-11,12-dione	Walgraeve et al. (2010)	-149.71	1.76
Benzo[a]pyrene-3,6-dione	Walgraeve et al. (2010)	-146.42	1.8
Benzo[a]pyrene-4,5-dione	Walgraeve et al. (2010)	-163.98	1.55
Benzo[a]pyrene-6,12-dione	Walgraeve et al. (2010)	-205.35	0.96
Benzo[a]pyrene-7,10-dione	Toriba et al. (2016)	-200.05	1.03
Benzo[a]pyrene-7,8-dione	Walgraeve et al. (2010)	-247.62	0.35
Benzo[c]phenanthrene-1,4-dione	Toriba et al. (2016)	21.49	4.22
Benzo[c]phenanthrene-5,6-dione	Toriba et al. (2016)	-121.8	2.16
Benzo[e]pyrene-4,5-dione	Toriba et al. (2016)	-141.8	1.87
Dibenzo[a,h]anthracene-5,6-dione	Toriba et al. (2016)	-248.42	0.34
Dibenzo[a,j]anthracene-7,14-dione	Toriba et al. (2016)	-387.49	0
2,3-Flouranthenequinone	Toriba et al. (2016)	-163.4	1.56
Tetramethyl-1,4-benzoquinone (Duro-quinone)	Toriba et al. (2016)	-59.83	3.05
6,12-Anthanthrenequinone	Durant et al. (1998)	-270.03	0
6,13-Pentacenequinone		-671.08	0
1,8-Dihydroxyphenanthraquinone		-291.93	0
1,3-Indanedione	Ehrenhauser et al. (2012)	-334.39	0
Cyclopenta[cd]pyrenedione	Hannigan et al. (1998)	-236.23	0.51
2-Hydroxy-benzo[a]pyrene-1,6-dione	Pöschl (2002)	-396.44	0
4H-Cyclopenta[def]phenanthrene-8,9-dione	Huang et al. (2020)	-381.09	0

Table S3. CNN predictions for reduction potentials of quinone species that act as metabolites formed in living systems, according to literature. The model CNN_Tabor_Kristensen is used to predict their reduction potential. By application of the equation in Roginsky et al. (1999) that associates reduction potentials of quinones with their log(k_{eff}) in an ascorbate assay, predicted reduction potentials are translated to OP proxies. The test RMSE of the CNN model is 116 mV. Propagation of errors results in an average error of 1.67 for the OP proxies of the individual species.

Quinone	Reference	CNN Prediction (H_2O/NHE System, offset: -575.78 mV) [mV]	OP Proxy: log(k_{eff} / [$M^{-1}s^{-1}$])
Benzo[a]pyrene-7,8-dione	Clergé et al. (2019)	-247.62	0.35
1,2-Naphthoquinone	Clergé et al. (2019)	-32.06	3.45
1,4-Naphthoquinone	Clergé et al. (2019)	-139.0	1.91
1,2-Phenanthrenequinone	Clergé et al. (2019)	55.92	0
Benzo[a]pyrene-1,6-dione	Clergé et al. (2019)	-253.94	0
Benzo[a]pyrene-3,6-dione	Clergé et al. (2019)	-146.42	1.8
Benzo[a]pyrene-6,12-dione	Clergé et al. (2019)	-205.35	0.96
1,4-Benzoquinone	O'Brien (1991)	64.42	0
2-Hydroxy-1,4-benzoquinone	O'Brien (1991)	-21.44	3.6
Benzo[a]anthracene-3,4-dione	Penning (2017)	29.1	4.33
5-Methyl-chrysene-1,2-dione	Penning (2017)	38.96	4.47
5-Methyl-chrysene-7,8-dione	Penning (2017)	-180.67	1.31

Table S4. CNN predictions for reduction potentials of quinones that are investigated with regards to lipid peroxidation, a pathway of toxicology, in Zhao et al. (2011). The model CNN_Tabor_Kristensen is used to predict their reduction potential. By application of the equation in Roginsky et al. (1999) that associates reduction potentials of quinones with their log(k_{eff}) in an ascorbate assay, predicted reduction potentials are translated to OP proxies. The test RMSE of the CNN model is 116 mV. Propagation of errors results in an average error of 1.68 for the OP proxies of the individual species.

Quinone	Reference	CNN Prediction (H_2O/NHE System, offset: -575.78 mV) [mV]	OP Proxy: log(k_{eff} / [$M^{-1}s^{-1}$])
1,4-Naphthtoquinone	Zhao et al. (2011)	-139.0	1.91
9,10-Anthraquinone	Zhao et al. (2011)	-501.06	0
9,10-Phenanthrenequinone	Zhao et al. (2011)	-189.54	1.18
4H-cyclopenta[def]phenanthrenequinone	Zhao et al. (2011)	-165.21	1.53
5,12-Tetracenequinone	Zhao et al. (2011)	-567.44	0
5,6-Chrysenequinone	Zhao et al. (2011)	-128.89	2.06
Benz[a]anthracene -7,12-dione	Zhao et al. (2011)	-402.9	0
4-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-346.32	0
5-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-714.51	0
6-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-453.43	0
8-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-449.23	0
9-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-450.6	0
10-Methylbenz[a]anthracene-7,12-dione	Zhao et al. (2011)	-475.04	0
1,2-Triphenylenequinone	Zhao et al. (2011)	18.36	4.17
1,2-Fluoranthenequinone	Zhao et al. (2011)	195.91	0
1,6-Pyrenequinone	Zhao et al. (2011)	-23.25	3.58
1,8-Pyrenequinone	Zhao et al. (2011)	121.2	0
6,13-Pentacenequinone	Zhao et al. (2011)	-671.08	0
Dibenzo[a,h]anthracene-7,14-dione	Zhao et al. (2011)	-355.7	0
Dibenzo[a,c]anthracene-9,14-dione	Zhao et al. (2011)	-397.2	0
Benzo[a]pyrene-1,6-dione	Zhao et al. (2011)	-253.94	0
Benzo[a]pyrene-3,6-dione	Zhao et al. (2011)	-146.42	1.8
Benzo[a]pyrene-6,12-dione	Zhao et al. (2011)	-205.35	0.96
Dibenzo[a,i]pyrene-5,8-dione	Zhao et al. (2011)	-361.9	0
Dibenzo[cd,jk]-pyrene-6,12-dione	Zhao et al. (2011)	-270.03	0
Benzo[ghi]perylene-7,8-dione	Zhao et al. (2011)	-68.55	2.92
Dibenzo[b,n]perylene-15,16-dione	Zhao et al. (2011)	-73.76	2.85

10

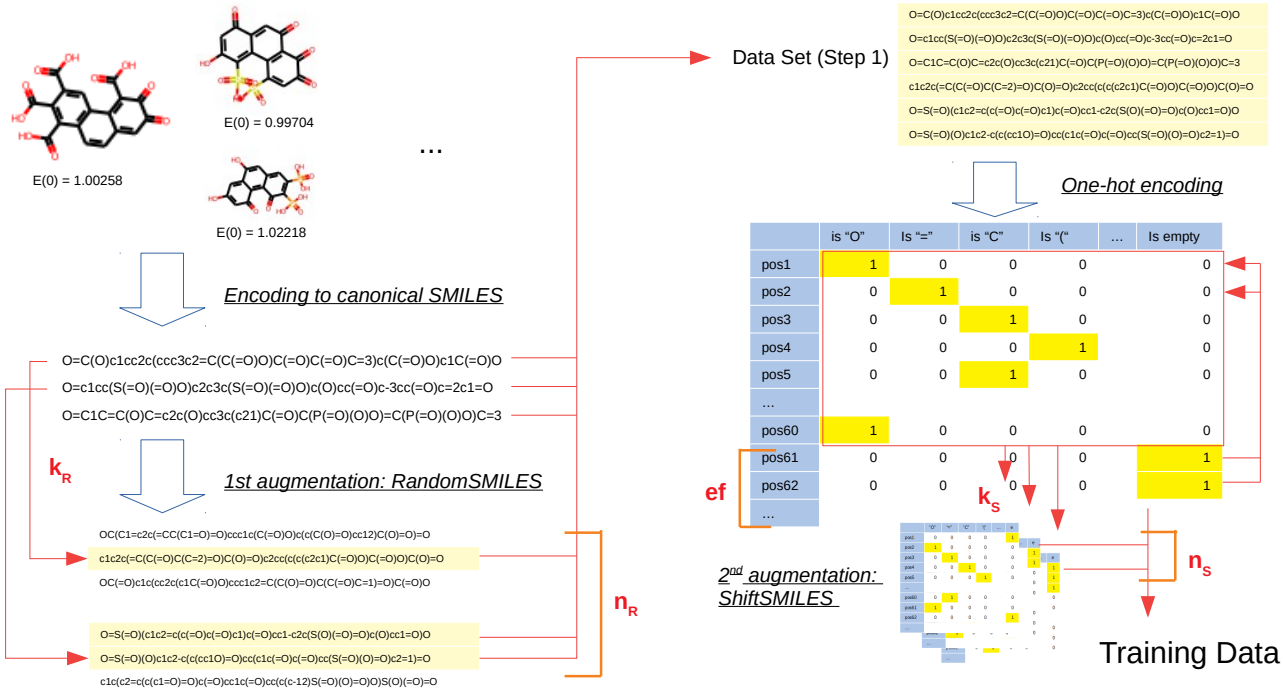

Figure S1. A methodic overview of the augmentation methods applied in this work. The relevant parameters are shown in red. The left side of the overview resembles the translation of molecules to SMILES and the first augmentation method, RandomSMILES, based on the semantics of SMILES encoding. m_r alternations for each molecule are generated and saved. Of all resulting augmentations, n_r are randomly picked, one-hot encoded (right column) and augmented via shiftSMILES, based on the empty positions of the overall reading frame, expanded by ef (extend frame). Of m_s augmentations for each molecule, n_s are again selected from the full collection and added to the training data. It is important to note, that additionally to the specified number of augmentations, the previously acquired original data is also fully added to the training data. Therefore, the number of SMILES-strings in the final training data is $n = original\_data + n_R + n_\mathrm{s}$.

References

Clergé, A., Le Goff, J., Lopez, C., Ledauphin, J., and Delépée, R.: Oxy-PAHs: occurrence in the environment and potential geno-toxic/mutagenic risk assessment for human health, Crit. Rev. Toxicol., 49, 302–328, 2019.

Delgado-Saborit, J. M., Alam, M. S., Pollitt, K. J. G., Stark, C., and Harrison, R. M.: Analysis of atmospheric concentrations of quinones and polycyclic aromatic hydrocarbons in vapour and particulate phases, Atmospheric Environ., 77, 974–982, 2013.

Durant, J. L., Lafleur, A. L., Plummer, E. F., Taghizadeh, K., Busby, W. F., and Thilly, W. G.: Human lymphoblast mutagens in urban airborne particles, Environ. Sci. Technol., 32, 1894–1906, 1998.

Ehrenhauser, F. S., Khadapkar, K., Wang, Y., Hutchings, J. W., Delhomme, O., Kommalapati, R. R., Herckes, P., Wornat, M. J., and Valsaraj, K. T.: Processing of atmospheric polycyclic aromatic hydrocarbons by fog in an urban environment, J. Environ. Monit., 14, 2566–2579, 2012.

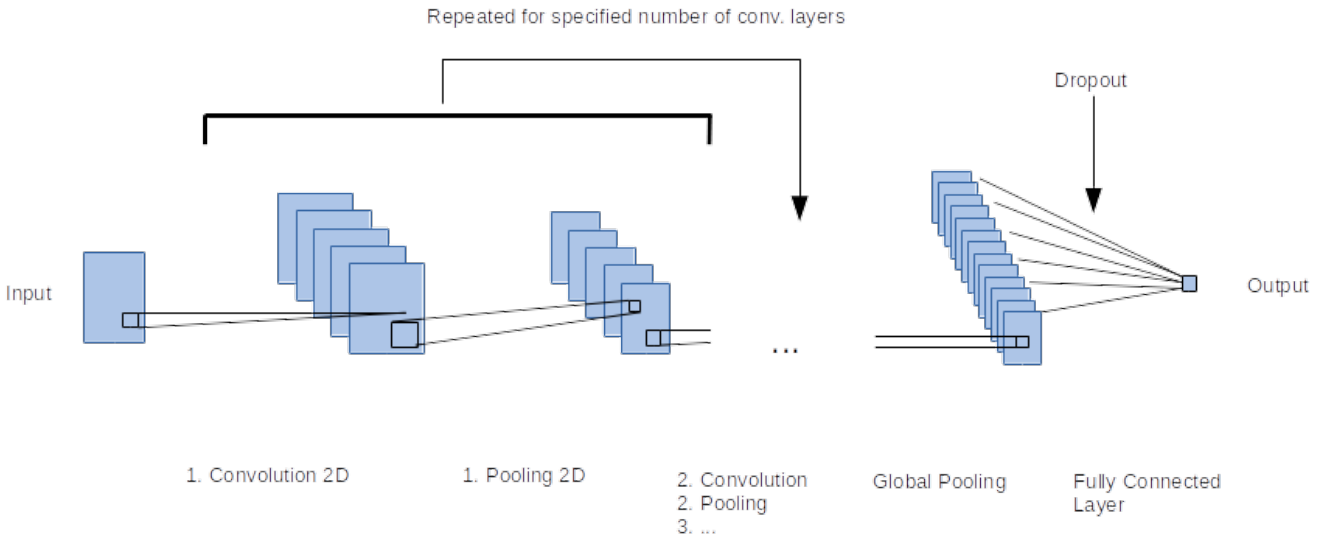

Figure S2. This schematic overview represents the basic architecture of the CNN model. An input layer is followed by a convolutional as well as pooling layer with specified number of filters. This sequence is repeated depending on the number of layers specified in the hyper-parameters. Following the last pooling layer, a global pooling layer, reducing the dimensionality of the data to 1 is located before the last fully connected output layer. A dropout rate is applied on this last fully connected layer, setting a specified fraction of weights to 0 in a training step, a method to ensure generalization of the model.

Feitelson, D., Rudolph, L., and Schwiegelshohn, U.: Job Scheduling Strategies for Parallel Processing: 9th International Workshop, JSSPP 2003, Seattle, WA, USA, June 24, 2003, Revised Papers, vol. 2862, Springer, 2003.

Gulli, A. and Pal, S.: Deep learning with Keras, Packt Publishing Ltd, 2017.

Hannigan, M. P., Cass, G. R., Penman, B. W., Crespi, C. L., Lafleur, A. L., Busby, W. F., Thilly, W. G., and Simoneit, B. R.: Bioassay-directed chemical analysis of Los Angeles airborne particulate matter using a human cell mutagenicity assay, Environ. Sci. Technol., 32, 3502–3514, 1998.

Heaton, J.: Ian goodfellow, yoshua bengio, and aaron courville: Deep learning, Springer, 2018.

Hirohara, M., Saito, Y., Koda, Y., Sato, K., and Sakakibara, Y.: Convolutional neural network based on SMILES representation of compounds for detecting chemical motif, BMC Bioinform., 19, https://doi.org/10.1186/s12859-018-2523-5, 2018.

Huang, R.-J., Yang, L., Shen, J., Yuan, W., Gong, Y., Guo, J., Cao, W., Duan, J., Ni, H., Zhu, C., et al.: Water-insoluble organics dominate brown carbon in wintertime urban aerosol of China: chemical characteristics and optical properties, Environ. Sci. Technol., 54, 7836–7847, 2020.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., et al.: PubChem in 2021: new data content and improved web interfaces, Nucleic acids research, 49, D1388–D1395, 2021.

Kristensen, S. B., van Mourik, T., Pedersen, T. B., Sørensen, J. L., and Muff, J.: Simulation of electrochemical properties of naturally occurring quinones, Sci. Rep., 10, https://doi.org/10.1038/s41598-020-70522-z, 2020.

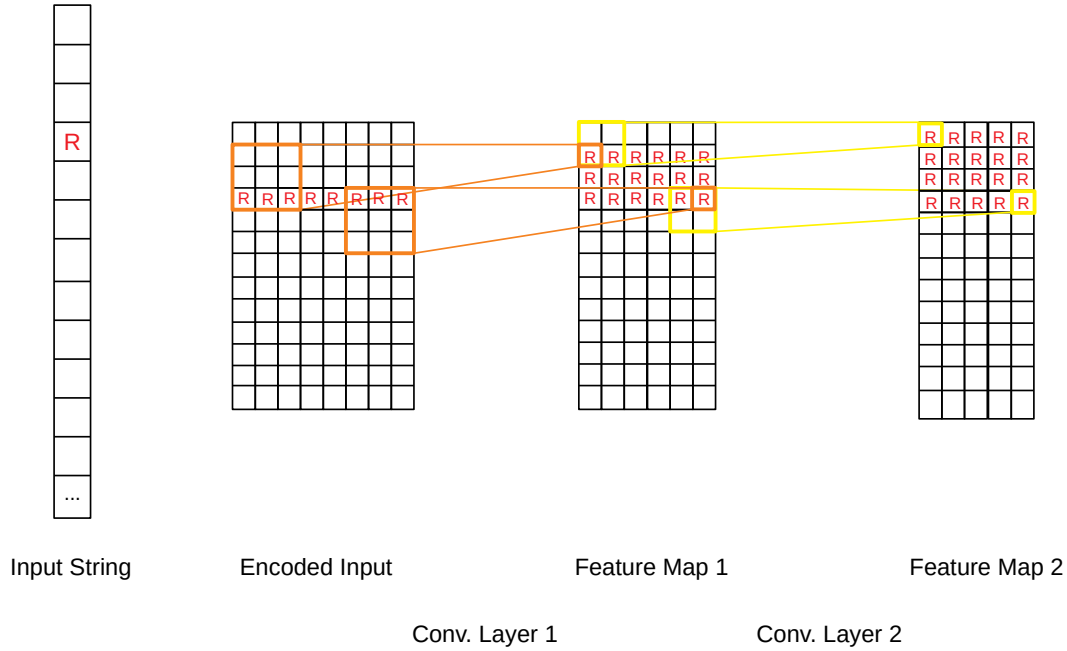

Figure S3. A schematic figure, representing the *area of effect* of a specific position of the input string on the feature maps in the first and second layer of a dummy CNN model with a filter size of 3 in the first and 2 in the second layer and a stride-size of 1 in both layers. The relevant positions in the one-and two- dimensional input representations are marked with an "R", as well as the fields in the feature maps influenced by this position. The convolution operation of the first layer is represented in orange, matching the first and last step addressing the relevant section, the second convolutional layer is equally represented in yellow.

Landrum, G. et al.: RDKit: Open-Source Cheminformatics Software, https://www.rdkit.org, 2010.

O'Brien, P.: Molecular mechanisms of quinone cytotoxicity, Chem. Biol. Interact., 80, 1–41, https://doi.org/https://doi.org/10.1016/0009-2797(91)90029-7, 1991.

Pence, H. E. and Williams, A.: ChemSpider: an online chemical information resource, 2010.

Penning, T. M.: Genotoxicity of ortho-quinones: reactive oxygen species versus covalent modification, Toxicol. Res., 6, 740–754, 2017.

Pöschl, U.: Formation and decomposition of hazardous chemical components contained in atmospheric aerosol particles, J. Aerosol Med. Pulm. Drug Delivery, 15, 203–212, 2002.

Roginsky, V. A., Barsukova, T. K., and Stegmann, H. B.: Kinetics of redox interaction between substituted quinones and ascorbate under aerobic conditions, Chem. Biol. Interact., 121, 177–197, https://doi.org/https://doi.org/10.1016/S0009-2797(99)00099-X, 1999.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., and Aspuru-Guzik, A.: Mapping the frontiers of quinone stability in aqueous media: Implications for organic aqueous redox flow batteries, J. Mater. Chem. A, 7, 12 833–12 841, https://doi.org/10.1039/c9ta03219c, 2019.

