Computational Methods to Quantify Nanoparticle Interactions

by

Jacob Charles Saldinger

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemical Engineering)
in The University of Michigan
2023

Doctoral Committee:

Professor Angela Violi, Chair
Professor André L. Boehman
Associate Research Scientist Paolo Elvati
Professor Nicholas A. Kotov
Professor Robert M. Ziff

Jacob C. Saldinger
jsald@umich.edu
ORCID iD: 0000-0001-5005-614X

Dedicated to my family who have shaped who I am today.

ACKNOWLEDGEMENTS

This dissertation is not only a product of my research at the University of Michigan but also speaks to the incredible amount of support I have had along my journey.

First, I would first like to acknowledge my advisor, Dr. Angela Violi, who has provided immeasurable support and allowed me to grow as a researcher at every point throughout my Ph.D. journey. I would also like to acknowledge Dr. Paolo Elvati for his valuable discussions and helping me gain the skills to work through the most challenging parts of my thesis work. I am fortunate to have been a member of the Violi lab and would like to acknowledge all the members for their personal support and technical expertise.

Furthermore, I would like to thank the committee members Dr. Nicholas Kotov, Dr. Andre Boehman, and Dr. Robert Ziff for their time, insights, and participation as members of my thesis committee. Their guidance and mentorship has shaped this thesis work.

These past five years would not have been possible without my friends in Michigan who have laughed and cried with me through all the ups and downs of this journey.

Finally I would like to acknowledge my family. My mom Elizabeth, dad Eric, step-mother Heather, brother Samuel, sister Elisheva, brother-in-law Osher, and many more extended family members. I am fortunate to have had such a loving and positive support network throughout my thesis and I am grateful they are in my life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF ABBREVIATIONS

AUC Area Under the Curve

CFD Computational Fluid Dynamics

CG Coarse-Grained

COM Center of Mass

CPSA Charged Partial Surface Area

DBD Docking Benchmark Database

FE Free Energy

FES Free Energy Surface

GQD Graphene Quantum Dot

HAB Height Above Burner

H/C Hydrogen-Carbon Ratio

kMC kinetic Monte Carlo

LC Length of Aliphatic Chain

LOOCV Leave-One-Out Cross Validation

LYS Lysine

MAE Mean Absolute Error

MAXPC Maximum partial Charge

MD Molecular Dynamics

MINPC Minimum partial Charge

ML Machine Learning

MW Molecular Weight

N5MR Number of Five-Membered Rings

N6MR Number of Six-Membered Rings

NeCLAS Neural Coarse-graining with Location Agnostic Sets

NN Neural Network

NOH Number of OH Groups

NR Number of Radical Sites

NRB Number of Rotatable Bonds

numC Number of Carbons

O/C Oxygen-Carbon Ratio

PAC Polycyclic Aromatic Compound

PDB Protein Data Bank

PNI Protein-Nanoparticle Interaction

PPI Protein-Protein Interaction

RFE Recursive Feature Elimination

RMSD Root Mean Squared Deviation

RMSE Root Mean Squared Error

SASA Solvent Accessible Surface Area

SCOP Structural Classification of Proteins Database

SNAPS2 Stochastic Nanoparticle Simulator 2

SVR Support Vector Regression

TPSA Total Positive Surface Area

VSA Van der Waal's Surface Area

WHIM Weighted Holistic Invariant Molecular Descriptor

WTMD Well-Tempered Metadynamics

ABSTRACT

Nanoparticles have emerged as a promising class of materials with potential applications in a variety of diverse fields. Central to the study of nanoparticles is understanding how these structures interact with their surrounding environment as these interactions can govern both how nanoparticles form and their specific functions. Computational methods show promise as a means to characterize these interactions on an atomic level as well as efficiently quantify a large number of potential nano-interactions. Still, many challenges exist in applying computational methods to the study of nanoscale interactions including the lack of available datasets, complex nanoscale chemistry, and heterogeneous nanoparticle environments. In this thesis, I show how multiple computational methods can be applied together to most effectively overcome these challenges and quantify nanoparticle interactions while maintaining a high level of chemical interpretability. Atomistic simulation provides a physically grounded means to produce nanoparticle interaction data, numerical descriptors provide a chemically relevant method to interpret atomistic simulations, while machine learning offers a computationally efficient tool to relate chemical descriptors to complex nanoscale interactions. I apply these methods to answer two broad questions: First, how do nanoscale interactions drive nanoparticle growth and the resulting properties of these nanoparticles. Second, how do the properties and chemistries of nanoparticles contribute to their function through the interactions in which they participate. In the first application, I focus on the chemical interactions leading to the formation of polycyclic aromatic compounds (PACs), a key class of structures in the creation of soot aggregates and the synthesis of gas-phase carbon nanoparticles. I show how kinetic Monte Carlo simulations of these interactions can reproduce

the diverse PAC chemical space in complex flame environments, while numerical descriptors and machine learning can help us better understand these processes. In the second application, I demonstrate how computational techniques can explain the physical interactions of these PAC nanostructures that lead to their aggregation into larger nanoparticles. Finally, I introduce a versatile nanoscale interaction prediction tool that uses machine learning to accurately predict interaction sites between nanostructures, showing how it can help understand the interactions and functions of liquid-phase biological nanoparticles. The wide variety of nanoparticle systems studied in this work underscores that these computational methods are not confined to a single class of nanostructure or type of interaction but rather provide a robust framework that can be applied to computationally quantify nanoparticle interactions in a diverse range of applications.

CHAPTER I

Introduction

1.1 Motivation

Nanoparticles describe a large class of structures which exist typically on scales between 1 nm and 100 nm. These molecules have a variety of unique chemistries due to their size-dependent properties, layered core-shell effects, complex shapes and morphologies, high surface areas, and diverse functionalizations[1–4]. These properties drive a range of diverse interactions which give nanoparticles a number of promising applications in areas such as manufacturing, electronics, pharmaceuticals, and energy[5].

The effectiveness of nanoparticles in these applications often depend on the nanoparticles' chemical features which drive their propensity to participate in very specific interactions. For example, nanoparticle drugs can be designed to target biological nanostrucutres such as proteins[6] while catalytic nanoparticles can exhibit selectivity towards certain reactants[7]. Furthermore, reliable synthesis of nanoparticles requires tailoring the interactions of nascent nanostructures with surrounding molecules in reactor systems[8,9]. Thus in order to properly design, synthesize, and optimize nanoparticles for specific applications, it is imperative to understand how the chemical features of nanoparticles result in desired and undesired interactions. In addition to the useful applications of these nanomaterials, anthropogenic nanoparticles formed

from human activity are also an area of active study due to their significant negative effects on human health and the environment. One of the largest sources of these harmful nanoparticles are soot and combustion nanoparticles which form from unburned hydrocarbon fuels during combustion[10,11]. These molecules have a number of adverse health effects[12–15] and are a key contributor to climate change[16]. Research has suggested that engine design[17,18] and fuel chemistry[19] both play a key role in the amount and types of these nanoparticles that form. Thus understanding the interactions that lead to the growth of these pollutant nanoparticles and their precursors is also paramount to designing systems which mitigate their effects.

Computational methods offer a valuable tool to study nanoparticle interactions. Broadly, these methods can be divided into atomistic simulations which simulate nanoparticle interactions using models grounded in physio-chemical principles, numerical descriptors which provide a quantification of both nanoparticle properties and interactions, and machine learning which discovers patterns in known nanoparticle interactions and extends these relationships to predict new interactions.

These computational methods have a number of unique advantages which make them particularly promising for the study of these interactions. In many cases, computational methods offer an efficient means to study nanoparticles. Computational methods can scale with computational resources and be highly automated which makes them particularly suited to analyzing large candidate nanomaterial databases and optimizing nanoparticle features for a specific application in order to produce highly targeted experimental studies. Furthermore, computational methods are often able to provide highly detailed atomic level information in dynamic processes of nanoparticle interactions such as atomic positions, bonds, forces, electrons, and thermodynamic energies. This is especially useful in highly reactive systems such as combustion environments which are difficult to measure due to their high reactivity and propensity of some measurement techniques to potentially alter the system

state[20,21]. Finally, the development of computational models of nanoparticle interactions is intrinsically linked to our understanding of nanoscale chemistry. A high fidelity atomistic simulation must incorporate the relevant physical phenomena in order to properly reproduce experimental measurements. Thus by assessing the areas of agreement and deviation with these measurements, we can gain quantitative insights into the accuracy of our chemical models and our assumptions about the process. When considering machine learning, supervised models require a representation of the nanoparticles which can be numerically related to the interaction. An accurate machine learning model suggests that the input representation includes the chemical features which drive the nano-interaction. Thus, this thesis looks to apply computational methods to better understand nanoparticle growth and function by characterizing nanoscale interactions.

A number of unique challenges exist towards applying computational methods towards the prediction of nanoscale interactions. These challenges include a lack of existing nanoparticle datasets, the heterogeneity of nanoparticle environments, the difficulty in representing complex nanoparticle chemistry, and the high computational cost of measuring nanoparticle properties. Ultimately, these challenges will influence the direction of this thesis, as it aims to demonstrate the necessity of employing multiple computational methods for overcoming the unique obstacles of predicting nanoscale interactions. With this guiding concept, the thesis will apply atomistic simulation, numerical descriptors, and machine learning to both human-engineered and naturally occurring nanostructures in order to quantify a diverse set of interactions. The wide breadth of computational methods and broad nanoparticle domains covered this work will offer a road-map which can be applied to computational study of nanoscale systems.

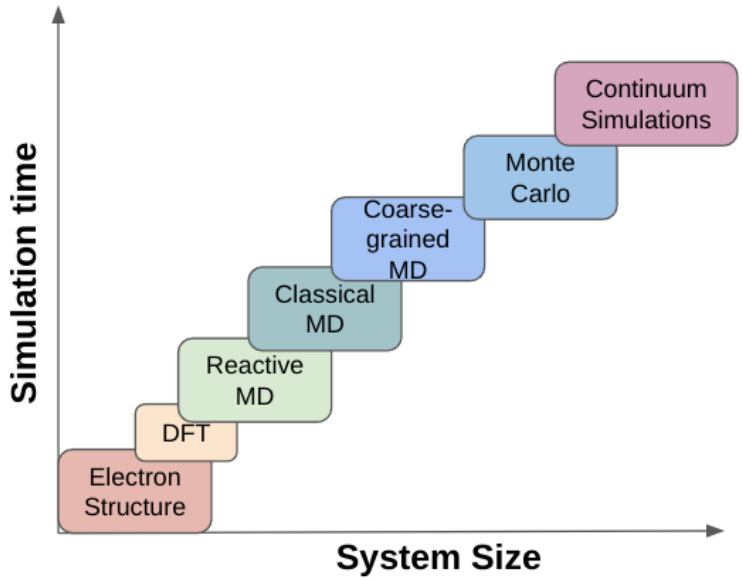

Figure 1.1: Comparison between different atomistic simulation techniques. Each simulation technique is plotted according to the relative number of atoms and timescale.

1.2 Atomistic Simulations to Predict Nanoscale Interactions

Atomistic simulations are a broad class of computational techniques which have been used to explore nano-interactions. The specific simulation methods which have been used to simulate these nano-interactions depend on the level of detail, size of the simulation, and length of simulation time[22,23]. A brief qualitative comparison between the levels of detail is given in figure 1.1. While nanomaterial studies have used simulations ranging from density functional theory calculation methods to describe nanoparticle reactivity[23] to finite element calculations to describe material properties[24], a wide variety of nanoparticle interactions occur on the scale of hundreds to thousands of atoms and picoseconds to nanoseconds of time. As such, three methods will be highlighted in this work which have been well suited to simulate these scales: molecular dynamics (MD), coarse-grained MD, and Monte Carlo methods.

Molecular dynamics simulates the motion of individual atoms in time by numerically solving Newton's equations of motion. A MD simulation undergoes a large

number of discrete time steps where the position and velocity of individual atoms are adjusted based on their current position, current velocity, and acceleration which is derived from a calculated force. At the center of MD is the force field which is a set of equations and numerical constants which describe the forces of each atom and can consider properties such as electrostatics, bond lengths, and bond torsion[25]. The force field is typically chosen for the specific system (*e.g.* organic or inorganic) and represents one of the costliest parts of the MD computation. MD simulations have revealed valuable information about nanoparticle interactions with biological structures[26,27], small molecules[28], and other classes of engineered nanoparticles[29,30]. Enhanced sampling MD techniques such as well-tempered Metadynamics and replica exchange MD[31–33] have also been introduced which allow sampling of rare events in nanomaterial interactions such as energetic barrier crossings[34,35].

One of the drawbacks of MD simulations is that they become computationally prohibitive for larger sized systems. For example, particle mesh Ewald electrostatic calculations in all-atom MD scales in complexity with the number of atoms n according to $O(n \log n)$[36]. In order to overcome this size limitation, a class of methods known as coarse-graining decomposes a nanoparticle into clusters of atoms known as beads. This effectively reduces the number of discrete particles which need to be accounted for in the simulation and can significantly reduce the computational cost. After this, a coarse-grained force field can be derived to describe the forces on these beads and a more computationally efficient MD simulation can be run. These coarse-grained simulations have been used extensively to study interactions between biological nanostructures such as proteins[37]. In these simulations, the proteins are typically coarse-grained according to their amino acids and applied to model nano-interactions in applications of self-interaction (folding)[38] and docking[39]. More recently, using physical relationships such as core-shell modeling, these methods have been extended to a broader class of nanomaterials[40].