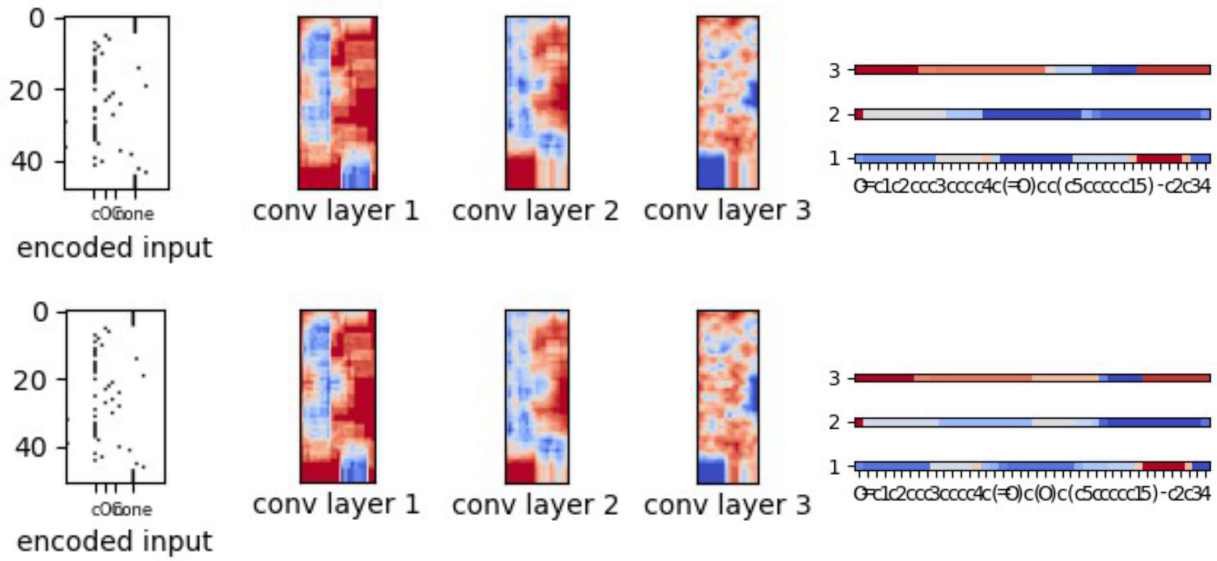

Figure S4. Feature maps of convolutional layers of the CNN model trained on the merged data set represent intermediate data structures that are passed from convolutional- to following layers. On the left, the input data (resembling a sparse, binary matrix) of two quinones are shown, which only differ regarding one hydroxy-group attached to the template quinone on top, Benzo(A)pyrene-6,12-dione. The following three heatmaps in each line represent the sum feature maps of each convolutional layer of this model. For each of the three sum feature maps, a one-dimensional positional activation map associated to the input string is presented on the right side. Dark red colors represent larger values, dark blue lower ones and are not transferable across the layers or molecules. For the first two layers, positional activations are values of local sum feature map field minima of the areas effected by the corresponding feature, but maxima for the third layer.

Toriba, A., Homma, C., Kita, M., Uozaki, W., Boongla, Y., Orakij, W., Tang, N., Kameda, T., and Hayakawa, K.: Simultaneous determination of polycyclic aromatic hydrocarbon quinones by gas chromatography-tandem mass spectrometry, following a one-pot reductive trimethylsilyl derivatization, J. Chromatogr. A, 1459, 89–100, 2016.

Walgraeve, C., Demeestere, K., Dewulf, J., Zimmermann, R., and Van Langenhove, H.: Oxygenated polycyclic aromatic hydrocarbons in atmospheric particulate matter: Molecular characterization and occurrence, Atmospheric Environ., 44, 1831–1846, 2010.

Zhao, Y., Xia, Q., Yin, J.-J., Yu, H., and Fu, P. P.: Photoirradiation of polycyclic aromatic hydrocarbon diones by UVA light leading to lipid peroxidation, Chemosphere, 85, 83–91, 2011.

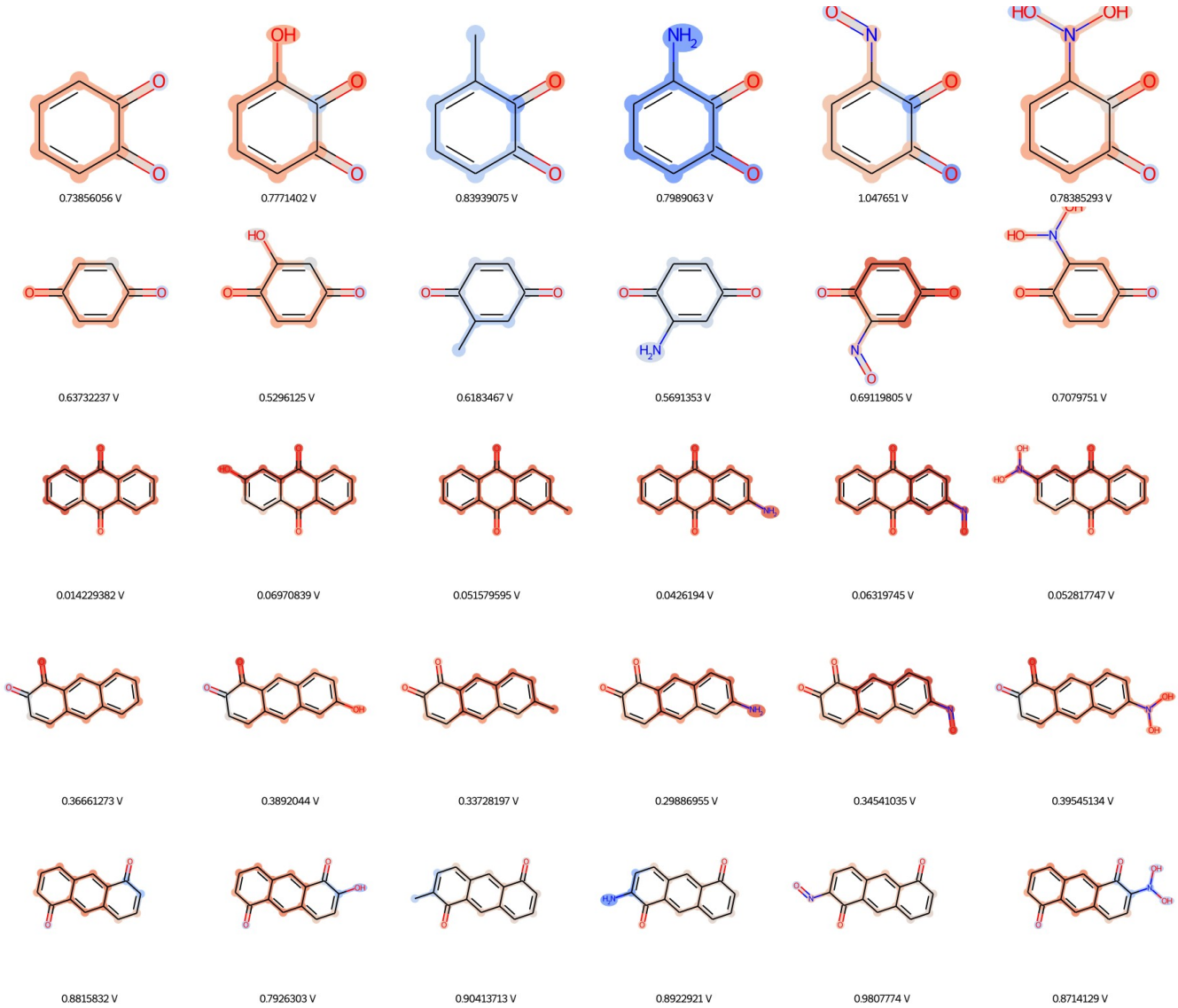

Figure S5. The one-dimensional feature map activation schemes of the third convolutional layer, as shown on the right in Fig. S4, associated with the one-dimensional SMILES representations of the synthetic input molecules are accordingly projected on the two-dimensional representation of their molecular structure. Columns represent groups of quinones, mostly identical in structure, the first molecule serving as template for the following molecules, obtained by the addition of a hydroxy-, methyl-, amine-, nitroso- or nitro-group at identical molecule positions. The template quinones shown are 1,2-Benzoquinone (ortho-), 1,4-Benzoquinone (para-), 9,10-Anthraquinone (para-), 1,2-Anthraquinone (ortho-) and 1,5-Anthraquinone (keto groups on multiple rings). Blue colors indicate low values of local maxima in the are of effect of associated input string positions, red colors large values. The mapping of values to colors is unified for all presented molecules, allowing comparison across the groups. Predictions of the CNN for the reduction potentials are provided for each of the shown molecules.

Supplement of Geosci. Model Dev., 18, 7357–7371, 2025
https://doi.org/10.5194/gmd-18-7357-2025-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^{2}NN)

Matteo Krüger et al.

Correspondence to: Manabu Shiraiwa (m.shiraiwa@uci.edu) and Thomas Berkemeier (t.berkemeier@mpic.de)

Table S1. Atom features represented in the feature map linked to individual nodes of the graph representation. To obtain the required features from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature name	Encoding	Possible values	Description
atom_type	OHE[1]	C, O, N, H, Cl, P, S, F, I, B, Br, Si	Element
n_heavy_neighbors	OHE[1]	0, 1, 2, 3, 4, MoreThanFour	Atom neighbors that are not H
formal_charge[2]	OHE[1]	-3, -2, -1, 0, 1, 2, 3, Extreme	Formal charge of atom
hybridisation_type	OHE[1]	S, SP, SP2, SP3, SP3D, SP3D2, OTHER	Atom hybridisation
is_in_a_ring	BOOL	0, 1	If atom is within ring structure
is_aromatic[2]	BOOL	0, 1	If atom is within conjugated structure
atomic_mass	FLOAT	-	Atomic mass in [u], scaled
vdw_radius	FLOAT	-	Van-der-Waals radius, scaled
covalent_radius	FLOAT	-	Covalent radius, scaled
chirality_type	OHE[1]	CHI_UNSPECIFIED, CHI_TETRAHEDRAL_CW, CHI_TETRAHEDRAL_CCW, CHI_OTHER	Chirality type
n_hydrogens	OHE[1]	0, 1, 2, 3, 4, MoreThanFour	Atom neighbors that are H

[1] One-hot-encoding [2] Omitted in confined data

Table S2. Bond features represented in the feature map linked to individual edges of the graph representation. To obtain the required features from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature name	Encoding	Possible values	Description
bond_type	OHE[1]	SINGLE, DOUBLE, TRIPLE, AROMATIC	Type of bond
bond_is_in_ring	BOOL	0, 1	If bond is within ring structure
bond_is_conj	BOOL	0, 1	If bond is conjugated
stereo_type	OHE[1]	Z, E, ANY, NONE	Stereo type of bond

[1] One-hot-encoding

Table S3: List of molecular descriptors passed to the group contribution component of GC^{2}NN models. To obtain the required molecular descriptors from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature	Description	Present in model
mass	Molar mass	all
NumAtoms	Number of atoms	all
NumBonds	Number of bonds	all
NumSingleBonds[1]	Number of single bonds	all
NumDoubleBonds[1]	Number of double bonds	all
NumTripleBonds[1]	Number of triple bonds	all
NumAromBonds[1]	Number of aromatic bonds	all
AromC	Number of aromatic carbon atoms	all
Charge	Formal charge	all
BertzCT	Bertz complexity index	all
Ipc	Structural information content	all
NumHDonors	Number of hydrogen donors	all
TPSA	Topological polar surface area	all
NHOHCount	Number of -NH and -OH groups	all
MolMR	Molar refractivity	all
VSA_EState3	EState indices for 3rd bin of VSA	all
AvgIpc	Average information content per atom	all
OC-ratio	Oxygen-carbon ratio	all
C^2	Carbon atoms	all
O^2	Oxygen atoms	all
N^2	Nitrogen atoms	all
Cl^2	Chlorine atoms	broad
I^2	Iodine atoms	broad
S^2	Sulfur atoms	broad
F^2	Fluorine atoms	broad
P^2	Phosphorus atoms	broad
Si^2	Silicon atoms	broad
Br^2	Bromine atoms	broad
B^2	Boron atoms	broad
hydroxyl	Hydroxyl groups	all

Table S3: (continued)

Feature	Description	Present in model
ester	Ester groups	all
carbonyl	Carbonyl groups	all
carboxyle	Carboxyl groups	confined, broad
ketone	Ketone groups	GeckoQ
hydroperoxide	Hydroperoxide groups	GeckoQ
nitrate	Nitrate groups	GeckoQ
aldehyde	Aldehyde groups	GeckoQ
carbonic acid	Carbonic acid groups	GeckoQ
peroxide	Peroxide groups	GeckoQ
carbonylperoxynitrate	Carbonylperoxynitrate groups	GeckoQ
ether	Ether groups	GeckoQ
nitro	Nitro groups	broad, GeckoQ
nitroester	Nitroester groups	GeckoQ
amine	Amine groups	broad
amide	Amide groups	broad
sulfide	Sulfide groups	broad
nitrile	Nitrile groups	broad

[1] Normalized as fraction of all bonds in compound [2] Normalized as fraction of all atoms in compound

Table S4. GC^2NN hyperparameter description and tested ranges.

Hyperparameter	Description	Tested range
num_conv_layers	Number of graph conv. layers	[2, 8]
num_conv_nodes	Number of nodes in each conv. layer	[8, 128]
num_hidden_layers	Number of additional fully-connected hidden layers	[0, 2]
hidden_layer_nodes	Number of nodes in each additional fully-connected layer	[8, 128]
num_merging_layers	Number of merging layers	[0, 2]
merging_layer_nodes	Number of nodes in each merging layer	[8, 128]
learning_rate	Learning rate during training	$[1 \times 10^{-4}, 1 \times 10^{-2}]$
lr_decay	Learning rate decay in each training epoch	[0.97, 1.0]
weight_decay	L2 regularization to avoid large weights	0 or $[1 \times 10^{-5}, 1 \times 10^{-2}]$
acivations	Activation of each conv. layer	'ReLU', 'LeakyReLU', 'Tanh' or 'Sigmoid'
layer_types	Types of conv. layers	'GCN'[1] or 'GAT'[2]
heads	Number of attention heads in conv. layer[3]	[1, 8]
pass_edge_attr	If edge (bond) attributes are passed to a layer[3]	0 or 1
batch_size	Number of molecules in each training batch	[4, 32]

[1] Graph convolution layers (Zhang et al., 2019) [2] Graph attention layers (Veličković et al., 2017) [3] Only applicable to GAT layers

Table S5. Selected hyperparameters for $fdGC^2NN$ models.

Hyperparameter	$fdGC^2NN$-confined	$fdGC^2NN$-broad	$fdGC^2NN$-GeckoQ
num_conv_layers	5	5	5
num_conv_nodes	[32, 64, 32, 32, 32]	[256, 128, 32, 256, 64]	[64, 32, 128, 32, 16]
num_hidden_layers	1	1	1
hidden_layer_nodes	32	32	32
learning_rate	1.94×10^{-3}	9×10^{-4}	4×10^{-3}
lr_decay	0.986	0.989	0.988
weight_decay	0	0	0
acivations	['Tanh', 'LeakyReLU', 'ReLU', 'Tanh', 'Tanh']	['Tanh', 'ReLU', 'ReLU', 'LeakyReLU', 'LeakyReLU']	['Tanh', 'Tanh', 'ReLU', 'ReLU', 'Tanh']
layer_types	[GAT, GAT, GCN, GAT, GCN]	[GCN, GAT, GCN, GAT, GAT]	[GAT, GCN, GCN, GCN, GAT]
heads	[4, 7, 0, 1, 0]	[0, 3, 0, 6, 5]	[5, 0, 0, 0, 3]
pass_edge_attr	[0, 1, 0, 1, 0]	[0, 0, 0, 0, 1]	[0, 0, 0, 0, 1]
batch_norm_layers	[1, 0, 0, 0, 0]	[0, 1, 0, 1, 0]	[1, 0, 1, 0, 0]
batch_size	32	16	64

Table S6. Selected hyperparameters for adGC^{2}NN models.

Hyperparameter	adGC^{2}NN
num_conv_layers	5
num_conv_nodes	[32, 16, 64, 16, 32]
num_hidden_layers	2
hidden_layer_nodes	32, 32
num_merging_layers	1
merging_layer_nodes	8
learning_rate	6.25×10^{-4}
lr_decay	0.985
weight_decay	0
acivations	['LeakyReLU', 'LeakyReLU', 'ReLU', 'ReLU', 'LeakyReLU']
layer_types	[GCN, GAT, GCN, GCN, GAT]
heads	[0, 6, 0, 0, 6]
pass_edge_attr	[0, 1, 0, 0, 1]
batch_norm_layers	[0, 0, 0, 0, 0]
batch_size	4

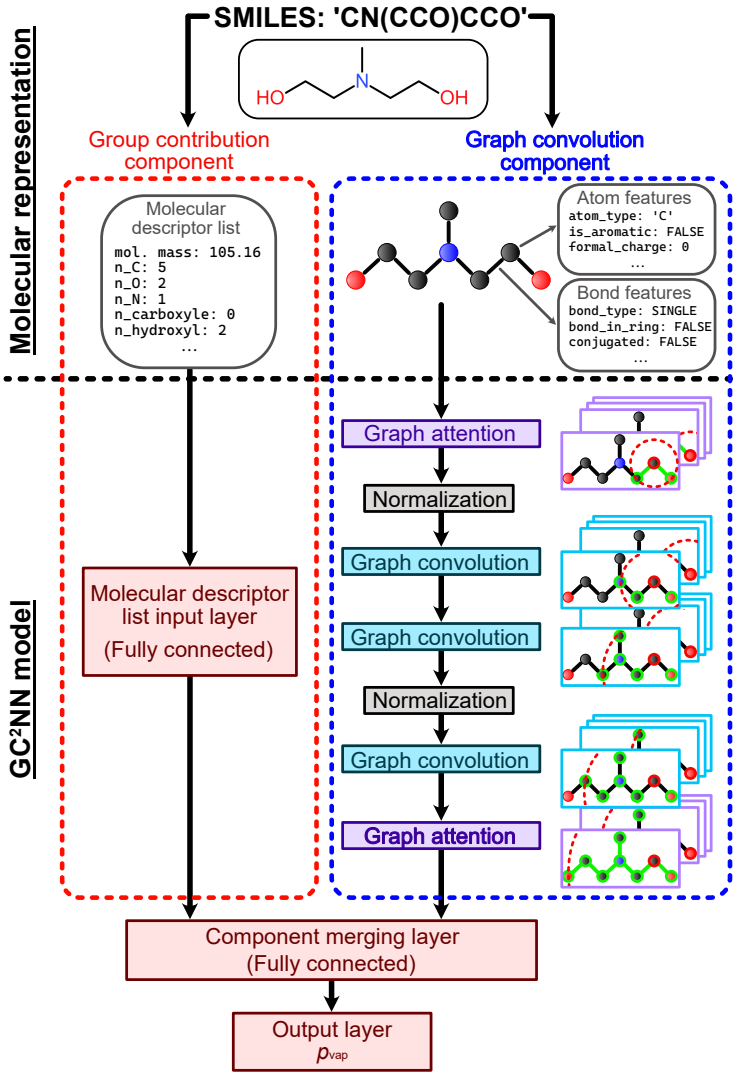

Figure S1. Schematic overview of molecular representation and model functionality of the fixed-depth GC^{2}NN (fdGC^{2}NN) model proposed in this work. Left: for the group contribution component, Simplified Molecular Input Line Entry System (SMILES) strings are used to derive holistic information on the molecule, such as its molar mass and the presence of atoms and functional groups (Tab. S3). Right: for the model's graph convolution component, SMILES strings are transformed into graph representations, encoded as adjacency matrices, node features, and edge features. This molecular representation is transformed using graph attention, graph convolution and batch normalization layers that normalize node or edge features across a batch, potentially stabilizing and accelerating the training. A fully-connected merging layer processes information from both model components and maps them to the single-node output layer, the p_{vap} prediction. Note that the displayed architecture represents model hyperparameters that were found optimal for a specific data set and model (fdGC^{2}NN-GeckoQ); the hyperparameters and thus architectures of other models presented in this study may deviate slightly in the type and order of layers in the graph convolution component (Tab. S5).