The final class of nanoparticle interaction simulation methods discussed herein are Monte Carlo methods. These methods are characterized by stochastic sampling of molecular configurations. With sufficient sampling it is possible to use Monte Carlo simulations to explore the molecular state space. Unlike the previously discussed MD methods, Monte Carlo methods rely on discrete stochastic transitions and do not require solving Newton's equations of motion. As a result, these simulations can be significantly more efficient[41]. A number of Monte Carlo sampling algorithms are based around Markov chains and focus on sampling states in equilibrium[42]. More applicable to the nanoparticle interactions discussed herein are kinetic Monte Carlo methods which explore the dynamics of an interactions by considering an evolving set of different transition rates. Using the Gillespie algorithm, a series of reactions or transitions are stochastically selected according to their relative rate along with a time step update inversely related to the sum of all transition probabilities[43]. The result is a history of a molecule's evolution in time. Kinetic Monte Carlo has shown success in characterizing a number of nano-interactions such as the chemical reactions leading to the growth of carbon nanostructures in flames[44–47] and the interactions of magnetic nanoparticles for biological applications[48].

1.3 Numerical Descriptors of Nanoparticles

While atomistic simulations can produce a large amount of data, there exists a need for methods which can post-process and characterize these atomistic simulations to provide quantitative interpretations. As simulations often produce high dimensional representations such as atomic coordinates or molecular structures evolving in time, numerical molecular descriptors are used to distill this information both to characterize these nano-interactions in a manner that is comprehensible to humans and to highlight properties of interest. This requires a set of descriptors which allows a molecule to be described by its properties. The diversity and unique properties of

nanoparticles makes this a non-trivial task.

Significant work has focused on finding these properties for small molecules, however, many of these descriptors are either overly simplistic (*e.g.* binary fingerprints[49]) or not universally applicable to all nanoparticle chemistries (*e.g.* numbers pertaining only to covalent bonds[50]). Still, a large number of publications[51–54] and software packages[55–57] exist on the subject of representing molecules with numerical descriptors. Broadly, there are four classes of chemical descriptors based on the information which is required to compute them (figure 1.2): 0-D Descriptors contain purely compositional information and can be derived solely from the molecular formula. 1-D Descriptors pertain to subgroups within the molecule such as the presence of a functional group or length of an aliphatic side chain. 2-D Descriptors are defined entirely by the atoms and their connectivity, this includes many graph based and topological descriptors[50]. Finally, 3-D Descriptors offer a 3-D definition of atoms and thus coordinates are required to compute these descriptors. While 3-D descriptors offer the most detailed description of nanoparticles, they require structural information and can be sensitive to configuration changes and changes to atomic positions of the nanoparticle in time. Higher dimensional descriptors such as those representing multiple conformers or of multiple descriptors also exist.

In providing interpretibility to MD simulations, descriptors have been applied to relate properties of nanoparticles to potential energies and free energies measured through simulation outputs[58–60]. Furthermore, 2-D descriptors can offer a means to detect complex interactions through graph-based descriptors based on proximity[61]. A number of descriptors have also been used to validate simulations by computing properties which can be related to experimental measurements[62,63].

In addition to interpreting atomistic simulation results, numerical descriptors can also be used as inputs into data-driven machine learning models[64,65]. Recently, attempts have been made to create nanoparticle-specific sets of descriptors that include

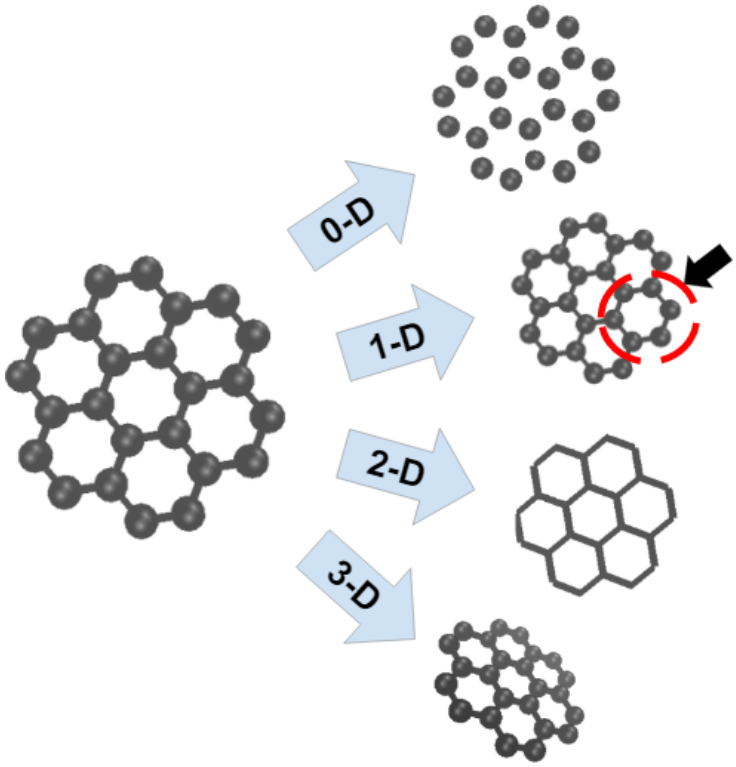

Figure 1.2: Visual representation of different descriptor types.

both structural representations and chemical features[66]. While these descriptors have typically been applied to small, homogeneous datasets and are limited in their specificity, they have demonstrated good predictive capabilities and produced chemical insights in their applications[67].

1.4 Machine Learning to Predict Nanoparticle Interactions

While atomistic simulations and numerical descriptors provide a means to study nanoparticle interactions, data-driven methods such as supervised machine learning are also a valuable tool to characterize these interactions. In this context, supervised machine learning, hereon referred to simply as machine learning, is a class of algorithms which can take as input existing nanoparticle interaction data along with a representation of the nanoparticles, typically a numerical vector of chemically meaningful features. The algorithm is then trained on this existing data, where parameters

are fit to relate the nanoparticle representation to the known interaction data. After this, the model can take as input the representation of any arbitrary nanoparticle and predict an interaction parameter of interest such as a free energy, rate constant, or interaction site.

Machine learning is a valuable tool for a number of reasons. First, machine learning offers a means to derive data where measurement by direct experiment or atomistic simulation is challenging. Experimental or atomistic simulation data is often costly to collect and machine learning allows a means to more cheaply obtain this nano-interaction data. One class of nanostructures where this has been particularly apparent is the development of machine learning methods for protein structure prediction. Protein amino acid sequences are significantly cheaper to measure than protein structures and as such, much experimental work has focused on using machine learning to predict self-interaction (*i.e.* folding) of these structures given protein sequences[68]. In other applications, machine learning has also offered a means to predict protein-nanoparticle interactions which are costly to measure with experiments[69,70]. Machine learning also fills a void where simulation is not practical due to computational cost. The binding energies of small molecules to catalytic metal nanoparticles has been studied with a number of machine learning methods which reduces the computational cost associated with density functional theory calculations[71,72] Recently, a study of silicon nanoparticle aggregation required millions of hours of MD simulations to simulate approximately one hundred different interactions, far below what is likely to occur in a real reactor system[73]. A machine learning model was able to be trained on this data and predict the interaction parameters of any arbitrary pair with negligible computational cost[74]. In addition to computational cost, machine learning also has been used to predict nanoparticle interactions which are difficult to model with simulation due to their complexity. For example, the formation of soot nanoparticles involves multiple chemical and physical interaction mechanisms occurring at differ-

ent length scales which would require a complex multi-scale simulation. However, machine learning has shown success finding complex quantitative relationships and using easily computed gas-phase chemistry to predict soot formation rates which are otherwise difficult to simulate[75,76].

Next, machine learning allows more targeted experiments of nanoparticle interactions. The nanomaterial design space is intractably large, and machine learning can identify promising areas for future study. Recently, machine learning was able to predict the biological interactions of nanoparticles with bacteria and make specific recommendations to design effective nanoparticle anti-microbials[64]. Another machine learning method identified two anti-cancer therapeutic nanoparticles which proved to be highly effective under experimental conditions[77]. A number of other machine learning works have yielded successful experiments which optimized nanoparticle synthesis[78], nano-medicine activity[64], and nanomaterial properties[67].

Finally, machine learning can also act as a tool to improve atomistic simulations of nanoparticle interactions. A number of works have focused on using novel atomic representations to develop more accurate and efficient MD force fields[79–81]. These works have been extended to enhance the accuracy of nanoparticle simulations[82]. In addition to force fields, machine learning has been used to provide more effective initial configurations for MD studies. A number of studies have used machine learning to predict protein interaction sites or configurations which allowed interaction simulations to proceed more efficiently than simulations which are naive to the protein structure and interacting residues[83,84]. Still the limited number of reliable machine learning prediction methods for non-protein nanostructures has precluded these methods being extended to the broader nanoparticle chemical space.

1.5 Challenges Extending Computational Frameworks to Nanoparticles

The application of computational methods for predicting nanoscale interactions presents several unique challenges, such as the absence of adequate nanoparticle datasets, the heterogeneity of nanoparticle environments, the complexity of representing nanoparticle chemistry, and the high computational cost associated with measuring many nanoparticle properties.

The first challenge is that nanoparticles often have significantly less data available than other domains. Molecules such as proteins and small molecules are characterized by large databases with significant information about chemical structure and properties[85,86] which has allowed for the rapid prediction these molecules' interactions[87–89]. Unlike proteins and small molecules, however, most nanoparticles do not have such available datasets. As such, computational works looking to broadly study nanoparticle interactions need to create equivalent datasets which can be used to quantify nanoparticle interactions or act as inputs into data-driven machine learning models. Creation of such datasets is not trivial and requires either adapting a database from the literature[90,91] or creating new data through atomistic simulation[65]. The former is time-consuming and depends on appropriate data being available. The later is often more practical as it can be generally applied to a nanoscale system but requires a simulation methodology which can capture the underlying chemistry.

While atomistic simulation data can be an effective way to create data to study nanoparticle interactions, it often needs to be combined with other computational methods in order to extract meaningful insights. Nanoparticles often do not exist in a system as a single structure but typically exist as an ensemble of molecules with a distribution of sizes, morphologies, and chemical features[92–95]. When this occurs, simulation results are often too complex to directly interpret. To remedy

this, numerical descriptors are needed to decompose a large number of individual molecular configurations into quantifiable values which describe the chemistry of the nanoparticles and their interactions.

In order to properly use these descriptors, however, one must account for the unique chemistry which occurs during nanoscale interactions. Many numerical descriptors which have typically been developed for organic small molecules[53,96] or ionic crystal structures[97] will often fail when applied to nanoparticles which often contain a mixture of ionic and covalent forces. Furthermore, the assumptions these descriptors make about molecular chemistry may not extend to nanointeractions which are governed by a unique mix of size-dependent interior contributions and surface effects[98]. In addition to all this, due to the ensemble of different nanoparticles in a system, a descriptor combination method is often needed which accounts for both the mean properties of the nanoparticles but also tails of the nanoparticle property distribution which may disproportionately contribute to the interactions[99]. For example, it has been determined that a few highly positive charged configurations of silver nanoclusters drive the interaction of these nanoparticles with carbon monoxide[100].

Finally, due to computational costs, simulations and descriptors alone are not enough to characterize all kinds of nanoparticle interactions. The large number of unique structures present in nanoparticle environments means that there are many potential interactions which might occur. For example, when considering binary interactions in a heterogeneous nanoparticle environment, the number of potential interactions which must be accounted for is $\frac{n^2+n}{2}$ where n is the number of unique nanoparticle structures (see appendix A). Such scaling often precludes direct atomistic simulation as the large number of simulations would be cost-prohibitive. Thus in many cases, in order to account for all potential interactions needed to accurately model a nanoparticle system, more efficient techniques such as machine learning are needed for characterizing large numbers of interactions. While techniques such as

machine learning, when properly implemented, can quantify these nano-interactions it still faces many of the same aforementioned challenges as it requires large amounts of data and reliable chemical descriptors.

Overall, each computational method discussed in this work is particularly suited to address some of these challenges but still faces significant limitations. Atomistic simulation allows for the creation of physically meaningful datasets of nanoparticle interactions but suffers from a high computational cost and still requires a means for interpretation. Chemical descriptors provide a meaningful representation of nanoparticles and their interactions but require additional computational techniques to create data for these descriptors and extend these numerical values to new systems. Finally machine learning allows for the rapid quantification of many nanoscale interactions, however also requires data and numerical nanoparticle representations in order to train and use the model. As a result of these limitations, nanoparticle interactions typically can not be sufficiently quantified with a single computational technique. Accurate characterization of nanoscale interactions requires multiple computational techniques in concert to overcome these challenges.