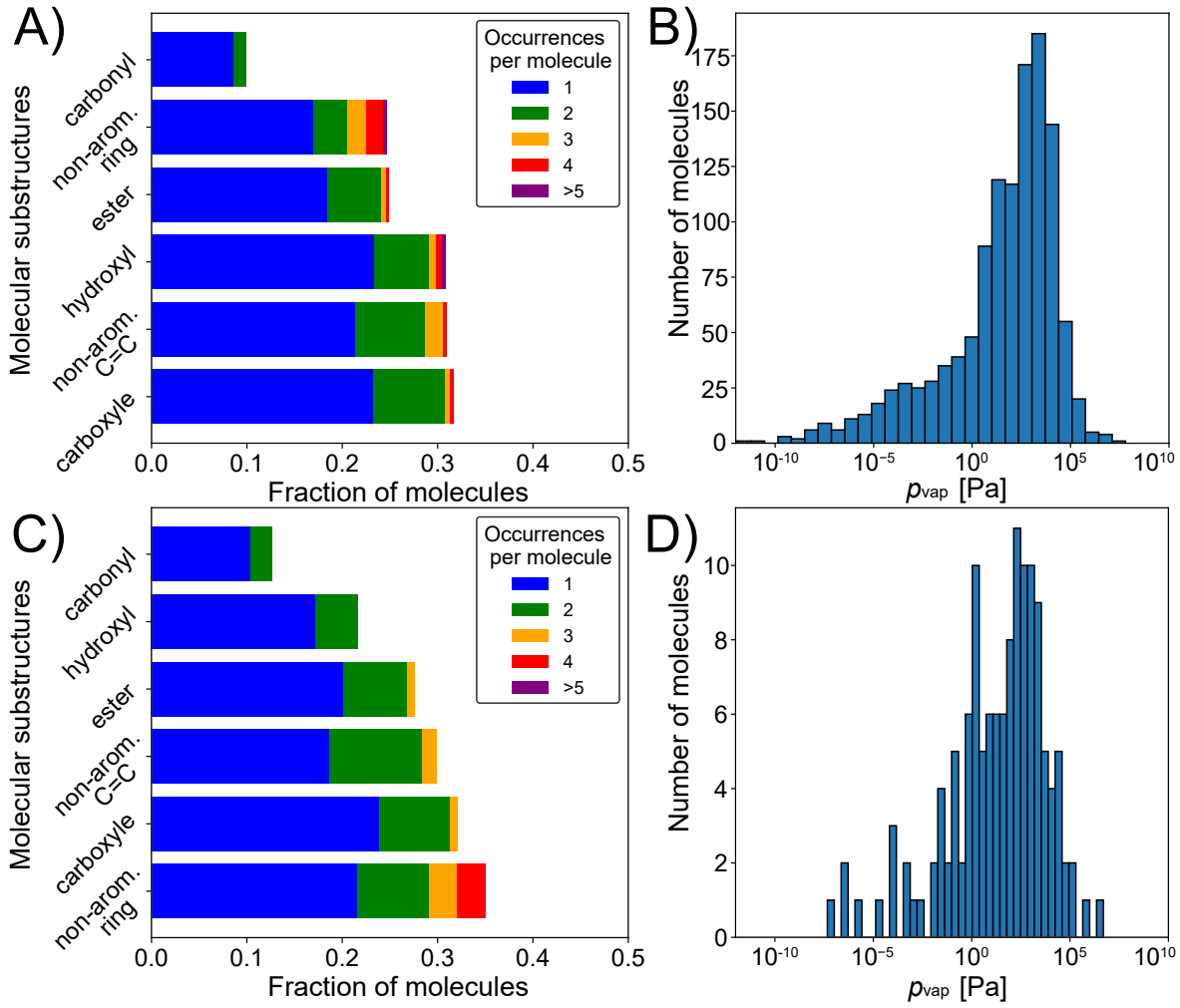

Figure S2. Occurrences of molecular substructures and vapor pressure measurements in the confined training plus validation (n = 1215; A, B) and test data set (n = 134; C, D), suitable for EVAPORATION (Compernolle et al., 2011). Panels A, and C show all substructures which are present in more than 1% of molecules in the respective data set. Panels B and D display histograms of experimental vapor pressure measurements in each data set. The distributions of molecular substructures and experimental vapor pressures are dissimilar, as 474 compounds present in the EVAPORATION training data are excluded from the test set.

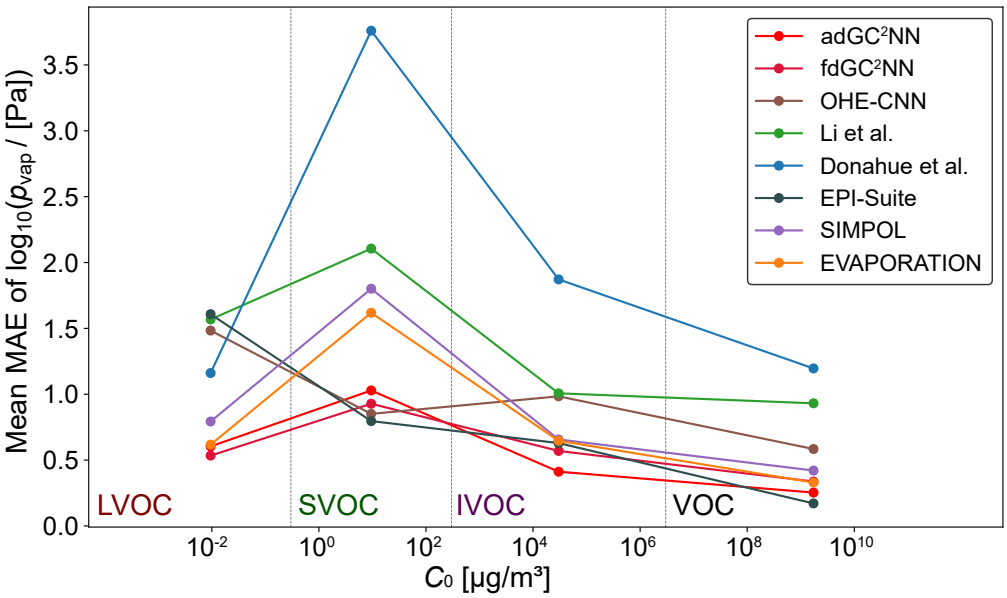

Figure S3. Mean confined test set prediction errors of four volatility bins as a function of experimental saturation concentration (C_0). Vertical dashed lines indicate interval borders of volatility bins. The number of compounds in each bin in the test set is ELVOC: 0, LVOC: 4, SVOC: 8, IVOC: 52, VOC: 70.

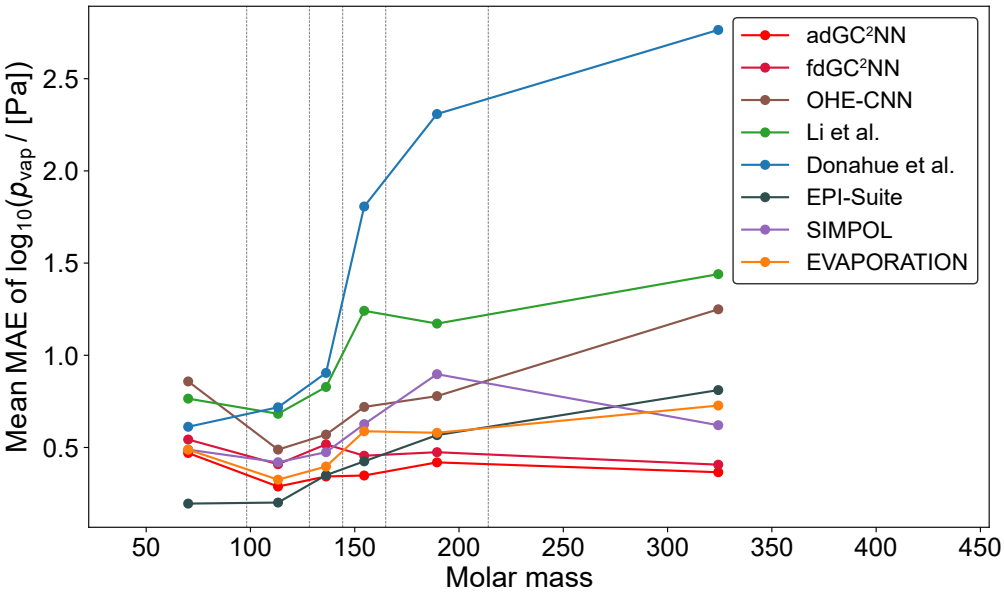

Figure S4. Mean confined test set prediction errors as a function of binned molecular masses. Vertical dashed lines indicate interval borders of mass bins. Bin intervals are selected so that each bin contains roughly 20 compounds from the test data set (22, 22, 21, 24, 22, 23).

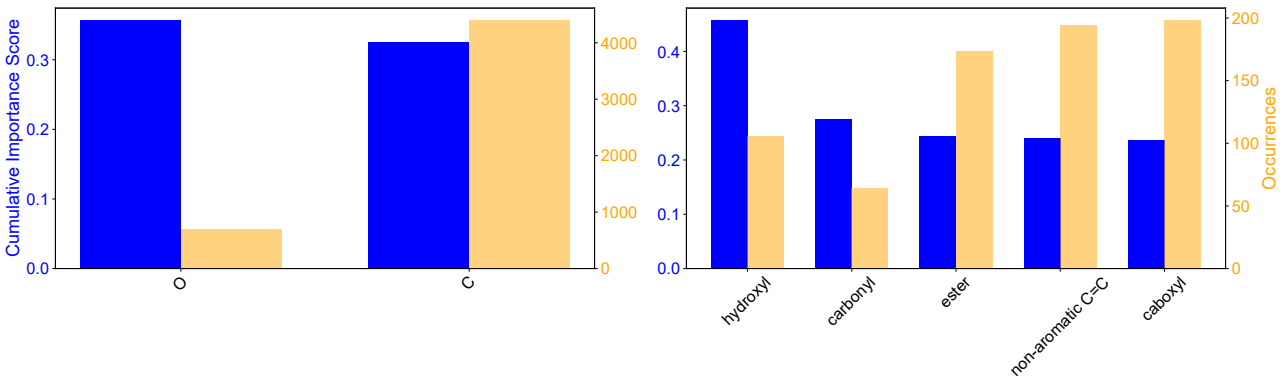

Figure S5. Cumulative importance scores and occurrences of atoms and functional groups in the confined test set (organic compounds with a limited set of functional groups), calculated in the second layer (graph attention layer) in the graph component of the trained T+V adGC^{2}NN-confined. Specifically, self-loop importances of the nodes attributed to various elements or functional groups are averaged to determine their relative importance among all neighboring nodes they are convoluted with.

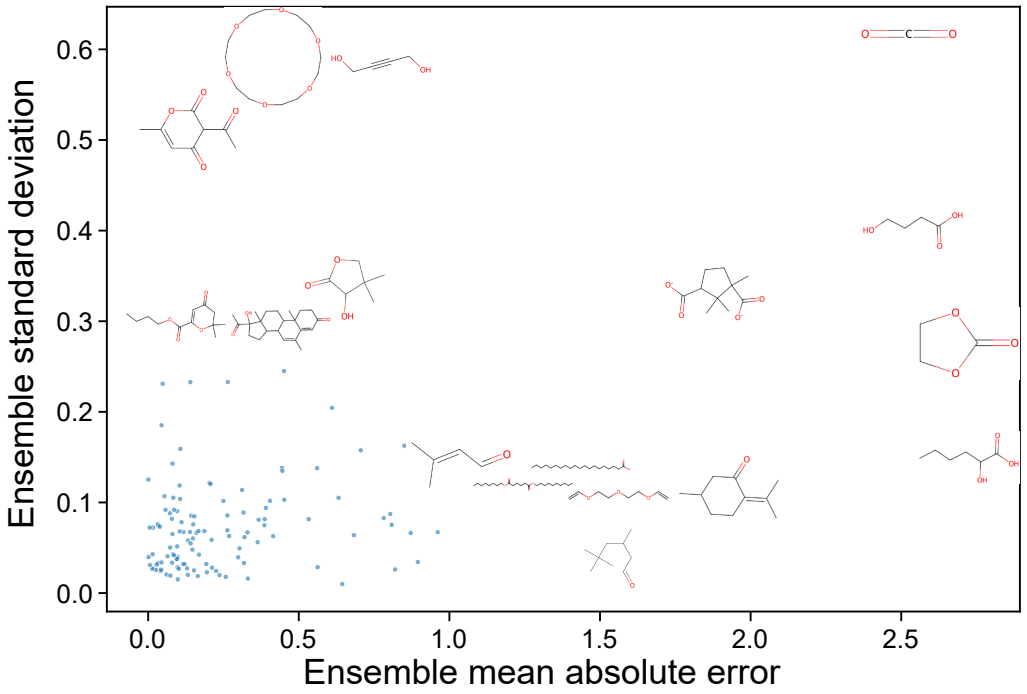

Figure S6. Ensemble standard deviation as a function of ensemble mean absolute error for the confined test set (organic compounds with a limited set of functional groups). Ensemble predictions originate from the confined adGC^{2}NN 5-fold cross validation models which are trained on different subsets of the training data. All compounds with an ensemble standard deviation larger than 0.3 or an ensemble mean absolute error larger than 1.0 are plotted as molecular structures. The compounds on the top of the figure are associated with large model uncertainty, while compounds in the bottom right have large errors despite small model uncertainty, a potential indicator for experimental uncertainty.

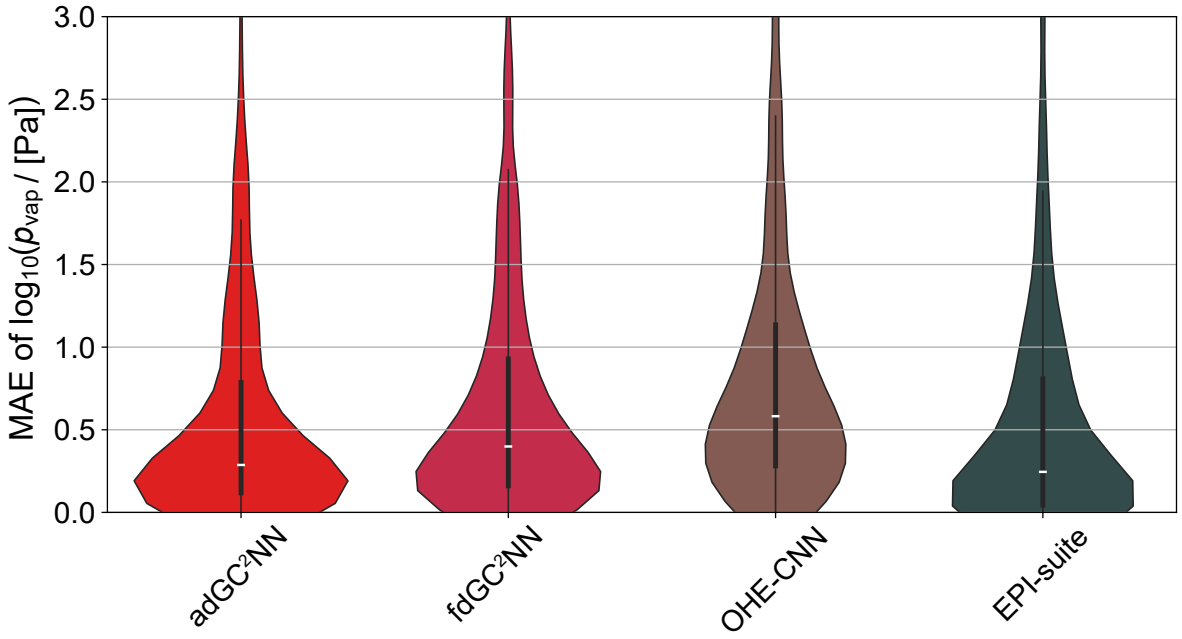

Figure S7. Violin plots representing broad test set error distribution of various models. Medians are shown as white markers, interquartile ranges as vertical wide black lines and 1.5 × interquartile ranges as narrow black lines.

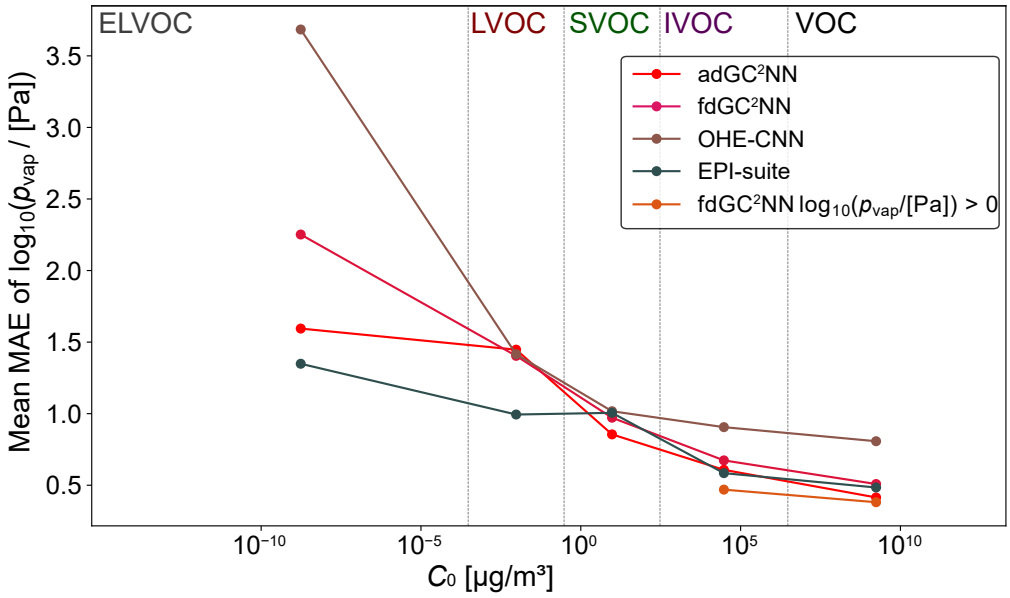

Figure S8. Mean broad test set prediction errors of five volatility bins as a function of experimental saturation concentration (C_0). Vertical dashed lines indicate interval borders of volatility bins. The number of compounds in each bin in the test set is ELVOC: 10, LVOC: 53, SVOC: 135, IVOC: 217, VOC: 202. An additional fdGC^{2}NN model is trained and tested on a subset of 3116 compounds (n_{train} = 2805, n_{test} = 311) with $\log_{10}(p_{vap}$ / [Pa]) > 0.

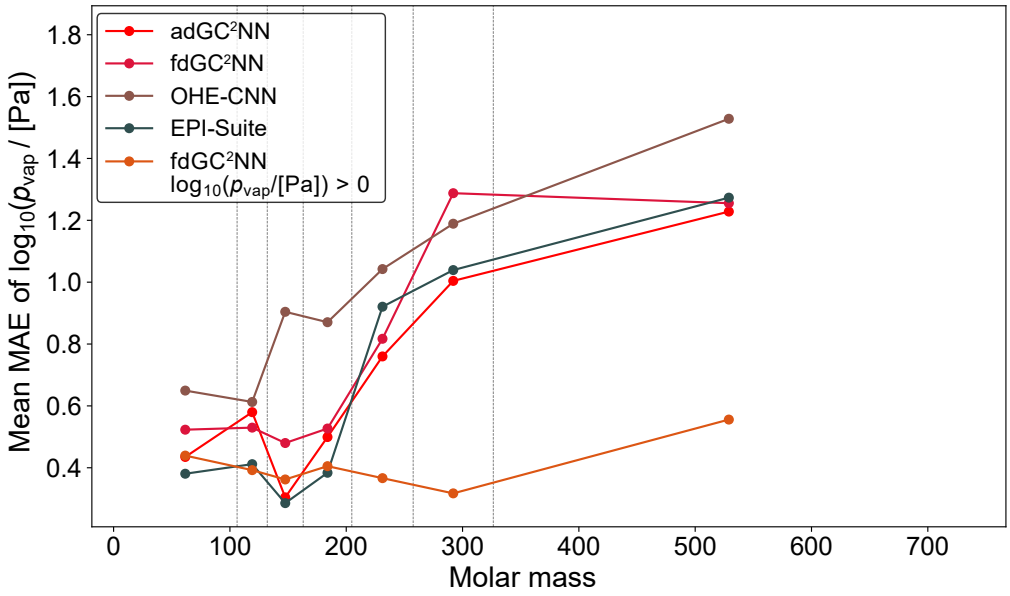

Figure S9. Mean broad test set prediction errors as a function of binned molecular masses. Vertical dashed lines indicate interval borders of mass bins. Bin intervals are selected so that each bin contains roughly 90 compounds from the test data set (88, 87, 88, 89, 88, 87, 89). An additional fdGC^{2}NN model is trained and tested on a subset of 3116 compounds (n_{train} = 2805, n_{test} = 311) with $\log_{10}(p_{vap}$ / [Pa]) > 0.