1.6 Dissertation Framework

This thesis will describe my work and findings in using computational methods to predict nanoscale interactions. I will focus on two broad scientific questions applied to a number of different nanoscale systems. First, how can computational methods quantify nanoparticle growth through the chemical and physical interactions in which they participate. Specifically, I will focus on the kinds of growth and properties that arise as a result of these interactions. Secondly, how can computational methods relate the properties of nanoparticles to their function by predicting the interactions which they participate in.

These overarching questions will be applied to three different systems: the chem-

ical interactions leading to the growth of nanoparticles in the gas-phase, the physical interactions of nanoparticles in the gas-phase, and liquid-phase interactions in biological systems.

For the first two gas-phase domains, I will focus on a class of nanostrucutres known as polycyclic aromatic compounds (PACs). I will show how computational methods can predict the types of nanostructures formed in combustion environments. I will use kinetic Monte Carlo simulations to represent chemical interactions and well-tempered Metadynamics to offer information about their physical aggregation. I will then show how numerical descriptors can quantify the properties of large ensembles of PACs to show the evolution of their growth in flame environments. Finally, I will demonstrate how data derived from my simulations and represented with these descriptors can be used as inputs into machine learning models both to predict the formation of larger combustion nanoparticles such as soot as well as predict the thermodynamic stability and rates of PAC physical clustering.

For the third domain, I will introduce a framework for better understanding the function of nanoparticles by predicting general nanoscale interactions. In contrast to the first two domains, this general prediction application looks to address a case where computational or experimental data for a specific nanoscale system may not be available and might be difficult to obtain. This approach will show how data can be leveraged from multiple sources with different levels of data availability. Significant discussion will focus on a multi-scale coarse-grained representation which can decompose any nano-structure into a set of sub-units which can then be represented by numerical descriptors to a machine learning model to predict interaction sites. I will then demonstrate how this method can be used to predict a number of diverse biological nano-interactions such as protein-protein interactions, protein-nanoparticle interactions, and nanoparticle-nanoparticle interactions.

The prediction of nano-interactions is a challenging problem due to their novel,

unique, and diverse chemistries. Across all the aforementioned systems, I will develop and apply multiple different kinds of computational methods together to overcome these challenges and accurately characterize how these nanoparticles interact in a variety of different contexts. The work in this thesis will provide a road map for using computational methods to quantify nanoscale interactions.

1.7 References

[1] Carlos A. Silvera Batista, Ronald G. Larson, and Nicholas A. Kotov. Nonadditivity of nanoparticle interactions. *Science*, 350(6257):1242477, 2015.

[2] Nadeem Joudeh and Dirk Linke. Nanoparticle classification, physicochemical properties, characterization, and applications: a comprehensive review for biologists. *J Nanobiotechnol*, 20(262), 2022.

[3] Nozomu Suzuki, Yichun Wang, Paolo Elvati, Zhi-Bei Qu, Kyoungwon Kim, Shuang Jiang, Elizabeth Baumeister, Jaewook Lee, Bongjun Yeom, Joong Hwan Bahng, Jaebeom Lee, Angela Violi, and Nicholas A. Kotov. Chiral graphene quantum dots. *ACS Nano*, 10(2):1744–1755, 2016.

[4] Susie Eustis and Mostafa A. El-Sayed. Why gold nanoparticles are more precious than pretty gold: Noble metal surface plasmon resonance and its enhancement of the radiative and nonradiative properties of nanocrystals of different shapes. *Chem. Soc. Rev.*, 35(3):209–217, 2006.

[5] Ibrahim Khan, Khalid Saeed, and Idrees Khan. Nanoparticles: Properties, applications and toxicities. *Arabian Journal of Chemistry*, 12(7):908–931, 2019.

[6] Michael J. Mitchell, Margaret M. Billingsley, Rebecca M. Haley, Marissa E. Wechsler, Nicholas A. Peppas, and Robert Langer. Engineering precision nanoparticles for drug delivery. *Nature Reviews Drug Discovery*, 20(2):101–124, 2021.

[7] Somnath Bhattacharjee, David M. Dotzauer, and Merlin L. Bruening. Selectivity as a function of nanoparticle size in the catalytic hydrogenation of unsaturated alcohols. *Journal of the American Chemical Society*, 131(10):3601–3610, 2009.

[8] Natalie Malikova, Isabel Pastoriza-Santos, Martin Schierhorn, Nicholas A. Kotov, and Luis M. Liz-Marzán. Layer-by-Layer Assembled Mixed Spherical and Planar Gold Nanoparticles: Control of Interparticle Interactions. *Langmuir*, 18(9):3694–3697, 2002.

[9] Reto Strobel and Sotiris E. Pratsinis. Flame aerosol synthesis of smart nanostructured materials. *Journal of Materials Chemistry*, 17(45):4743, 2007.

[10] Andrea D'Anna. Combustion-formed nanoparticles. *Proceedings of the Combustion Institute*, 32(1):593–613, 2009.

[11] Hai Wang and Michael Frenklach. A detailed kinetic modeling study of aromatics formation in laminar premixed acetylene and ethylene flames. *Combust. Flame*, 110(1):173–221, 1997.

[12] Per Gerde, Bruce A Muggenburg, Margot Lundborg, and Alan R Dahl. The rapid alveolar absorption of diesel soot-adsorbed benzo[a]pyrene: bioavailability, metabolism and dosimetry of an inhaled particle-borne carcinogen. *Carcinogenesis*, 22(5):741–749, 2001.

[13] Thomas R Barfknecht. Toxicology of soot. *Prog. Energy Combust. Sci.*, 9(3):199–237, 1983.

[14] Dana Loomis, Yann Grosse, Béatrice Lauby-Secretan, Fatiha El Ghissassi, Véronique Bouvard, Lamia Benbrahim-Tallaa, Neela Guha, Robert Baan, Heidi Mattock, and Kurt Straif. The carcinogenicity of outdoor air pollution. *Lancet Oncol.*, 14(13):1262, 2013.

[15] Aleksandra Stanković, Dragana Nikić, and Maja Nikolić. Relationship between exposure to air pollution and occurrence of anemia in pregnancy. *Facta Univ Ser Med Biol*, 13(1):54–57, 2006.

[16] Drew Shindell, Johan CI Kuylenstierna, Elisabetta Vignati, Rita van Dingenen, Markus Amann, Zbigniew Klimont, Susan C Anenberg, Nicholas Muller, Greet Janssens-Maenhout, Frank Raes, et al. Simultaneously mitigating near-term climate change and improving human health and food security. *Science*, 335(6065):183–189, 2012.

[17] Dale R. Tree and Kenth I. Svensson. Soot processes in compression ignition engines. *Progress in Energy and Combustion Science*, 33(3):272–309, 2007.

[18] Peng Ye and André L. Boehman. An investigation of the impact of injection strategy and biodiesel on engine NOx and particulate matter emissions with a common-rail turbocharged DI diesel engine. *Fuel*, 97:476–488, 2012.

[19] Perrine Pepiot-Desjardins, H. Pitsch, Ripudaman Malhotra, Stephen R. Kirby, and Andre L. Boehman. Structural group analysis for soot reduction tendency of oxygenated fuels. *Combustion and Flame*, 154(1):191–205, 2008.

[20] Allan N. Hayhurst and David B. Kittelson. Mass spectrometric sampling of ions from atmospheric pressure flames—III: Boundary layer and other cooling of the sample. *Combust. Flame*, 28:137–143, 1977.

[21] Ulf Struckmeier, Patrick Oßwald, Tina Kasper, Lena Böhling, Melanie Heusing, Markus Köhler, Andreas Brockhinke, and Katharina Kohse-Hoeinghaus. Sampling probe influences on temperature and species concentrations in molecular beam mass spectroscopic investigations of flat premixed low-pressure flames. *Z. Phys. Chem.*, 223(4-5):503–537, 2009.

[22] Tommaso Casalini, Vittorio Limongelli, Mélanie Schmutz, Claudia Som, Olivier Jordan, Peter Wick, Gerrit Borchard, and Giuseppe Perale. Molecular Modeling for Nanomaterial–Biology Interactions: Opportunities, Challenges, and Perspectives. *Frontiers in Bioengineering and Biotechnology*, 7:268, 2019.

[23] Priyanka Makkar and Narendra Nath Ghosh. A review on the use of DFT for the prediction of the properties of nanomaterials. *RSC Advances*, 11(45):27897–27924, 2021.

[24] Raza Ansari, Jalal Torabi, and Amir Norouzzadeh. Bending analysis of embedded nanoplates based on the integral formulation of Eringen's nonlocal theory using the finite element method. *Physica B: Condensed Matter*, 534:90–97, 2018.

[25] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.

[26] Changjiang Liu, Paolo Elvati, Sagardip Majumder, Yichun Wang, Allen P. Liu, and Angela Violi. Predicting the Time of Entry of Nanoparticles in Lipid Membranes. *ACS Nano (accepted)*, 2019.

[27] Yichun Wang, Usha Kadiyala, Qu Zhibei, Paolo Elvati, Christopher Altheim, Nicholas A. Kotov, Angela Violi, and J. Scott VanEpps. Anti-biofilm activity of graphene quantum dots via self-assembly with bacterial amyloid proteins. *J. Phys. Chem. A*, 13(4):4278–4289, 2019.

[28] Jeffrey Comer, Ran Chen, Horacio Poblete, Ariela Vergara-Jaque, and Jim E. Riviere. Predicting Adsorption Affinities of Small Molecules on Carbon Nanotubes Using Molecular Dynamics Simulation. *ACS Nano*, 9(12):11761–11774, 2015.

[29] Ayse Cetin and Mine Ilk Capar. Functional-Group Effect of Ligand Molecules on the Aggregation of Gold Nanoparticles: A Molecular Dynamics Simulation Study. *The Journal of Physical Chemistry B*, 126(29):5534–5543, 2022.

[30] Paolo Elvati, Elizabeth Baumeister, and Angela Violi. Graphene quantum dots: effect of size, composition and curvature on their assembly. *RSC Advances*, 29, 2017.

[31] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. *Physical Review Letters*, 100(2):020603, 2008.

[32] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual Review of Physical Chemistry*, 67(1):159–184, 2016.

[33] Cameron Abrams and Giovanni Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2014.

[34] Paolo Elvati and Angela Violi. Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains. *Proceedings of the Combustion Institute*, 34(1):1837–1843, 2013.

[35] Francesco Tavanti, Alfonso Pedone, and Maria Cristina Menziani. Disclosing the interaction of gold nanoparticles with aβ(1–40) monomers through replica exchange molecular dynamics simulations. *International Journal of Molecular Sciences*, 22(1), 2021.

[36] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[37] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The martini force field: Coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.

[38] Gia G. Maisuradze, Patrick Senet, Cezary Czaplewski, Adam Liwo, and Harold A. Scheraga. Investigation of Protein Folding by Coarse-Grained Molecular Dynamics with the UNRES Force Field. *The Journal of Physical Chemistry A*, 114(13):4471–4485, 2010.

[39] Paulo C. T. Souza, Sebastian Thallmair, Paolo Conflitti, Carlos Ramírez-Palacios, Riccardo Alessandri, Stefano Raniolo, Vittorio Limongelli, and Siewert J. Marrink. Protein–ligand binding with the coarse-grained Martini model. *Nature Communications*, 11(1):3714, 2020.

[40] Ankush Singhal and G. J. Agur Sevink. A Core-Shell Approach for Systematically Coarsening Nanoparticle–Membrane Interactions: Application to Silver Nanoparticles. *Nanomaterials*, 12(21):3859, 2022.

[41] Herma M. Cuppen, Leendertjan J. Karssemeijer, and Thanja Lamberts. The kinetic monte carlo method as a way to solve the master equation for interstellar grain chemistry. *Chemical Reviews*, 113(12):8840–8871, 2013.

[42] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[43] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[44] Jason Y. W. Lai, Paolo Elvati, and Angela Violi. Stochastic atomistic simulation of polycyclic aromatic hydrocarbon growth in combustion. *Phys. Chem. Chem. Phys.*, 16(17):7969, 2014.

[45] Angela Violi. Modeling of soot particle inception in aromatic and aliphatic premixed flames. *Combust. Flame*, 139(4):279–287, 2004.

[46] Michael Frenklach, Zhenyuan Liu, Ravi I. Singh, Galiya R. Galimova, Valeriy N. Azyazov, and Alexander M. Mebel. Detailed, sterically-resolved modeling of soot oxidation: Role of O atoms, interplay with particle nanostructure, and emergence of inner particle burning. *Combust. Flame*, 188:284–306, 2018.

[47] Abhijeet Raj, Matthew Celnik, Raphael Shirley, Markus Sander, Robert Patterson, Richard West, and Markus Kraft. A statistical approach to develop a detailed soot growth model using PAH characteristics. *Combust. Flame*, 156(4):896–913, 2009.

[48] RP Tan, Julian Carrey, and Marc Respaud. Magnetic hyperthermia properties of nanoparticles inside lysosomes using kinetic monte carlo simulations: Influence of key parameters and dipolar interactions, and evidence for strong spatial variation of heating power. *Phys. Rev. B*, 90:214421, 2014.