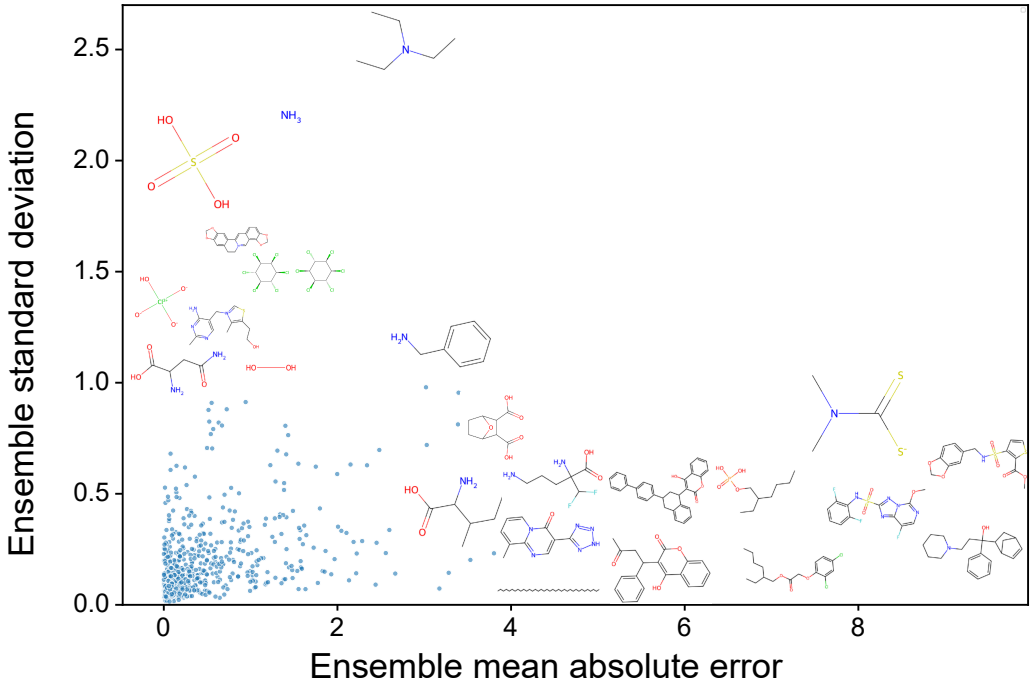

Figure S10. Ensemble standard deviation as a function of ensemble mean absolute error for the broad test set. Ensemble predictions originate from the broad adGC^{2}NN 5-fold cross validation models which are trained on different subsets of the training data. All compounds with an ensemble standard deviation larger than 1.0 or an ensemble mean absolute error larger than 4.0 are plotted as molecular structures. The compounds on the top of the figure are associated with large model uncertainty, while compounds in the bottom right have large errors despite small model uncertainty, a potential indicator for experimental uncertainty.

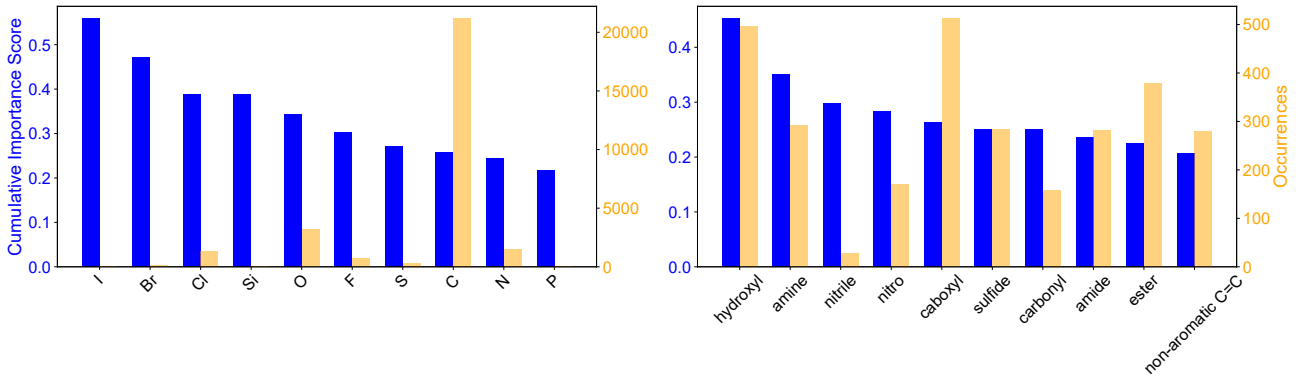

Figure S11. Cumulative importance scores and occurrences of atoms and functional groups in the broad test set (including inorganic compounds), calculated in the second layer (graph attention layer) in the graph component of the trained T+V adGC^{2}NN-broad. Specifically, self-loop importances of the nodes attributed to various elements or functional groups are averaged to determine their relative importance among all neighboring nodes they are convoluted with.

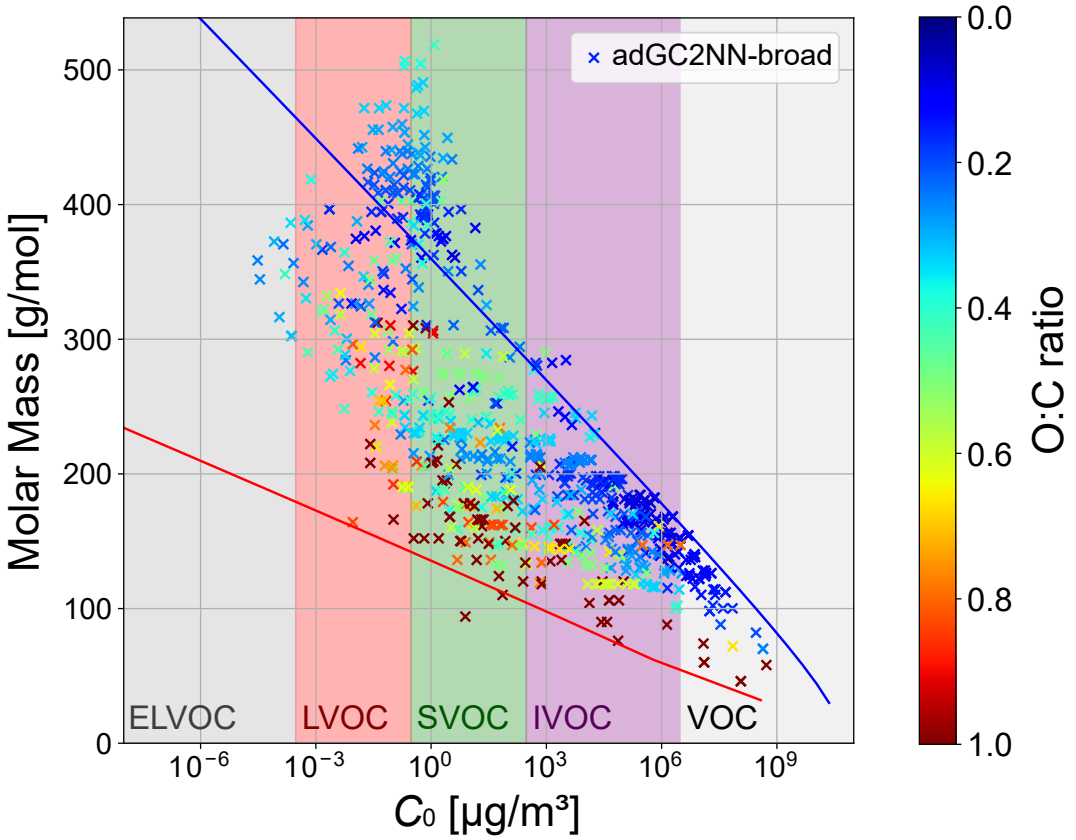

Figure S12. Molecular corridor plots following Shiraiwa et al. (2014). Application of the adGC^{2}NN-broad model to a data set of atmospherically relevant compounds (Shiraiwa et al., 2014). Blue and red boundary lines correspond to the volatility of n-alkanes and sugar alcohols (as determined by EVAPORATION), respectively.

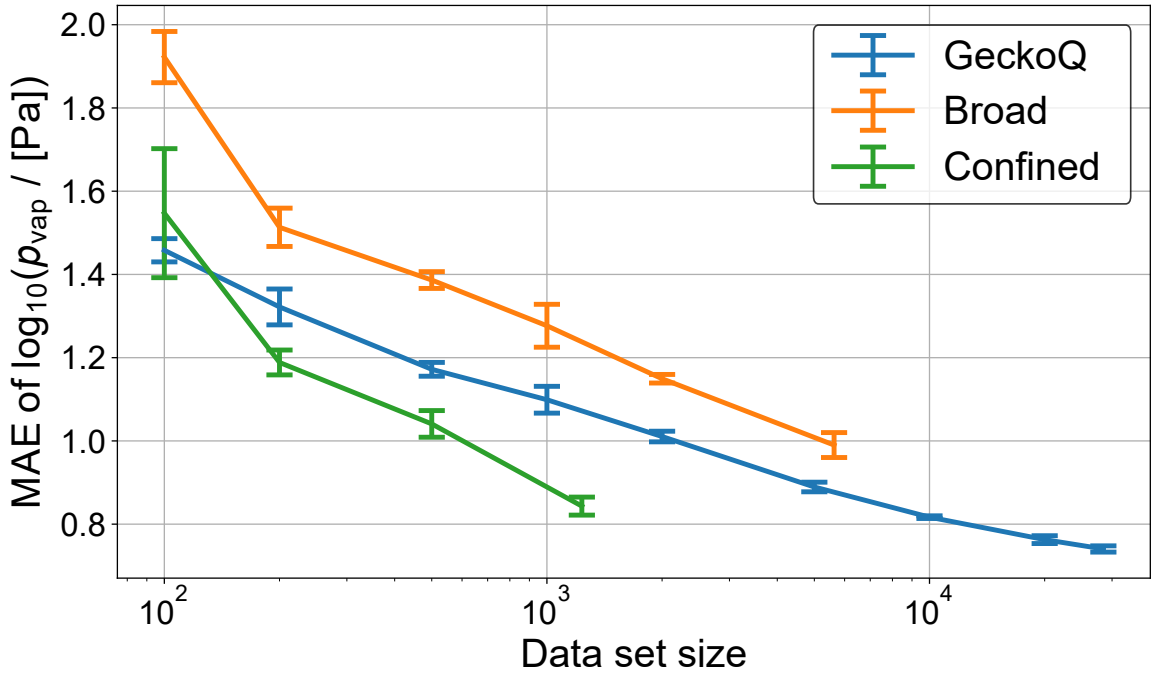

Figure S13. Mean absolute error (MAE) for independent test sets (confined: n = 137; broad: n = 625; GeckoQ: n = 3,163), as a function of training data set size of graph-only GCNN models trained on subsets of the three data sets. The experiment is performed by sampling subsets of various size from each of the respective data sets and training GCNN models on these. Hyperparameter tuning is performed for each subset. Shown are the average test set log unit MAE of five cross-validation models in each subset. Error bars represent standard deviations among the cross-validation folds.

References

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmos. Chem. Phys., 11, 9431–9450, https://doi.org/10.5194/acp-11-9431-2011, 2011.

Landrum, G.: RDKit: Open-source cheminformatics, Release, 1, 4, https://www.rdkit.org, 2013.

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmos. Chem. Phys., 14, 8323–8341, https://doi.org/10.5194/acp-14-8323-2014, 2014.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph Attention Networks, https://doi.org/10.48550/ARXIV.1710.10903, version Number: 3, 2017.

Zhang, S., Tong, H., Xu, J., and Maciejewski, R.: Graph convolutional networks: a comprehensive review, Comput Soc Netw, 6, 11, https://doi.org/10.1186/s40649-019-0069-y, 2019.

Supplement of Geosci. Model Dev., 16, 2037–2054, 2023
https://doi.org/10.5194/gmd-16-2037-2023-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks

Thomas Berkemeier et al.

Correspondence to: Thomas Berkemeier (t.berkemeier@mpic.de)

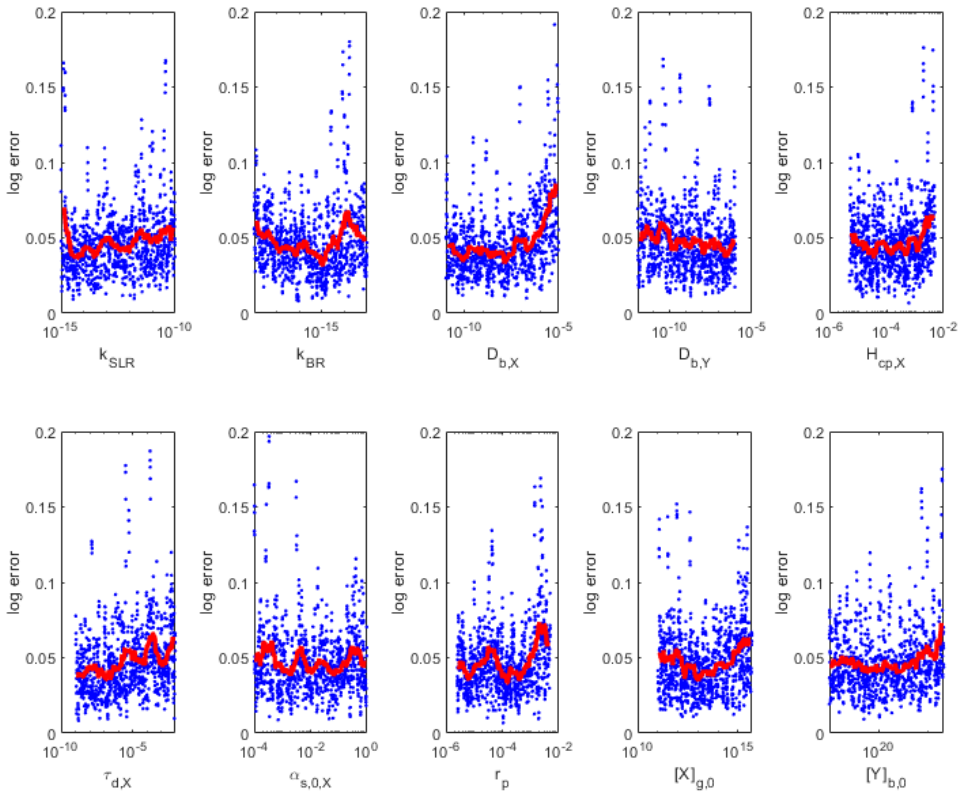

Figure S1. Absolute logarithmic error of the surrogate model for the 50 % reaction time (chemical half-life) in the test data set (N=1000) as a function of all 10 model input parameters. Due to the large size of the test data set and the variability in model error, errors are depicted as moving averages between the n=5 (blue dots) and n=100 (red lines) closest neighbours in a sorted list with ascending parameter values, respectively. No significant dependence of surrogate model error on parameter values is observed.

Supplementary Information for "A numerical compass for experiment design in chemical kinetics and molecular property estimation"

Matteo Krüger[1], Ashmi Mishra[1], Peter Spichtinger[2], Ulrich Pöschl[1], and Thomas Berkemeier[1]

[1]Multiphase Chemistry Department, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany
[2]Institute for Atmospheric Physics (IPA), Johannes Gutenberg University, Johann-Joachim-Becher-Weg 21, 55128 Mainz, Germany

Correspondence: Thomas Berkemeier (t.berkemeier@mpic.de)

Supplementary Note 1: Equations for process models, fit ensembles and prediction ensembles

The numerical compass (NC) method can be evaluated with any predictive process model M:

$$\mathrm{M} : \mathbb{R}^{n_\lambda + n_\mathrm{s}} \rightarrow \mathbb{R}^{n_z} \tag{S.1}$$

$$\mathrm{M}(\lambda, \mathrm{s}) = z \tag{S.2}$$

where λ are the kinetic input parameters $(\lambda_p)_{p=1,\ldots,n_\lambda}$, s the experimental (environmental or system) parameters $(\mathrm{s}_q)_{q=1,\ldots,n_\mathrm{s}}$ and z the model outputs $(z_m)_{m=1,\ldots,n_z}$.

With a specified acceptance threshold θ, we define the model solution space K as the set of fits, sets of kinetic parameter values in agreement with a given set of experiments with associated experimental parameters $\mathrm{S_{exp}}$ and measurements Y:

$$\mathrm{K}_{\mathrm{M},\mathrm{S_{exp}},Y} := \{\lambda : \Delta(\mathrm{M}(\lambda, \mathrm{S_{exp}}), Y) < \theta\} \tag{S.3}$$

where Δ is an error or distance metric for corresponding model outputs and experimental data, for instance a mean squared (absolute) logarithmic error (Eq. S.8).

From this model solution space, a finite number of ordered kinetic parameter sets λ with increasing error is described as fit ensemble (FE):

$$\mathrm{FE} = (\mathrm{FE}_l)_{l=1,\ldots,n_\mathrm{FE}} : \mathrm{FE}_l \in \mathrm{K} \wedge (\Delta(\mathrm{M}(\mathrm{FE}_l, \mathrm{S_{exp}}), Y) < \Delta(\mathrm{M}(\mathrm{FE}_{l+1}, \mathrm{S_{exp}}), Y)) \tag{S.4}$$

We define the sequence of outputs of a model M for all kinetic parameter sets of a FE of size n_{FE} in association with a single set of experimental parameters s as ensemble solution (ENS):

$$\mathrm{ENS} = (\mathrm{M}(\mathrm{FE}_l, \mathrm{s}))_{l=1,\ldots,n_{\mathrm{FE}}} \tag{S.5}$$

For the evaluation of the NC method, collections of ensemble solutions with varying experimental parameters are generated, a single ensemble solution is used to obtain an ensemble spread or a parameter constraint potential.

Supplementary Note 2: Equations for ensemble mean and standard deviation

Ensemble mean and ensemble standard variation for the determination of the ensemble spread are defined as:

$$\overline{Z_m} = \frac{\sum_{l=1}^{n_{\mathrm{FE}}} Z_{lm}}{n_{\mathrm{FE}}} \tag{S.6}$$

$$\sigma_m = \sqrt{\frac{1}{n_{\mathrm{FE}} - 1} \sum_{l=1}^{n_{\mathrm{FE}}} (Z_{lm} - \overline{Z_m})^2} \tag{S.7}$$

where n_{FE} is the number of fits in the fit ensemble and outputs in the associated ensemble solution, and Z_{lm} the output of the model for fit l in the ensemble and data point m.

Supplementary Note 3: Parameter boundary constraint potential metric with reduced sample density

In order to reduce the computational effort required for the evaluation of the parameter constraint potential metric, we sort ensemble solutions according to their error towards the mean of all predictions, and take into account if the difference of 50 % of $N_{\mathrm{Y},0}$ between the individual prediction and the ensemble solution average is negative or positive. Or simply put, if the fit's associated chemical half-life is smaller or larger than the average chemical half-life of the ensemble solution. By applying this order of fits on the ensemble solution, the parameter constraint potential can be approximated by evaluating a much smaller, evenly distributed fraction of all predictions as subset-forming elements, as demonstrated in Fig. S1.

Note that the comparison of parameter constraint potentials across various kinetic parameters or fit ensemble is problematic, as the measure is sensitive to the overall distribution of their individual values in the fit ensemble. We suggest to take this into account by subtracting the "background value" of the parameter constraint potential at its minimum in the constraint potential map from all parameter constraint potential values of this parameter. Additionally, the parameter constraint potential can be normalized by dividing the values by the logarithmic difference of the previous upper and lower boundary of the corresponding parameter. Such post-processing steps strongly depend on the distributions of kinetic parameter values in the underlying fit ensemble.

Furthermore, the initial assumption that every fit in the fit ensemble has a similar probability to represent the true physical values only applies, if the acceptance threshold for fit acceptance is selected sufficiently low. In applications where this is not the case, a weight could be associated with each subset, inversely proportional to the subset-forming fit's error in comparison with the previous experiments. This way, the larger probability of better fits in the fit ensemble to represent or resemble the physical truth is taken into account for parameter constraint potential calculation.

Supplementary Note 4: Oleic acid ozonolysis system applied in this study

We select r_p and $[\mathrm{O_3}]_\mathrm{g,0}$ (Tab. 1) as variable system parameters. The third system parameter $[\mathrm{OL}]_\mathrm{b,0}$ is set to $1.89 \cdot 10^{21}$ cm^{-3} to simplify this exemplary application and avoid the curse of dimensionality. In the range [$1 \cdot 10^{-6}$, $1 \cdot 10^{-2}$ cm] for r_p and [$1 \cdot 10^{10}$, $1 \cdot 10^{16}$ cm^{-3}] for $[\mathrm{O_3}]_\mathrm{g,0}$, we define a 100×100 log-uniform grid of potential experiments. For each point on this grid, we obtain a PE based on a pre-sampled FE with the associated experimental parameters r_p and $[\mathrm{O_3}]_\mathrm{g,0}$. Large values on the associated ES map are considered system parameters for potential experiments that are likely to lead to a large reduction of model solution space. Naturally, we propose the r_p and $[\mathrm{O_3}]_\mathrm{g,0}$ values associated with the absolute ES maximum to be the best experimental set-up to restrict the model solution space. Based on the system parameters, as well as the model outputs, we furthermore reject proposed experiments that would be difficult or impossible to conduct in a laboratory. To be accepted, suggested experiments may not exceed the following boundaries:

- 50 nm < r_p < 100 μm
- 10^{12} cm^{-3} < $[\mathrm{O_3}]_\mathrm{g,0}$ < 10^{16} cm^{-3}
- 1 s < predicted experiment duration < 3 d

As the experiment time can only be derived from the model outputs and is dependent on its kinetic parameters, we compute all simulated experiments with KM-SUB beforehand. All combinations of experimental parameters where at least one set of kinetic parameters leads to a simulated experiment with measurements exceeding the boundaries are not accepted as proposed experiments during the evaluation of the NC. The resulting boundaries are visualized in Fig. S3.

Supplementary Note 5: Surrogate model training

For the generation of the SM, we use feedforward multilayer-perceptrons with a maximum of three hidden layers provided by the Python library Keras (Chollet et al., 2015) and compute 1×10^6 random KM-SUB samples in log-uniform parameter space as training (990,000 samples) and test data (10,000 samples). The individual steps in KM-SUB sampling, data pre-processing, neural network model training and validation are elaborated in detail in Berkemeier et al. (2023). As the required SM is nearly identical with the one presented in this previous work with regards to in- and outputs of the template model, we adapt the suggested hyperparameters and only perform very basic hyperparameter tuning (<10 tested hyperparameter sets) applying 5-fold cross-validation to avoid over-fitting.

The SM selected for further evaluation achieves a test set mean square error (MSE) of 2.46×10^{-3}. The average test set MSE of the five cross-validation models is slightly larger at 3.06×10^{-3} and error variance of the five models low at 1.80×10^{-7}, an indication for no significant over-fitting. Average training time for an individual model on one NVIDIA GeForce GTX 1080 Ti is 6587.0 s (< 2 h). In comparison with the best-performing SM for KM-SUB presented in Berkemeier et al. (2023), we achieved a significant reduction of test errors by focusing brief hyperparameter tuning on the optimization of the individual layers' dropout rates. The hyperparameters of the model selected for this study are: Number of hidden layers: 2, numbers of neurons in layers: (4096, 4096), layer activations: ('relu', 'relu'), layer dropout rates: (0.2, 0.2), learning rate: 0.0001, learning rate decay: no, batch size: 16, epochs: 32.

Supplementary Note 6: Fit ensemble acquisition with KM-SUB and SM

In this study, we use seven experimental data sets of the ozonolysis of oleic acid aerosol available in the literature (Hearn and Smith, 2004; Ziemann, 2005; Gallimore et al., 2017; Müller et al., 2022) and a mean square logarithmic error (MSLE) to quantify the error of a matrix Z of model outputs $Z_{i,j}$ for experiment i with specified experimental parameters, and data point j, in comparison with the matrix Y of the corresponding experimental data $Y_{i,j}$ (Berkemeier et al., 2023):

$$\mathrm{MSLE}(Z,Y) = \frac{\sum_{i=1}^{n_{\mathrm{exp}}} \frac{\sum_{j=1}^{n_{\mathrm{d}}} (\log_{10}(Z_{ij}) - \log_{10}(Y_{ij}))^2}{n_{\mathrm{d}}}}{n_{\mathrm{exp}}} \tag{S.8}$$

where n_{exp} is the number of experimental data sets and n_{d} the number of data points in each set. Note that the measured decomposition steps in the experimental data are not always equal to the default output of the simplified KM-SUB that we use and may require an interpolation. We add an additional data point at $x = 1$, $z = 0$ to all individual model output sequences, as these initial conditions apply in every case (no decomposition at time 0) and apply a second order spline interpolation on the model outputs.

We use the two compared models, KM-SUB and the SM in turn to acquire a fit ensemble of 500 kinetic parameter sets each, using random batch sampling in log-uniform parameter space. Sampled parameter sets which result in a model output with a MSLE falling below θ = 0.0105 are added to the associated fit ensemble. Visualizations of seven ensemble solutions for experimental conditions corresponding to the seven experiments used in this study are shown in Fig. 3. A contrariwise cross-evaluation of each fit ensemble with the opposite model allows an estimation of "false-positive" and "false-negative" errors of the SM, and is visualized in Fig. S4. Plot matrices visualizing the input parameter distributions and densities of both fit ensembles are provided in in Fig. S5 and S6.

Supplementary Note 7: Uncertainty calibration and simulated experiments

For the testing of methods that suggest experiments, we perform simulations that include the generation of artificial experimental results based on KM-SUB assuming a single kinetic parameter set from the fit ensemble as the simulated physical truth. We

add uncertainty to the synthetic data in the form of uncertainty in experimental input variables and output values, mimicking the effect of errors in experimental setup and measurement, respectively. To model experimental uncertainty, we sample the value for each variable experimental parameter of the model and each output from a normal distribution in logarithmic space with the original value as the distribution's mean and a defined Σ_{unc} as its standard deviation. Individual values of Σ_{unc} are used for each experimental parameter (Σ_{rad} and Σ_{O_3}) and one for the outputs Σ_{out}. We define $\text{UC}(\text{I}, \Sigma_{\text{unc}})$ as the function that maps a single or multiple input values I to their uncertainty-values under consideration of the corresponding uncertainty parameter Σ_{unc}.

Σ_{rad}, Σ_{O_3} and Σ_{out} are calibrated with a method that tracks error development throughout multiple random simulated experiments. First, a single kinetic parameter set FE_l is selected and its error in association with the real experimental data Y obtained:

$$\delta_{\text{exp},l} = \text{MSLE}(\text{M}(\text{FE}_l, \text{S}_{\text{exp}}), Y) \tag{S.9}$$

For a matrix of n_{sim} randomly selected experimental parameter sets S_{sim}, under consideration of the boundaries and conditions presented in Suppl. Note 4, and for n_r repetitions, we obtain the following errors:

$$\delta_{\text{sim},l,i} = \frac{\sum_{u=1}^{n_r} \text{MSLE}(\text{M}(\text{FE}_l, \text{S}_{\text{sim},i}), \text{UC}(\text{M}(\text{FE}_l, \text{UC}(\text{S}_{\text{sim},i}, \Sigma_{\text{rad}}, \Sigma_{\text{X}}), \Sigma_{\text{out}})}{n_r} \tag{S.10}$$

where i is the index of an individual experimental parameter set $(\text{s}_q)_{q=1,\ldots,n_{\text{s}}}$ in S_{sim} and l the index of the kinetic parameter set initially selected. Simply put, we quantify the average error between an unmodified model output and multiple artificial experimental outputs with a set of pre-selected Σ_{unc} based on a single set of kinetic parameters from the fit ensemble. For visualization purposes, we apply an arbitrarily order to the simulated experiments and calculate the average error of all - real and simulated - experiments at each iteration v of the simulation:

$$\delta_{l,v} = \frac{\delta_{\text{exp},l} * n_{\text{exp}} + \sum_{i=1}^{v} \delta_{\text{sim},l,i}}{n_{\text{exp}} + n_{\text{sim}}} \tag{S.11}$$

The sequence $(\delta_{l,v})_{v=1,\ldots,n_{\text{sim}}}$ represents the hypothetical error development of multiple experiment simulations where only the simulated uncertainty Σ_{unc} contributes to the error, not the uncertainty of the model mechanism or the kinetic parameter set. Since these errors do contribute to the initial error $\delta_{\text{exp},l}$ to an unknown extent, we suggest a combination of Σ_{unc} that leads to a minor error decrease for kinetic parameter sets with a low $\delta_{\text{exp},l}$ and a larger decrease for those with a large $\delta_{\text{exp},l}$. Three examples are presented in Fig. S10, including the combination of Σ_{unc} that has been selected for this study (Σ_{rad} = 0.05; Σ_{O_3} = 0.02; Σ_{out} = 0.07; panel B). Note that only the concerted effect of the three uncertainty parameters can be calibrated with this approach. Based on individual experimental methods, and associated limitations in accuracy, we only consider combinations for $\Sigma_{\text{out}} > \Sigma_{\text{rad}} > \Sigma_{\text{O}_3}$.

Supplementary Note 8: Sensitivity analysis

To compare the proposed methods with a baseline strategy of experiment selection, we derive sensitivity maps based on the kinetic parameters in the fit ensembles after each simulated experiment, following an intuitive approach of testing conditions where kinetic model parameters are most sensitive. To obtain a total sensitivity map for all kinetic parameters, we vary parameters individually and sum up the absolute model residuals. In detail, we apply the normalized sensitivity models M_{SP} (partial sensitivity) or M_{ST} (total sensitivity) to generate ensemble solutions for the grid of experimental conditions and KM-SUB fit ensemble:

$$M_{SP}(\lambda, s, p) = \left| \frac{\frac{M(\lambda,s) - M(\lambda_{\bar{p}},s)}{M(\lambda,s)}}{\frac{\lambda_p - \lambda_{\bar{p},p}}{\lambda_p}} \right| \tag{S.12}$$

$$M_{ST}(\lambda, s) = \sum_{p=1}^{n_\lambda} \frac{M_{SP}(\lambda, s, p)}{n_\lambda} \tag{S.13}$$

where $\lambda_{\bar{p}}$ are parameter sets with parameter λ_p varied: $\lambda_{\bar{p},p} = \lambda_p * 1.2$. The ensemble solutions are then evaluated in the existing framework of the numerical compass with a simple constraint potential metric that selects the maximum of the average normalized sensitivities of associated experimental conditions.

Supplementary Note 9: Computational effort

The application of the NC method requires evaluations of the applied model in two consecutive steps, fit ensemble acquisition and ensemble solution generation. The application of SM in the fit ensemble acquisition step has been demonstrated in Berkemeier et al. (2023). SM accuracy is dependent on the training data size and has been tested for a wide range of such in Berkemeier et al. (2023). In this study, we arbitrarily selected 1×10^6 as the number of samples in the SM training data, but expect similar results from a SM trained on fewer, e.g., 1×10^5 samples, which scored a nearly identical accuracy (Berkemeier et al., 2023). The overall computational effort associated with the NC method also strongly depends on the choice of dimensions and resolution of the constraint potential map, which makes general statements regarding SM-acceleration difficult. The following numbers are derived from the exemplary application showcased in this manuscript:

Fit acquisition

To obtain 500 fits, the SM sampled 5.71×10^6 kinetic parameter sets in 86,908 s ($\sim$ 1 d) of CPU-time, KM-SUB sampled 2.53×10^6 kinetic parameter sets in $\sim$ 600 d of total CPU-time, distributed onto many CPU on a computer cluster.

Table S1. Simulation argument descriptions and selected values for this study.

Function argument	Selected value	Description
gridsize	[100, 100]	Number of individual values for $[\mathrm{X}]_{\mathrm{g},0}$ and r_p on simulated grid
rem_ground_truth	No	If parameter set selected as ground truth is removed from fit ensemble
ignore_ES_frame	1	Minimum distance of simulated experiments from edges of grid
Σ_{rad}	0.05	Simulated experimental uncertainty for r_p $[log_{10}(\mathrm{cm})]$ [1]
Σ_{X}	0.02	Simulated experimental uncertainty for $[\mathrm{O_3}]_{\mathrm{g},0}$ $[log_{10}(\mathrm{cm}^{-3})]$ [1]
Σ_{out}	0.07	Simulated experimental uncertainty for measurements $[log_{10}(\mathrm{s})]$ [1]
filter_theshold	0.0105	Remove fits from fit ensemble above this acceptance threshold (MSLE, Eq. S.8)
exp_distance	0.2	Minimal distance between two experiments on log_{10} plane for $[\mathrm{O_3}]_{\mathrm{g},0}$ and r_p
restrict_exp_duration	[1s, 3d]	Minimal and maximal KM-SUB-predicted duration of experiment to be selected
revive_fits	Yes	If removed fits of the fit ensemble are re-evaluated in later iterations

[1] Defined as standard deviation of normal distribution based on logarithmic original value from which new value is sampled randomly (Suppl. note 7).

Numerical compass method

We apply the NC with a total of 10,000 combinations of experimental parameters (100×100 grid). Each of these 10,000 ensemble solutions contain model predictions for each kinetic parameter set in the fit ensemble (here: 500). The resulting 5×10^6 model evaluations represent a major fraction of the overall computational effort in our workflow. In contrast, 10,000 evaluations of the ensemble spread or parameter constraint potential metrics (one for each ensemble solution) fall within feasibility range on a personal computer within few hours of time (when using the reduced sample density for the parameter constraint potential, Suppl. Note 3). The initial CPU time needed for the generation of 10,000 ensemble solutions is roughly a day for the SM, and more than a year for KM-SUB.

Note that the SM used in this study is trained on only 1×10^6 KM-SUB samples, in contrast to 5×10^6 model evaluations necessary to perform the NC method on a grid with the selected size. Given that the computation of the relevant data with the SM is negligible compared with the computation using KM-SUB, we achieve a speed-up by a factor of ~ 5 for the NC using the KM/SM-hybrid approach. If the SM is also applied for fit ensemble acquisition (i.e., a SM-only application), the speed-up increases to a factor of ~ 7.5.

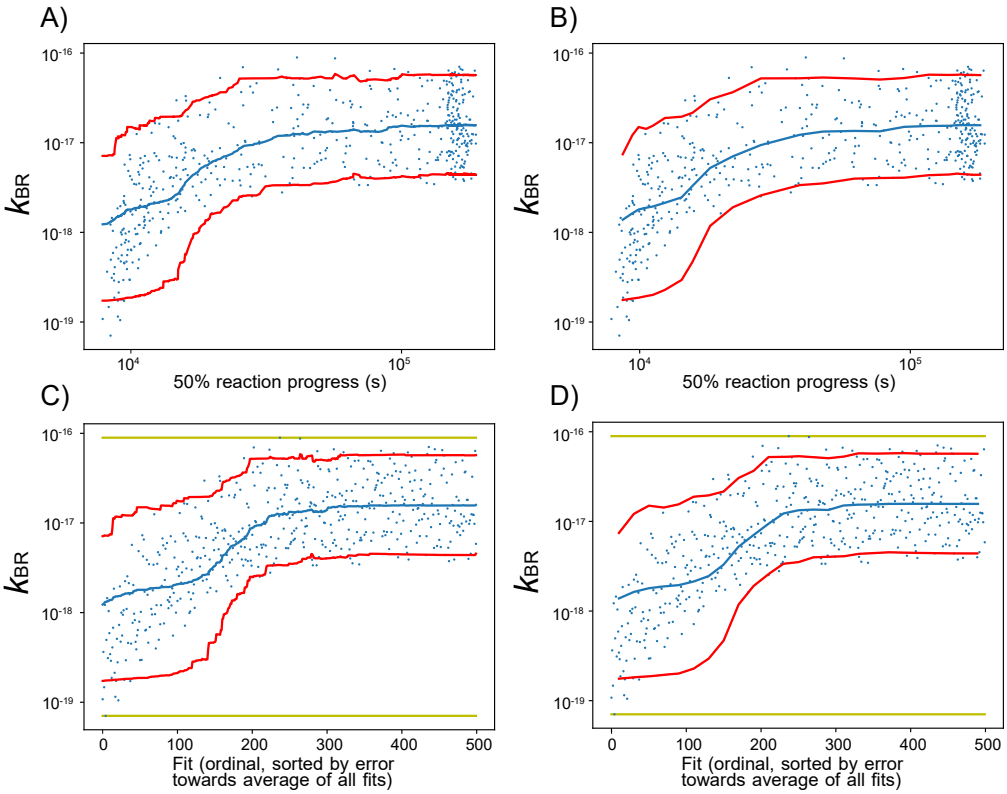

Figure S1. Visualization of the parameter constraint potential metric for the kinetic parameter k_{BR}, KM-SUB, the KM-SUB fit ensemble and the experimental parameters $r_{\mathrm{p}} = 10$ nm, $[\mathrm{O_3}]_{\mathrm{g},0} = 4.67{\times}10^{-1}$ ppb and $[\mathrm{OL}]_{\mathrm{b},0} = 1.89 \cdot 10^{21}$ cm^{-3}. Panels A and B show the distribution of k_{BR} as a function of chemical half-lives, C and D as a function of an ordinal order of fits according to the MSLE and under consideration, if the chemical half-life is smaller or larger than the one of the ensemble mean. Blue dots represent the values of k_{BR} of individual fits in the sorted fit ensemble. The blue line shows the associated subset average and the red lines the subset 5 and 95 percentiles. The yellow lines represent the absolute minimum and maximum of k_{BR} in the fit ensemble. The parameter constraint potential can be described as the area between the 5-percentile and the minimum plus the area between the 95-percentile and the maximum for the ordinal x-axis (C, D). Panel A and C are based on a fit/subset ratio of 1, panel B and D on a ratio of 20. While the computational effort (with pre-sampled model predictions) to obtain figures B and D is only 5 % in comparison to A and C, only minor differences are visible for the percentiles. Consequently, resulting parameter constraint potentials are almost identical for the two cases, 865.788 for the fit/subset ratio of 1, and 865.793 for the fit/subset ratio of 20.

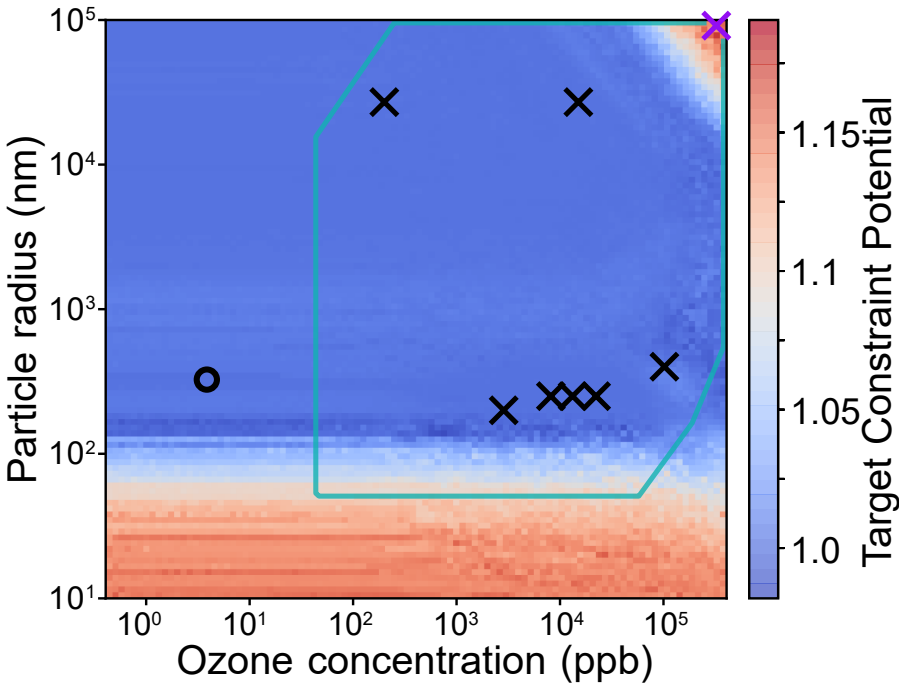

Figure S2. Constraint potential map for the target constraint potential, evaluated by the KM, based on the KM-SUB fit ensemble. The target constraint potential utilizes subsets of hypothetically accepted fits to calculate each subset's ensemble spread at the selected target condition. The target constraint potential is the average ensemble spread at the target condition of all subsets. The selected target in this case represents atmospherically relevant conditions (particle radius: $10^{2.5}$ nm; ozone concentration: 10^{11} cm^{-3}; black circle). Black crosses represent the experimental parameters of the seven real experiments that are used for the initial acquisition of the fit ensemble. The purple cross represents the ensemble spread maximum with satisfied experimental constraint conditions. In basic tests for the oleic acid ozonolysis system, we observed that target constraint potential maps show high similarities to ensemble spread maps if the number of fits in the ensemble is large. Note that this figure was made using the *KineticCompass* module for the Julia programming language.