[49] Ling Xue, Jeffrey W Godden, Florence L Stahura, and Jurgen Bajorath. Design and Evaluation of a Molecular Fingerprint Involving the Transformation of Property Descriptor Values into a Binary Classification Scheme. *J. Chem. Inf. Comput. Sci.*, 43(4):1151–1157, 2003.

[50] Ovidiu Ivanciuc, Teodora Ivanciuc, and Alexandru T. Balaban. Design of Topological Indices. Part 10.s Parameters Based on Electronegativity and Covalent Radius for the Computation of Molecular Graph Descriptors for Heteroatom-Containing Molecules. *Journal of Chemical Information and Computer Sciences*, 38(3):395–401, 1998.

[51] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settingsq. *Advanced Drug Delivery Reviews*, pages 3–26, 2001.

[52] Lowell H. Hall and Lemont B. Kier. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Modeling*, 35(6):1039–1045, 1995.

[53] Scott A. Wildman and Gordon M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

[54] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry. Wiley, 1 edition, 2000.

[55] Greg Landrum. Rdkit: Open-source cheminformatics.

[56] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. 2015.

[57] Jie Dong, Dong-Sheng Cao, Hong-Yu Miao, Shao Liu, Bai-Chuan Deng, Yong-Huan Yun, Ning-Ning Wang, Ai-Ping Lu, Wen-Bin Zeng, and Alex F. Chen. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics*, 7(1), 2015.

[58] Paolo Elvati, Kirk Turrentine, and Angela Violi. The role of molecular properties on the dimerization of aromatic compounds. *Proceedings of the Combustion Institute*, 37(1):1099–1105, 2019.

[59] Jeffrey S. Lowe, Jason Y.W. Lai, Paolo Elvati, and Angela Violi. Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity. *Proc. Combust. Inst.*, 35(2):1827–1832, 2015.

[60] Bangquan Li, Jing Li, Xiaoqiang Su, and Yimin Cui. Molecular dynamics study on structural and atomic evolution between au and ni nanoparticles through coalescence. *Scientific Reports*, 11:15432, 2021.

[61] Fabio Pietrucci and Wanda Andreoni. Graph theory meets ab initio molecular dynamics: Atomic structures and transformations at the nanoscale. *Phys. Rev. Lett.*, 107:085504, 2011.

[62] Qi Wang, Paolo Elvati, Doohyun Kim, K. Olaf Johansson, Paul E. Schrader, Hope A. Michelsen, and Anegla Violi. Spatial dependence of the growth of polycyclic aromatic compounds in an ethylene counterflow flame. *Carbon*, 149:328–335, 2019.

[63] Moshen Ramezanpour, Sherry S. W. Leung, Karelia H. Delgado-Magnero, BYM. Bashe, Jenifer Thewalt, and Dirk P. Tieleman. Computational and experimental approaches for investigating nanoparticle-based drug delivery systems. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1858(7, Part B):1688–1709, 2016.

[64] Mahsa Mirzaei, Irini Furxhi, Finbarr Murphy, and Martin Mullins. A machine learning tool to predict the antibacterial capacity of nanoparticles. *Nanomaterials*, 11(7):1774, 2021.

[65] Alex K. Chew, Joel A. Pedersen, and Reid C. Van Lehn. Predicting the physicochemical properties and biological activities of monolayer-protected gold nanoparticles using simulation-derived descriptors. *ACS Nano*, 16(4):6282–6292, 2022.

[66] Xiliang Yan, Alexander Sedykh, Wenyi Wang, Xiaoli Zhao, Bing Yan, and Hao Zhu. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale*, 11(17):8352–8362, 2019.

[67] Xiliang Yan, Alexander Sedykh, Wenyi Wang, Bing Yan, and Hao Zhu. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nature Communications*, 11(1):2519, 2020.

[68] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[69] Nicholas Ouassil, Rebecca L. Pinals, Jackson Travis Del Bonis-O'Donnell, Jeffrey W. Wang, and Markita P. Landry. Supervised learning model predicts protein adsorption to carbon nanotubes. *Science Advances*, 8(1):eabm0898, 2022.

[70] Minjeong Cha, Emine S.T. Emre, Xiongye Xiao, Ji-Young Kim, Ppaul Bogdan, J. Scott VanEpps, Angela Violi, and Nicholas A. Kotov. Unifying structural descriptors for biological and bioinspired nanoscale complexes. *Nature Computational Science*, 2:243–252, 2022.

[71] Ryosuke Jinnouchi and Ryoji Asahi. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *The Journal of Physical Chemistry Letters*, 8(17):4279–4283, 2017.

[72] Zachary W. Ulissi, Michael T. Tang, Jianping Xiao, Xinyan Liu, Daniel A. Torelli, Mohammadreza Karamad, Kyle Cummins, Christopher Hahn, Nathan S. Lewis, Thomas F. Jaramillo, Karen Chan, and Jens K. Nørskov. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO $_2$ Reduction. *ACS Catalysis*, 7(10):6600–6608, 2017.

[73] Xuetao Shi, Paolo Elvati, and Angela Violi. On the growth of si nanoparticles in non-thermal plasma: physisorption to chemisorption conversion. *Journal of Physics D: Applied Physics*, 54(36):365203, 2021.

[74] Paolo Elvati, Jacob C. Saldinger, Matt Raymond, Jonathan Lin, Xuetao Shi, and Angela Violi. Machine learning models for si nanoparticle growth in non-thermal plasma. *In Preparation*, 2023.

[75] Mehdi Jadidi, Stevan Kostic, Leonardo Zimmer, and Seth B Dworkin. An artificial neural network for the low-cost prediction of soot emissions. *Energies*, 13(18):4787, 2020.

[76] Mehdi Jadidi, Luke Di Liddo, and Seth B Dworkin. A long short-term memory neural network for the low-cost prediction of soot concentration in a time-dependent flame. *Energies*, 14(5):1394, 2021.

[77] Daniel Reker, Yulia Rybakova, Ameya R. Kirtane, Ruonan Cao, Jee Won Yang, Natsuda Navamajiti, Apolonia Gardner, Rosanna M. Zhang, Tina Esfandiary, Johanna L'Heureux, Thomas von Erlach, Elena M. Smekalova, Dominique Leboeuf, Kaitlyn Hess, Aaron Lopes, Jaimie Rogner, Joy Collins, Siddartha M. Tamang, Keiko Ishida, Paul Chamberlain, DongSoo Yun, Abigail Lytton-Jean, Christian K. Soule, Jaime H. Cheah, Alison M. Hayward, Robert Langer, and Giovanni Traverso. Computationally guided high-throughput design of self-assembling drug nanoparticles. *Nature Nanotechnology*, 16(6):725–733, 2021.