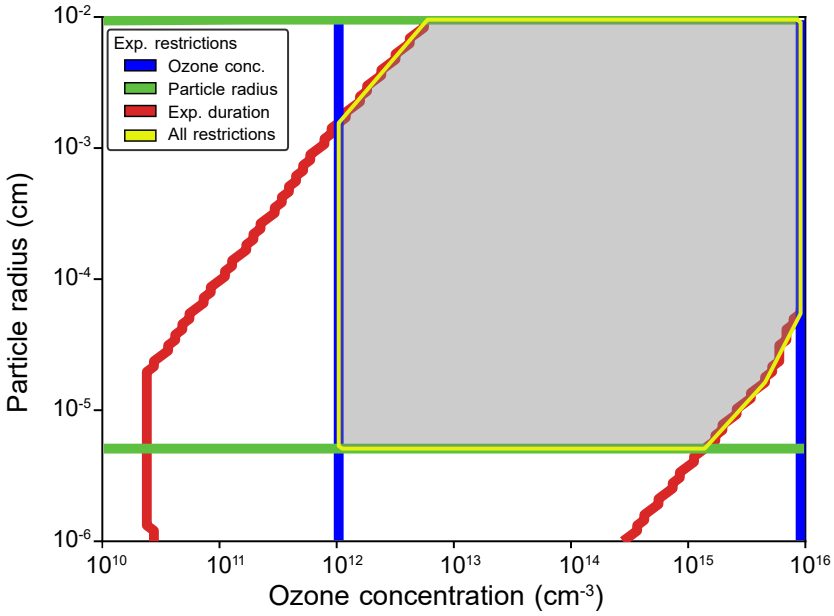

Figure S3. Restrictions for constraint potential maps with regards to experimental feasibility in this study. Blue and green lines show the boundary conditions for the experimental parameters ozone concentration and particle radius, respectively. The red lines frame combinations of parameters where KM-SUB predictions based on all fits in the KM-SUB fit ensemble fall within the required experiment duration (1s < exp_dur < 3d). The yellow box with gray filling shows the area where experiments are accepted for simulation, if proposed by the NC.

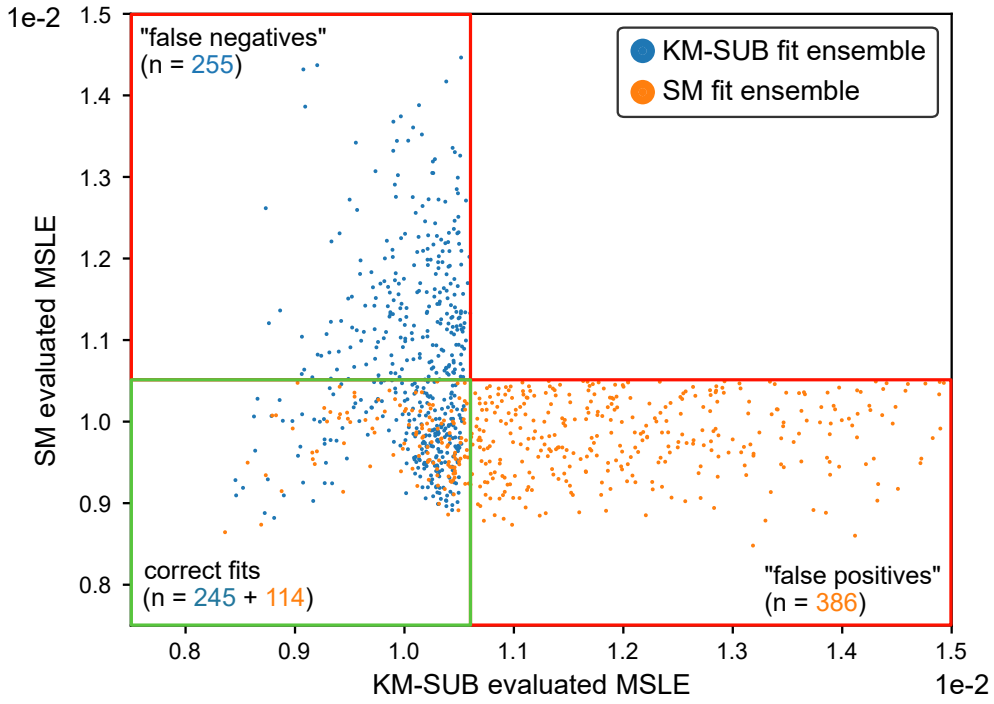

Figure S4. Contrariwise cross-evaluation of the KM-SUB fit ensemble (blue) and the neural network surrogate model (SM) fit ensemble (orange) with regards to mean squared (absolute) logarithmic error (MSLE) in comparison of model outputs with the seven experimental data sets used for initial fit ensemble acquisition. False negative fits in the top left rectangle are KM-SUB fits that are not recognized as fits by the SM. False positive fits in the bottom right rectangle are, in contrary, SM suggested fits with KM-SUB predictions that exceed the associated acceptance threshold $\theta = 0.0105$. In contrast to false positives, false negatives can not be eliminated by re-sampling of the fit ensemble with the KM, and represent a general uncertainty in SM applications (Berkemeier et al., 2023).

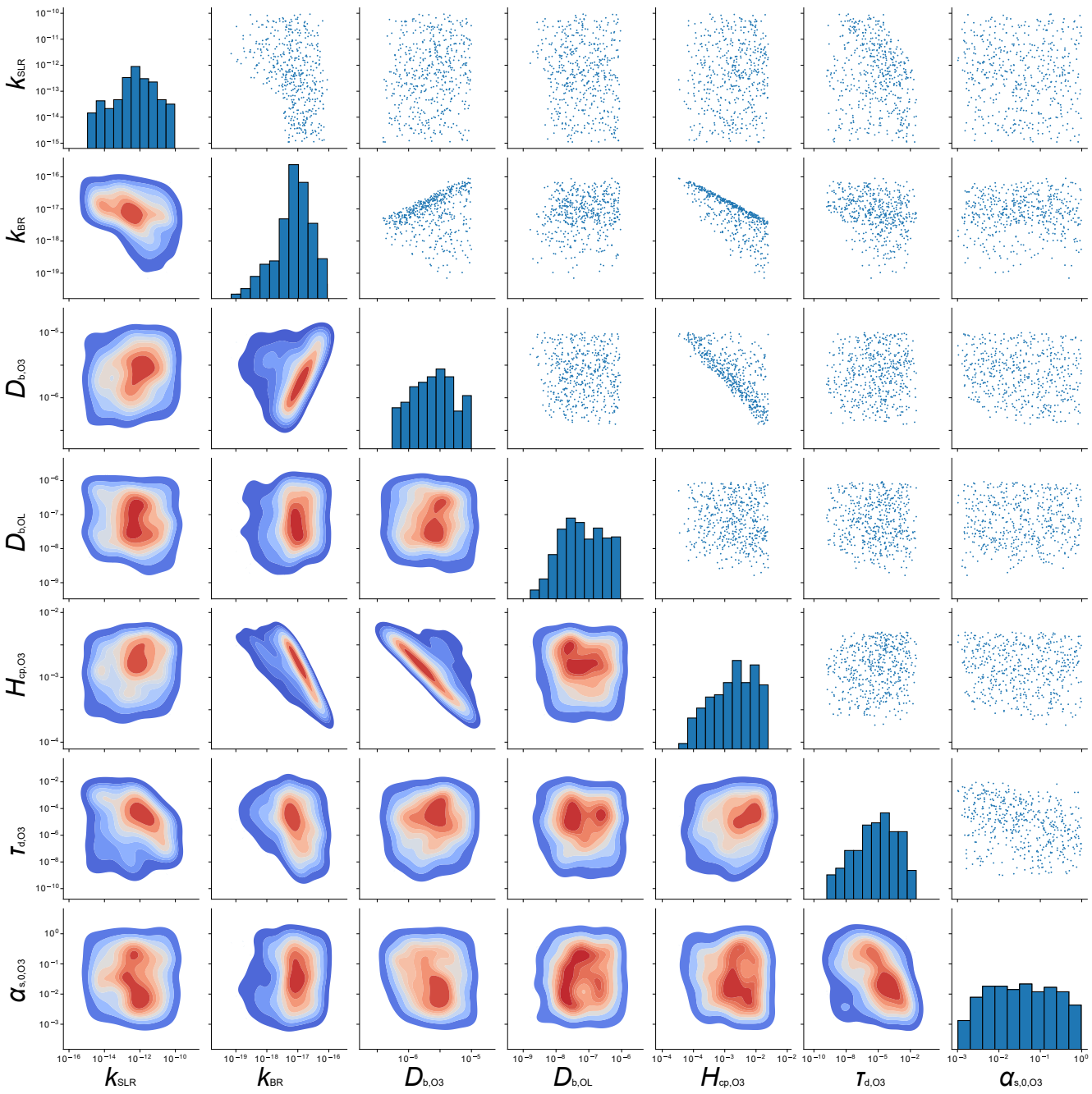

Figure S5. Scatter plot matrix of the KM-SUB fit ensemble (n = 500) with an acceptance threshold θ of 0.0105. The diagonal elements are histograms showing the distributions of the seven kinetic input parameters. The off-diagonal elements are scatter plots (top right) or densities (bottom left) of all combinations of two parameters occurring in the KM-SUB fit ensemble.

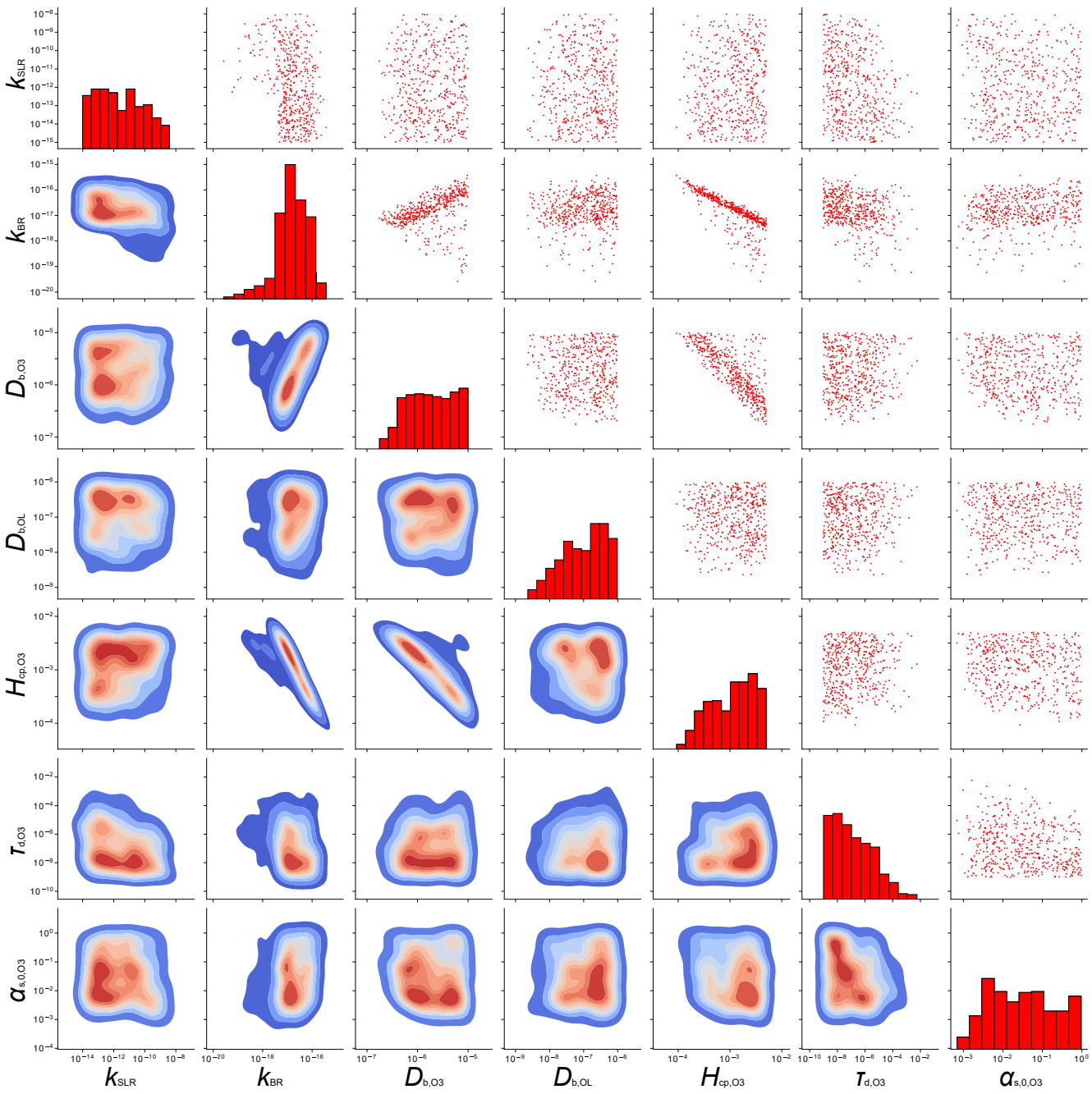

Figure S6. Scatter plot matrix of the SM fit ensemble (n = 500) with an acceptance threshold θ of 0.0105. The diagonal elements are histograms showing the distributions of the seven kinetic input parameters of KM-SUB. The off-diagonal elements are scatter plots (top right) or densities (bottom left) of all combinations of two parameters occurring in the SM fit ensemble.

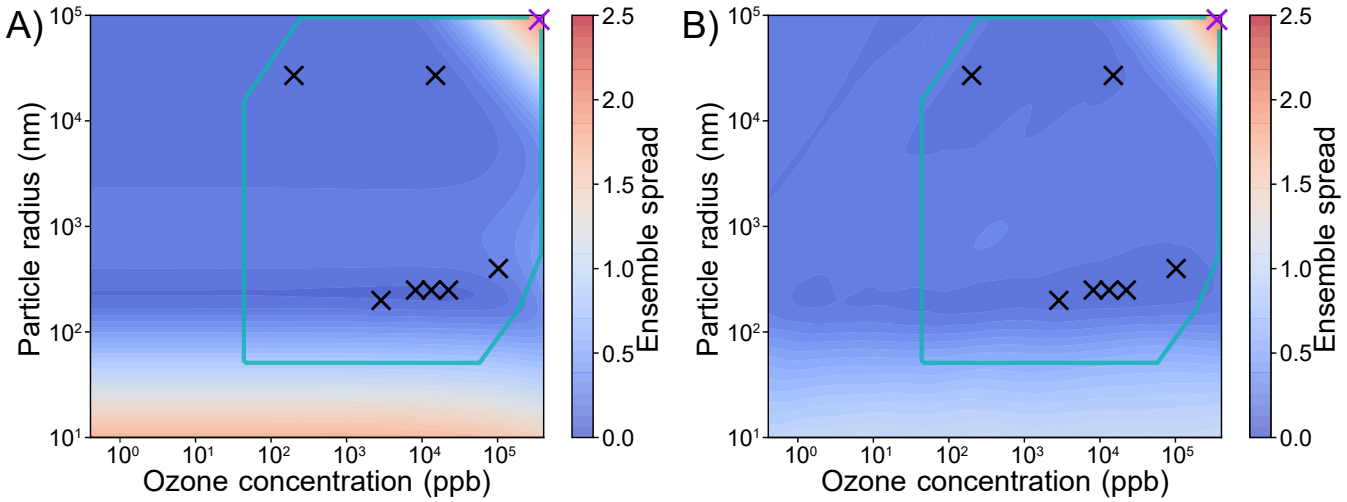

Figure S7. Constraint potential maps for the ensemble spread, evaluated by (A) KM-SUB and (B) SM, based on the KM-SUB fit ensemble and SM fit ensemble, respectively. The teal box frames the area of experimentally accessible conditions with regards to particle radius, ozone concentration and predicted experiment duration (Suppl. Note 4). Black crosses represent the experimental parameters of the seven real experiments that are used for the initial acquisition of the fit ensemble. Purple crosses represent the ensemble spread maximum in each grid with satisfied experimental constraint conditions.

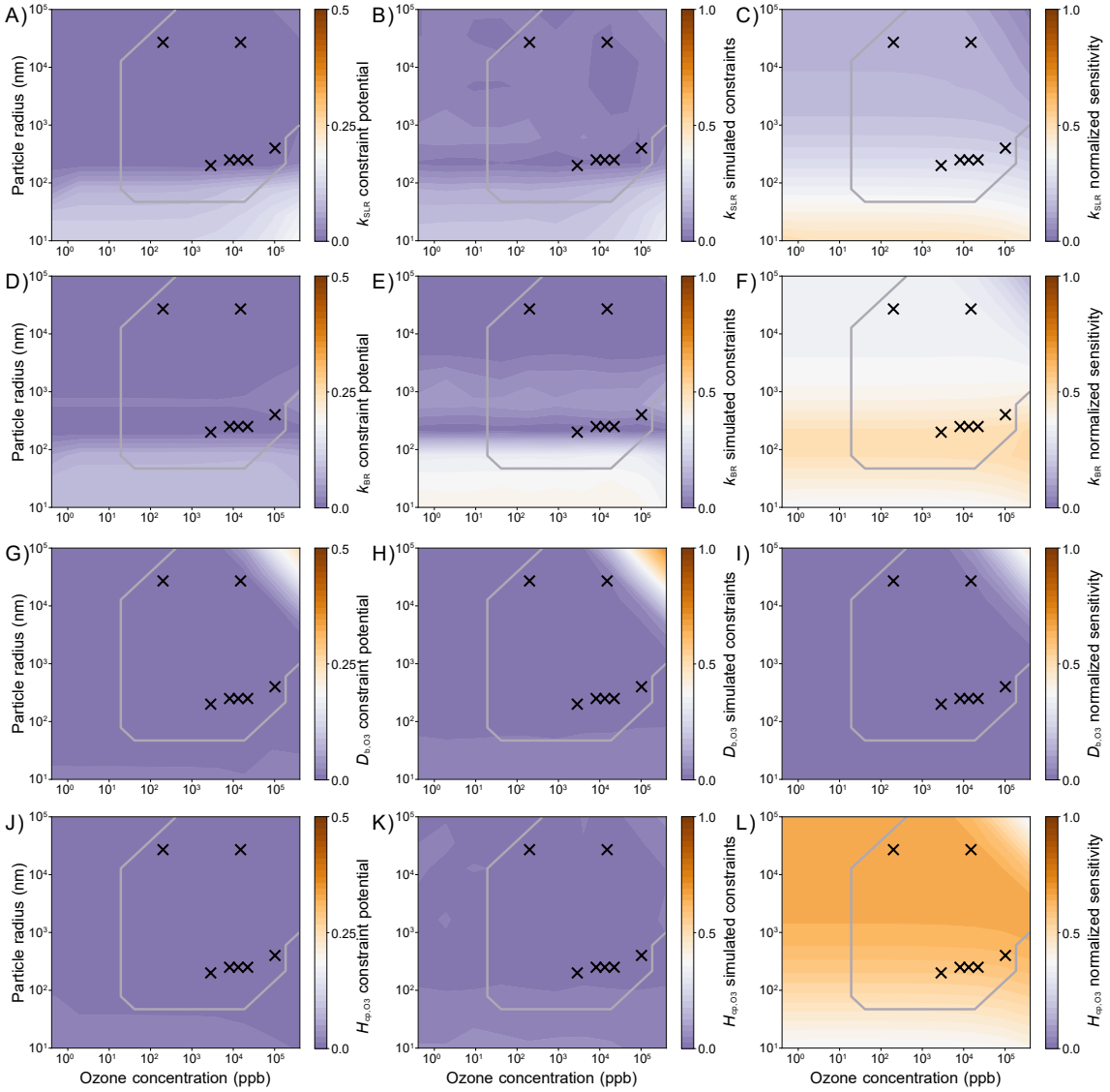

Figure S8. Panels in columns display maps for different methods to select experiments based on approximated constraints for individual parameters: Normalized parameter constraint potential (A, D, G, J), average constraints of 5-95-percentile ranges of parameters in the fit ensemble for brute-force simulation across all fits as simulated truths (B, E, H, K), and normalized partial sensitivities (Eq. S.12) in the KM-SUB fit ensemble (C, F, I, L). Panels in rows show these maps for four kinetic parameters k_{SLR} (A, B, C), k_{BR} (D, E, F), $D_{b,OL}$ (G, H, I), $H_{cp,O3}$ (J, K, L). The gray boxes frame the area of experimentally accessible conditions with regard to particle radius and ozone concentration (Suppl. Note 4). Black crosses represent the experimental parameter sets of the seven real experiments that are used for the initial acquisition of the fit ensemble. The calculations are performed on a reduced 10×10 grid of experimental conditions.

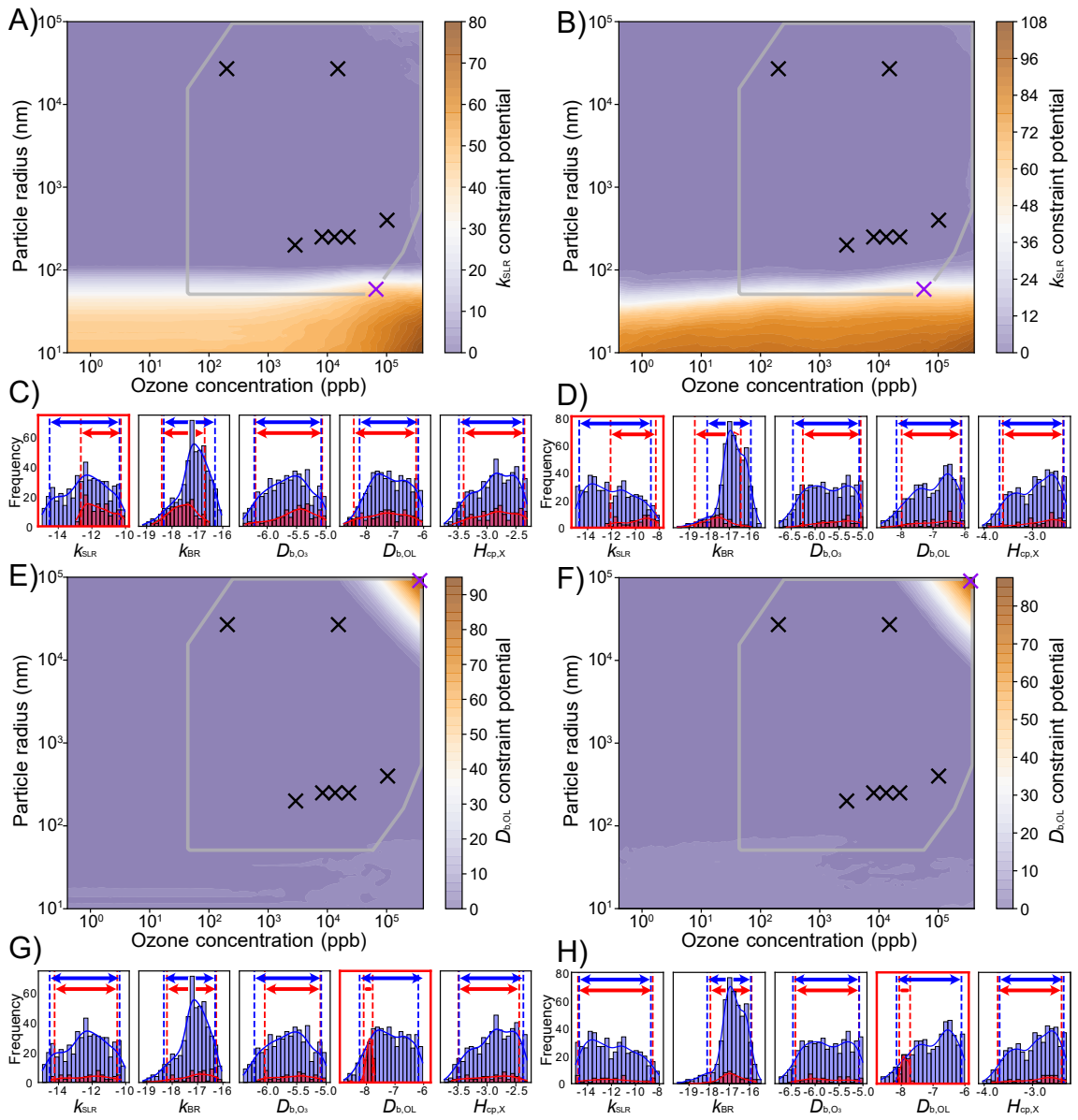

Figure S9. Constraint potential maps for the kinetic parameters k_{SLR} (A, B) and $D_{b,OL}$ (E, F). In (A) and (E), the model KM-SUB is used (KM-only approach), while in (B) and (F), the SM is employed (SM-only approach). The gray box frames the area of experimentally accessible conditions with regard to particle radius and ozone concentration (Suppl. Note 4). Black crosses represent the experimental parameter sets of the seven real experiments that are used for the initial acquisition of the fit ensemble. The purple cross represents the parameter constraint potential maximum with satisfied experimental constraint conditions. The suggested experimental conditions are used to obtain synthetic experimental data by evaluating KM-SUB for the best fit in the KM-SUB fit ensemble. Frequency distributions are shown for individual kinetic parameters in the KM-SUB fit ensemble (C, G) and SM fit ensemble (D, F), before (blue) and after (red) fit filtering with acceptance threshold $\theta = 0.0105$. Blue and red dotted lines and arrows visualize the 5-95 percentile range of each distribution.

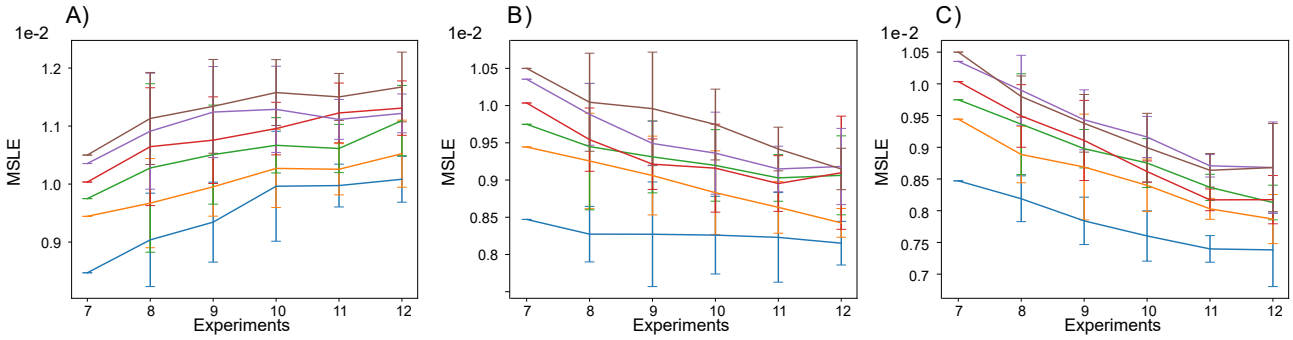

Figure S10. Visualization of the uncertainty calibration method for three sets of uncertainty parameters (Panel A: Σ_{rad} = 0.06; Σ_X = 0.02; Σ_{out} = 0.1; panel B: Σ_{rad} = 0.05; Σ_X = 0.02; Σ_{out} = 0.07; panel C: Σ_{rad} = 0.05; Σ_X = 0.03; Σ_{out} = 0.06) and six selected sets of kinetic parameters that resemble the original $\delta_{exp,l}$ distribution in the fit ensemble. The first data point (7) represents the original $\delta_{exp,l}$ for the seven real experiments of each selected kinetic parameter set (Eq. S.9). In the following, mean $\delta_{l,v}$ errors as well as their standard deviations (error bars) for 20 repetitions and five iterations of the uncertainty calibration are displayed (Eq. S.11). While panels A) and C) show "runaway errors", panel B represents the desired slight error decrease for the best fit and large decreases for fits with a larger $\delta_{exp,l}$. We select these values of Σ_{unc} for the entire study.

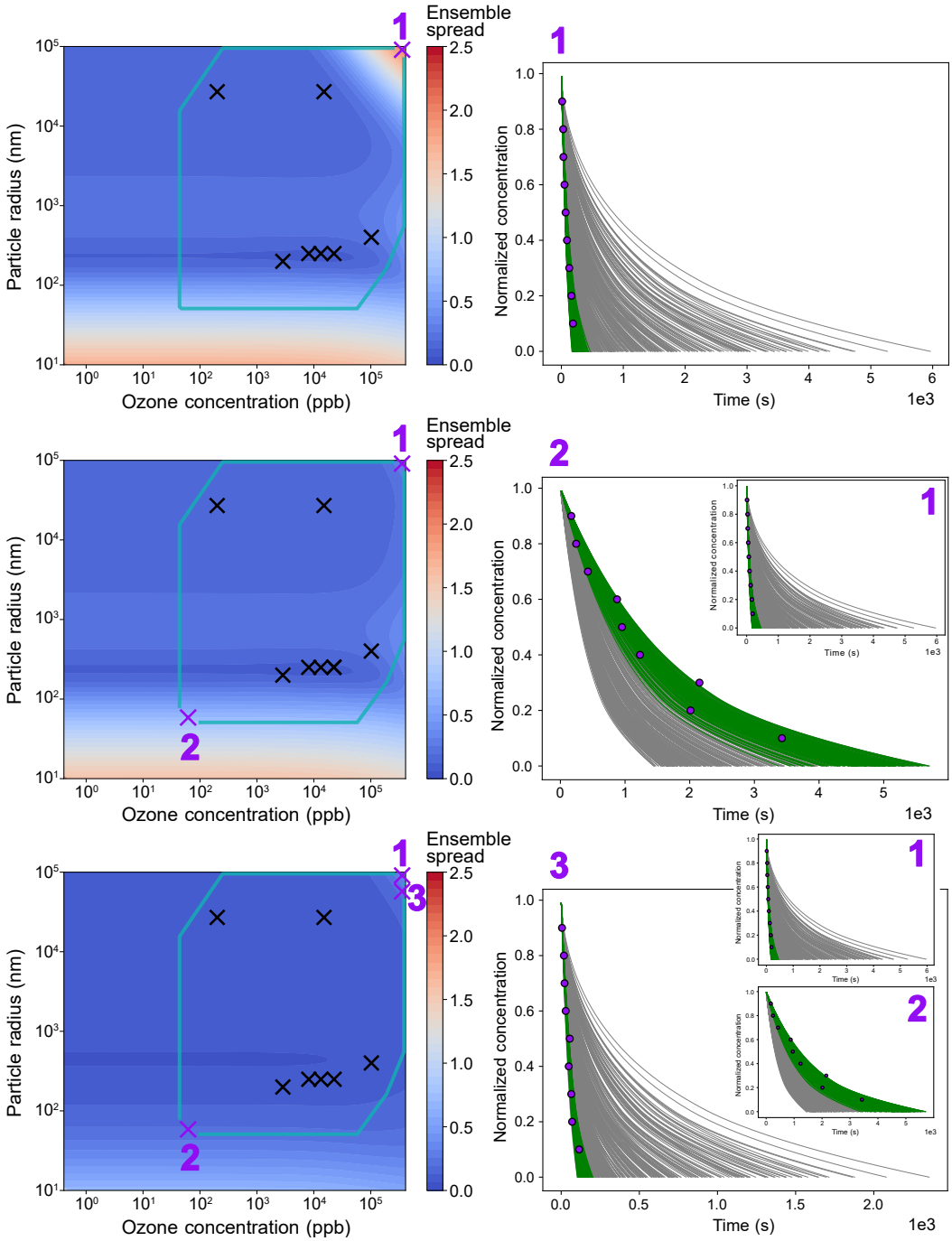

Figure S11. Three iterations of an example simulation for the NC, evaluating the ensemble spread metric with KM-SUB from the KM-SUB fit ensemble. On the left, constraint potential maps for each of three iterations are shown. Plots on the right show ensemble solutions for the selected experimental parameters with the simulated experiment (purple markers), accepted fits (green) and rejected fits (gray) at each iteration. The parameter set selected as simulated truth is the same as in Fig. S13 for the KM/SM-hybrid application. The number of accepted fits in each iteration is 217, 136 and 137, and thus comparably high.

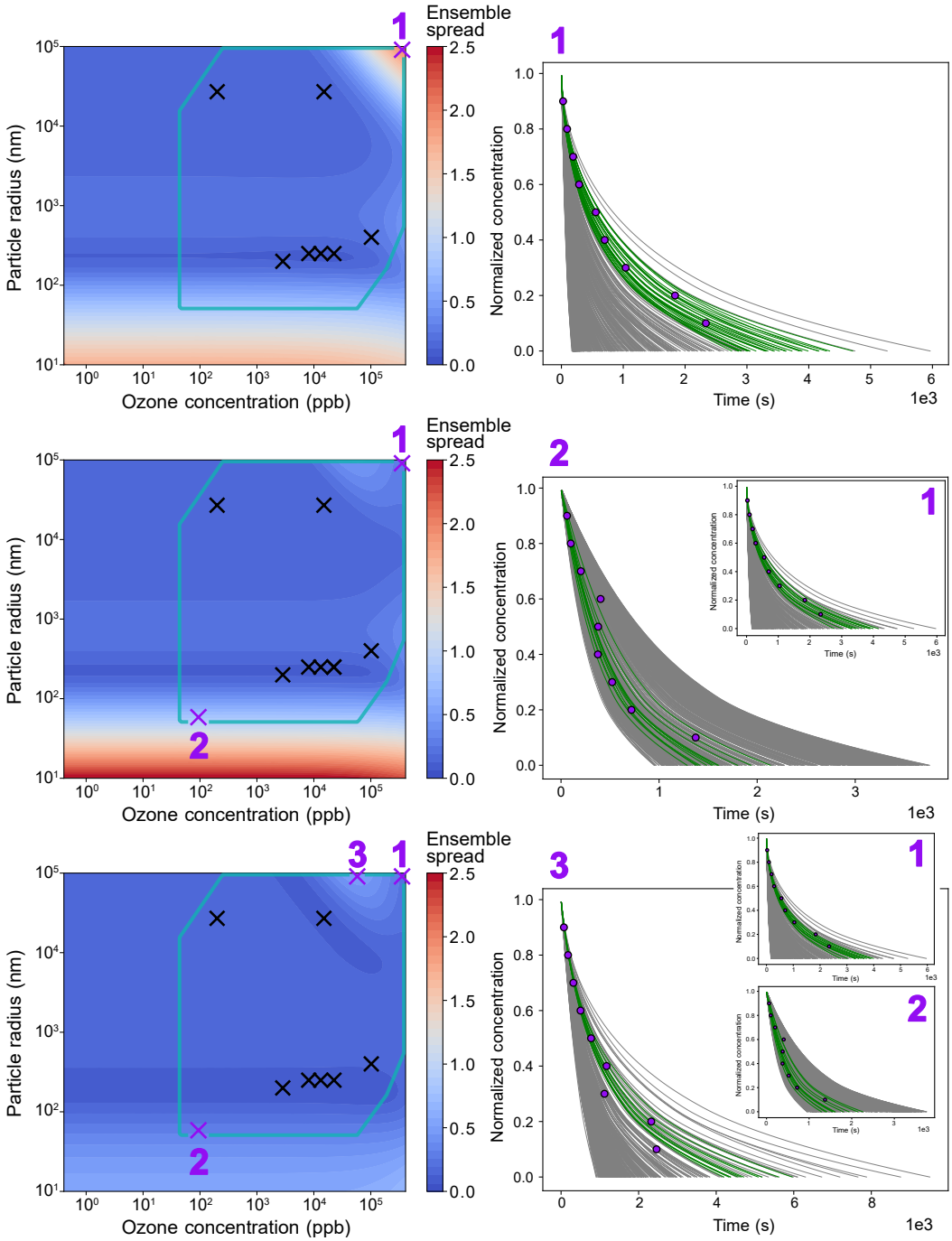

Figure S12. Three iterations of an example simulation for the NC, evaluating the ensemble spread metric with KM-SUB from the KM-SUB fit ensemble. On the left, constraint potential maps for each of three iterations are shown. Plots on the right show ensemble solutions for the selected experimental parameters with the simulated experiment (purple markers), accepted fits (green) and rejected fits (gray) at each iteration. The parameter set selected as simulated truth is the same as in Fig. S14 for the KM/SM-hybrid application. The number of accepted fits in each iteration is 31, 12 and 11, and thus comparably low.

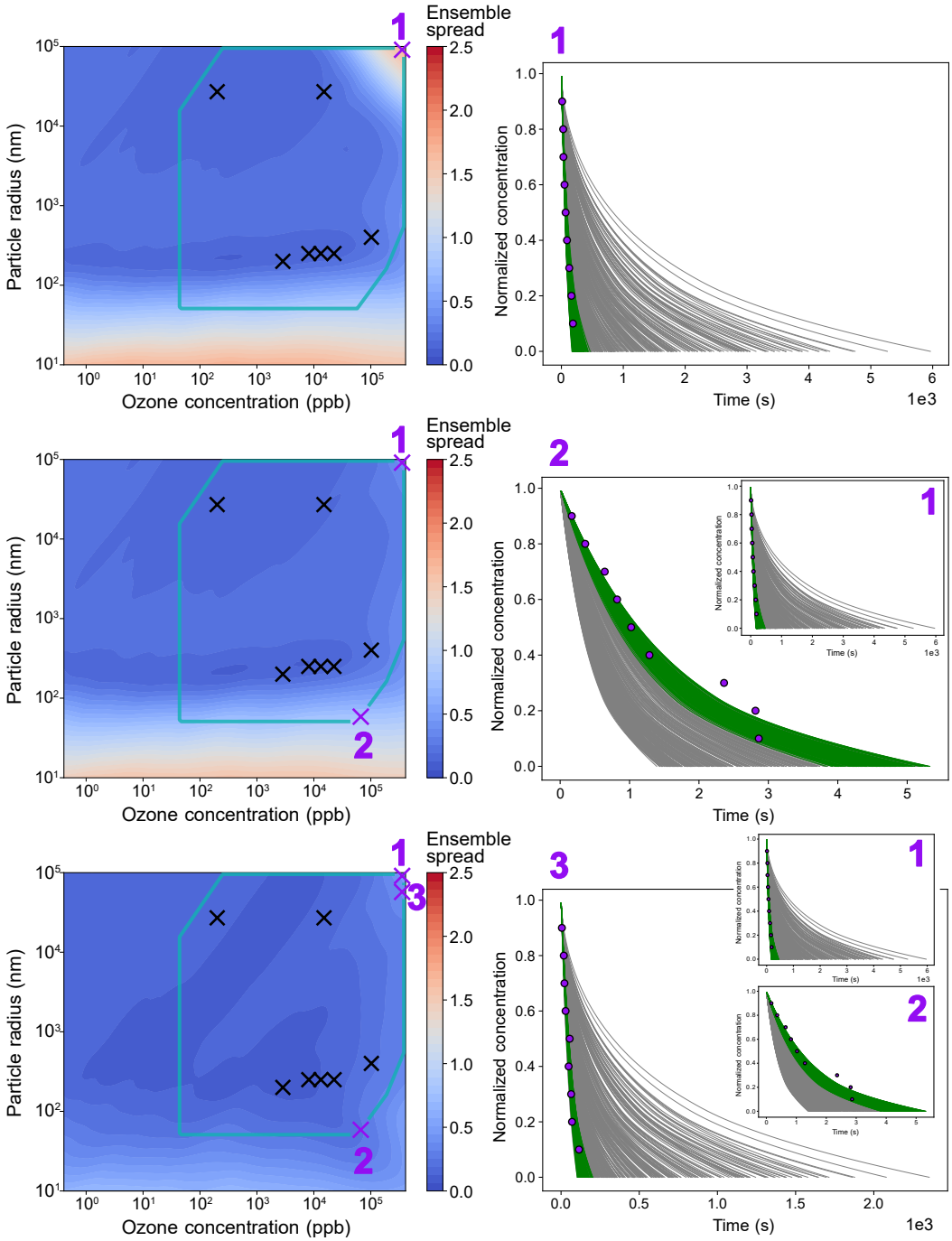

Figure S13. Three iterations of an example simulation for the NC, evaluating the ensemble spread metric with the SM from the KM-SUB fit ensemble. On the left, constraint potential maps for each of three iterations are shown. Plots on the right show ensemble solutions for the selected experimental parameters with the simulated experiment (purple markers), accepted fits (green) and rejected fits (gray) at each iteration. The number of accepted fits in each iteration is 217, 136 and 133, and thus comparably high.

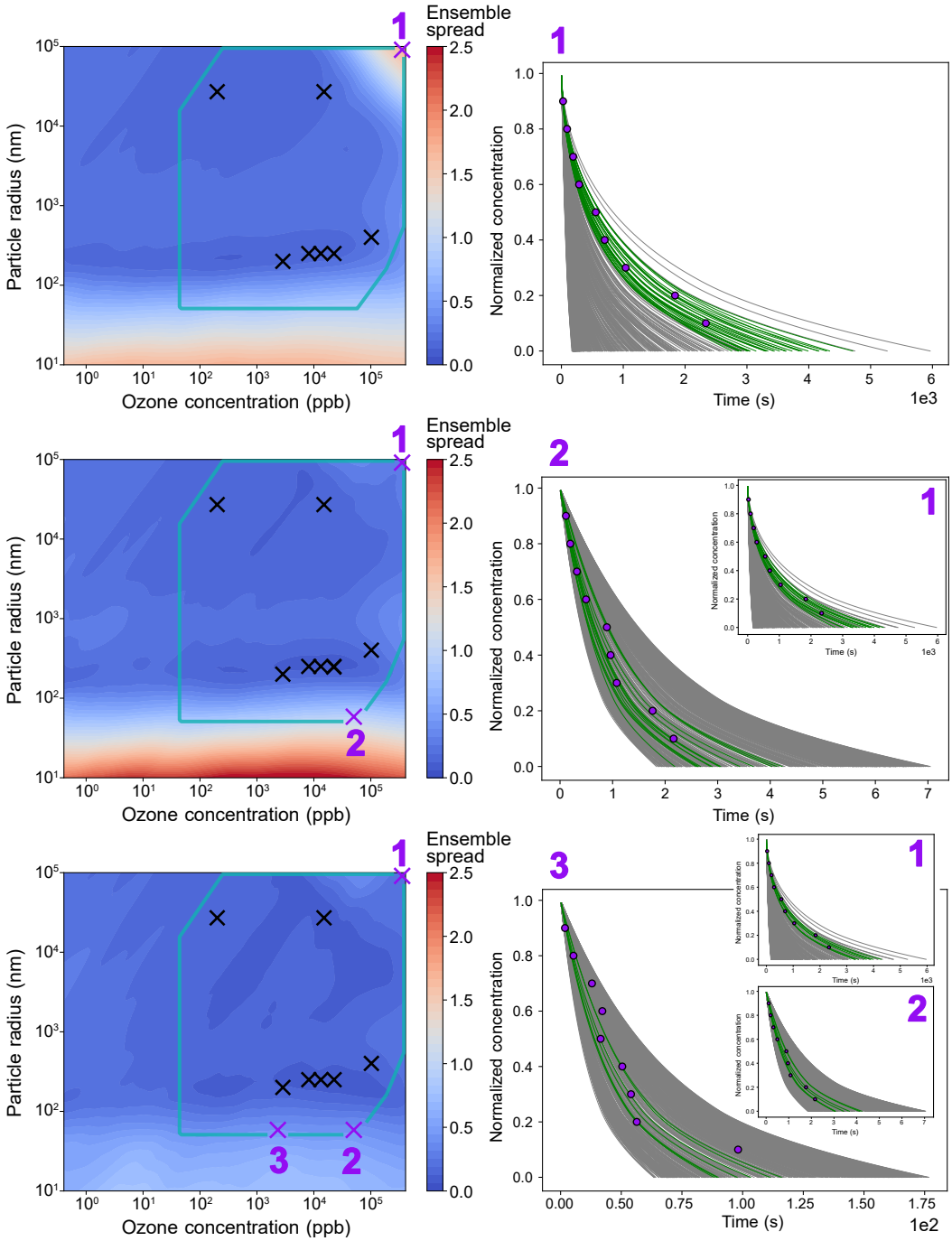

Figure S14. Three iterations of an example simulation for the NC, evaluating the ensemble spread metric with SM from the KM-SUB fit ensemble. On the left, constraint potential maps for each of three iterations are shown. Plots on the right show ensemble solutions for the selected experimental parameters with the simulated experiment (purple markers), accepted fits (green) and rejected fits (gray) at each iteration. The number of accepted fits in each iteration is 31, 15 and 7, and thus comparably low.

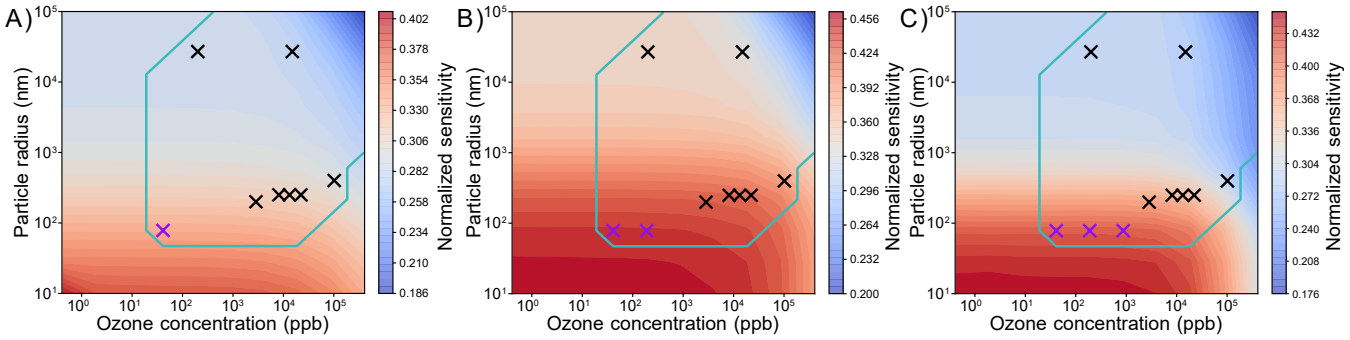

Figure S15. Maps of total sensitivity (Eq. S.13) applied to the fit ensemble during three iterations of an example simulation for the NC, evaluating the normalized parameter sensitivities of KM-SUB using the KM-SUB fit ensemble on a 10×10 grid of experimental conditions. The teal boxes frame the area of experimentally accessible conditions with regards to particle radius, ozone concentration and predicted experiment duration (Suppl. Note 4). Black crosses represent the experimental parameters of the seven real experiments that are used for the initial acquisition of the fit ensemble. The purple crosses represent simulated experiments at the sensitivity maxima with satisfied experimental constraint conditions. The experimental conditions (up to three in panel C) are selected successively in each repetition of the simulation and independent of the simulated truth, synthetic experimental outcome and resulting constraint on the fit ensemble.

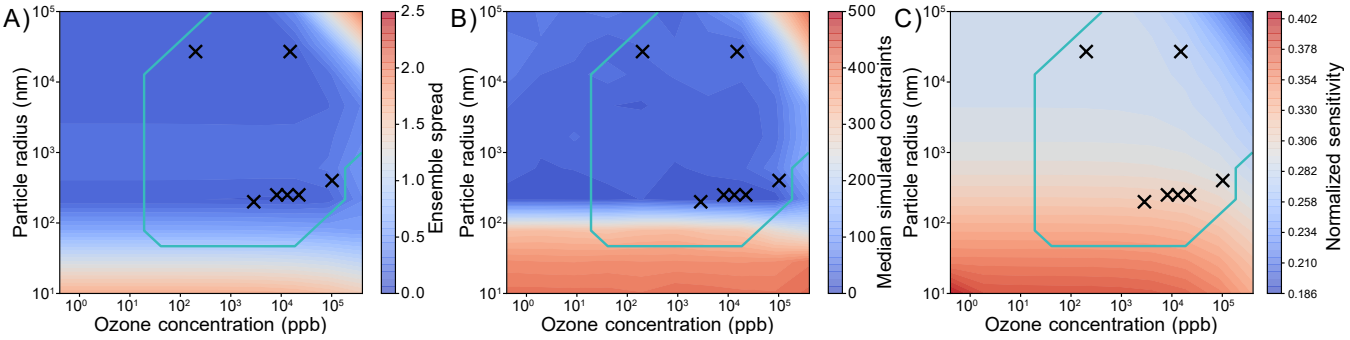

Figure S16. Ensemble spread (KM-SUB; panel A), median brute-force simulated constraints (panel B) and total sensitivities (KM-SUB; panel C) for the KM-SUB fit ensemble in a 10×10 grid of experimental conditions. For the brute force simulation, each fit in the fit ensemble is selected as simulated truth, and the median numbers of fits rejected from the fit ensemble are plotted on the map of experimental conditions in a similar fashion than the constraint potential maps. The map represents the median constraints that can be achieved for the given fit ensemble and assumptions made in the simulation.

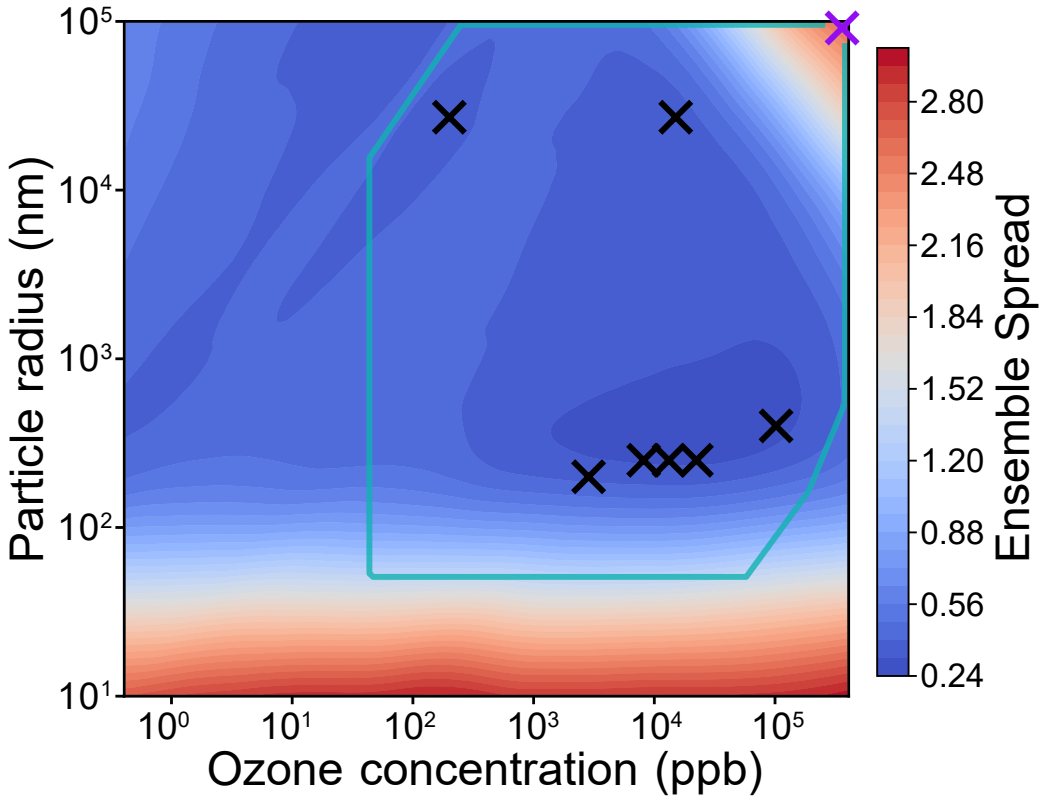

Figure S17. Constraint potential map for the ensemble spread, evaluated by the SM, based on a newly acquired SM fit ensemble with an acceptance threshold of 0.021, a factor two larger than the one used elsewhere in this study. The teal box frames the area of experimentally accessible conditions with regards to particle radius, ozone concentration and predicted experiment duration (Suppl. Note 4). Black crosses represent the experimental parameters of the seven real experiments that are used for the initial acquisition of the fit ensemble. The purple cross represents the ensemble spread maximum with satisfied experimental constraint conditions.

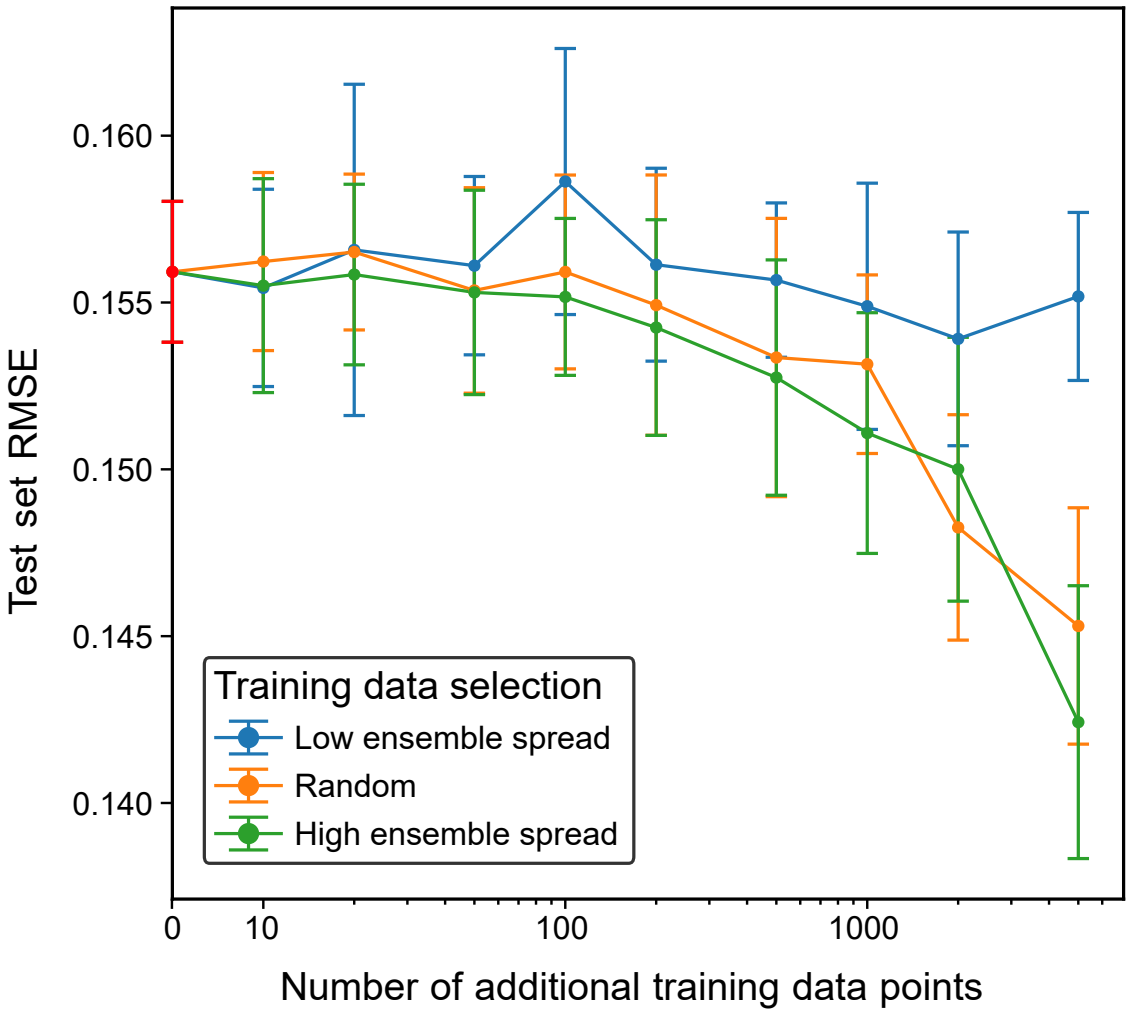

Figure S18. Effect of ensemble spread in additional training data on the QSAR model accuracy of a newly trained model. We train models that parameterize the reduction potential of quinones based on SMILES strings representing molecular structure (Krüger et al., 2022) on a subset of the Tabor_nosulf data (10,000 quinones) (Tabor et al., 2019), using 10-fold cross-validation, and select 1000 quinones as independent test set for all compared models. For the remaining 58,599 quinones in the original data, we determine the ensemble spread using the ensemble predictions of the models from the first step. New models with identical hyper-parameters are then trained on data sets comprised of the original 10,000 quinones plus an additional 10, 20, 50, 100, 200, 500, 1000, 2000 or 5000 quinones from the remaining data. We compare the selection of quinones by largest ensemble spread (green) and lowest ensemble spread (blue) with three different random samples (orange). Markers show mean test set RMSE of the 10 cross-validation models (3×10 for the random selection) in each run, error bars the standard deviation. By addition of molecules with a high ensemble spread, the strongest improvement of the newly trained QSAR models is achieved for almost all data set sizes. However, only a slight difference is observed between the random and high ensemble spread selection.

References

Berkemeier, T., Krüger, M., Feinberg, A., Müller, M., Pöschl, U., and Krieger, U. K.: Accelerating models for multiphase chemical kinetics through machine learning with polynomial chaos expansion and neural networks, Geosci. Model Dev., 16, 2037–2054, https://doi.org/10.5194/gmd-16-2037-2023, 2023.

Chollet, F. et al.: Keras, https://keras.io, 2015.

Gallimore, P. J., Griffiths, P. T., Pope, F. D., Reid, J. P., and Kalberer, M.: Comprehensive modeling study of ozonolysis of oleic acid aerosol based on real-time, online measurements of aerosol composition: Organic Aerosol Model and Measurements, J. Geophys. Res. Atmos., 122, 4364–4377, https://doi.org/10.1002/2016JD026221, 2017.

Hearn, J. D. and Smith, G. D.: Kinetics and Product Studies for Ozonolysis Reactions of Organic Particles Using Aerosol CIMS, J. Phys. Chem. A, 108, 10 019–10 029, https://doi.org/10.1021/jp0404145, 2004.

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B. A. M., Lammel, G., Schmidt, B., Pöschl, U., and Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, Nat. Sci., 2, e20220 016, https://doi.org/10.1002/ntls.20220016, publisher: John Wiley & Sons, Ltd, 2022.

Müller, M., Mishra, A., Berkemeier, T., Hausammann, E., Peter, T., and Krieger, U. K.: Electrodynamic balance–mass spectrometry reveals impact of oxidant concentration on product composition in the ozonolysis of oleic acid, Phys. Chem. Chem. Phys., 24, 27 086–27 104, https://doi.org/10.1039/D2CP03289A, 2022.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., and Aspuru-Guzik, A.: Mapping the frontiers of quinone stability in aqueous media: Implications for organic aqueous redox flow batteries, J. Mater. Chem. A, 7, 12 833–12 841, https://doi.org/10.1039/c9ta03219c, 2019.

Ziemann, P. J.: Aerosol products, mechanisms, and kinetics of heterogeneous reactions of ozone with oleic acid in pure and mixed particles, Faraday Discuss., 130, 469, https://doi.org/10.1039/b417502f, 2005.

Supplementary Information for: Towards an annual carbon balance of biological soil crusts: parametric equations and neural networks to model gas exchange and net primary productivity

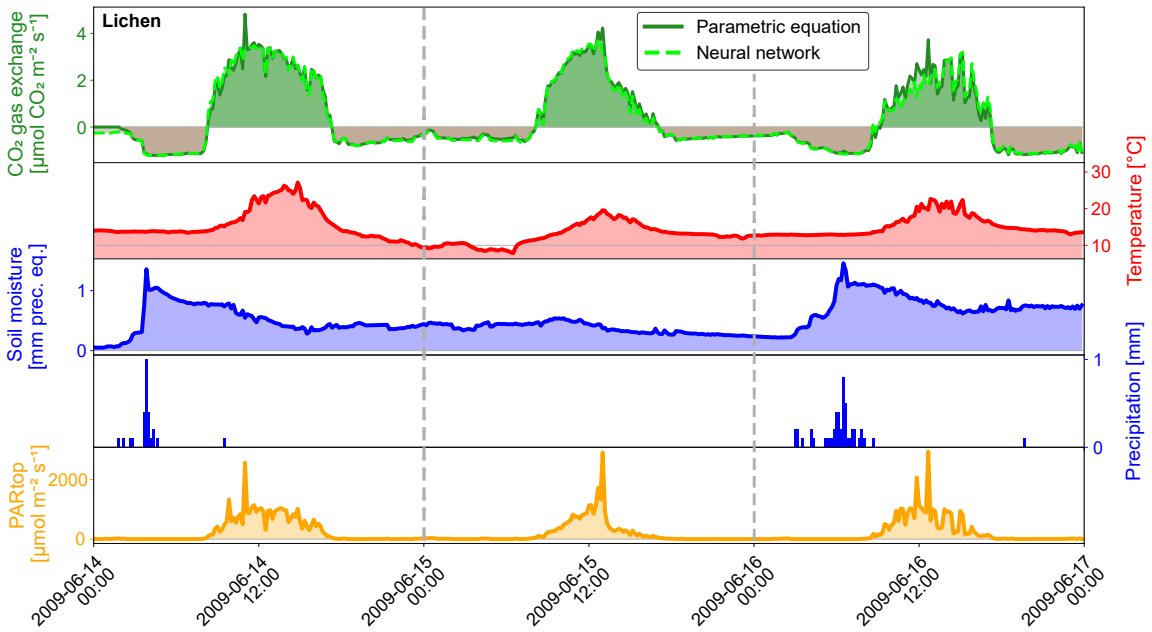

Figure 1: *Application of the parametric equation and neural network models for lichen on micrometeorological data from Soebatsfontein, Succulent Karoo, South Africa over 3 days from June 14 to June 16, 2009. Red, yellow and blue lines show temperature, light intensity and soil moisture profiles, blue bars represent precipitation, the green lines display CO_2 gas exchange of lichen biocrusts predicted by the respective models. Brown areas in the top panel indicate periods where respiration surpasses photosynthetic activity, while green areas show times of net photosynthetic productivity.*