[78] Flore Mekki-Berrada, Zekun Ren, Tan Huang, Wai Kuan Wong, Fang Zheng, Jiaxun Xie, Isaac Parker Siyu Tian, Senthilnath Jayavelu, Zackaria Mahfoud, Daniil Bash, Kedar Hippalgaonkar, Saif Khan, Tonio Buonassisi, Qianxiao Li, and Xiaonan Wang. Two-step machine learning enables optimized nanoparticle synthesis. *npj Computational Materials*, 7(1):55, 2021.

[79] Jorg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011.

[80] Michael Gastegger, Ludwig Schwiedrzik, Marius Bittermann, Florian Berzsenyi, and Philipp) Marquetanda. wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.*, 148, 2018.

[81] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nature Communications*, 11(1):5461, 2020.

[82] Claudio Zeni, Kevin Rossi, Aldo Glielmo, and Francesca Baletto. On machine learning force fields for metallic nanoparticles. *Advances in Physics: X*, 4(1):1654919, 2019.

[83] Yao Sun, Yanqi Jiao, Chengcheng Shi, and Yang Zhang. Deep learning-based molecular dynamics simulation for structure-based drug design against SARS-CoV-2. *Computational and Structural Biotechnology Journal*, 20:5014–5027, 2022.

[84] Roy Nassar, Emiliano Brini, Sridip Parui, Cong Liu, Gregory L. Dignon, and Ken A. Dill. Accelerating Protein Folding Molecular Dynamics Using Inter-Residue Distances from Machine Learning Servers. *Journal of Chemical Theory and Computation*, 18(3):1929–1935, 2022.

[85] John J Irwin and Brian K Shoichet. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.*, 45(1):177–182, 2006.

[86] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[87] Ayoun Cho, Hongseok Yun, Jin Hwan Park, Sang Yup Lee, and Sunwon Park. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Syst Biol.*, 4(35), 2010.

[88] Matthias Brüstle, Bernd Beck, Torsten Schindler, William King, Timothy Mitchell, and Timothy Clark. Descriptors, Physical Properties, and Drug-Likeness. *Journal of Medicinal Chemistry*, 45(16):3345–3355, 2002.

[89] Yu-Chen Lo, Stefano E. Rensi, Wen Torng, and Russ B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug Discovery Today*, 23(8):1538–1546, 2018.

[90] Mayank Baranwal, Abram Magner, Paolo Elvati, Jacob Saldinger, Angela Violi, and Alfred O Hero. A deep learning architecture for metabolic pathway prediction. *Bioinformatics*, 36(8):2547–2553, 2020.

[91] Mayank Baranwal, Abram Magner, Jacob C. Saldinger, Emine S. Turali-Emre, Paolo Elvati, Shivani Kozarekar, J. Scott VanEpps, Nicholas A. Kotov, Angela Violi, and Alfred O. Hero. Struct2graph: a graph attention network for structure based predictions of protein–protein interactions. *BMC Bioinformatics*, 23(370), 2022.

[92] Bocheng Zhang, Heng Liu, Xiangyi Huang, Chaoqing Dong, and Jicun Ren. Size distribution of nanoparticles in solution characterized by combining resonance light scattering correlation spectroscopy with the maximum entropy method. *Analytical Chemistry*, 89(22):12609–12616, 2017.

[93] Zhen H. Li and Donald G. Truhlar. Nanothermodynamics of metal nanoparticles. *Chemical Science*, 7:2605–2624, 2014.

[94] Hai Wang. Formation of nascent soot and other condensed-phase materials in flames. *Proceedings of the Combustion Institute*, 33(1):41–67, 2011.

[95] Lutz Mädler, Hendrik K. Kammler, Roger Mueller, and Sotiris E. Pratsinis. Controlled synthesis of nanostructured particles by flame spray pyrolysis. *Journal of Aerosol Science*, 33(2):369–389, 2002.

[96] Paul Labute. A widely applicable set of descriptors. *J. Mol. Graph*, 18(4-5):464–477, 2000.

[97] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8(1):15679, 2017.

[98] Younjin Min, Mustafa Akbulut, Kai Kristiansen, Yuval Golan, and Jacob Israelachvili. The role of interparticle and external forces in nanoparticle assembly. *Nature Materials*, 7:527–538, 2008.

[99] Paolo Elvati, V. Tyler Dillstrom, and Angela Violi. Oxygen driven soot formation. *Proceedings of the Combustion Institute*, 36(1):825–832, 2017.

[100] Kaining Duanmu and Donald G. Truhlar. Partial ionic character beyond the pauling paradigm: Metal nanoparticles. *The Journal of Physical Chemistry C*, 118(48):28069–28074, 2014.

CHAPTER II

Chemical Growth of Gas-Phase Nanoparticles

2.1 Summary

In this chapter, I discuss how the chemical reactions of precursor nanostructures lead to the formation of combustion nanoparticles in flame environments. I focus on a class of precursor molecules known as polycyclic aromatic compunds (PACs) and their growth through chemical interactions with small molecules in the gas-phase. In this section, I apply computational methods to describe how these nanostructures form through chemical interactions, how different chemical interactions contribute to the properties' of PACs, and the implications of these properties for the broader transition of these precursors into larger combustion nanoparticles. To address these areas, I study a diverse set of flame systems and apply *kMC* simulations, descriptors, and machine learning in a number of different contexts. After introducing the problem and explaining the common methodology, each section in this chapter describes at least one of my publications regarding the computational study of a specific flame system and application[1–5]. Overall, I show how computational methods can validate PAC simulations against experiment, spatially characterize the development of chemical properties, explain observed PAC growth phenomena, and quantitatively relate PAC properties to more difficult to measure combustion nanoparticle growth phenomena.

2.2 Introduction

With the ubiquitous role of hydrocarbons in energy, transportation, and manufacturing, the formation of combustion byproducts remains a pressing concern. Unburned hydrocarbons react within the flame to form multi-ringed aromatic structures known as polycyclic aromatic compounds (PACs)[6]. These molecules are widely accepted to be a major precursor to combustion nanoparticles[7] which have negative impacts on both humans[8] and the environment[9]. Furthermore, as flame synthesis and flame spray pyrolysis are some of the most effective methods for high-throughput synthesis of nanoparticles[10], there is also a desire to better understand the role these PACs play in the growth of engineered carbon nanoparticles[11,12].

Despite decades of research on PACs, however, there is a poor understanding of the types of structures formed in flame environments and the conditions leading to their formation. Past studies have determined that PACs grow through a variety of competing pathways[13–19]. These works highlight the large number of possible PAC growth reactions and molecules present in combustion. Still, the majority of research neglects a diverse range of structures that have been experimentally observed[20] focusing instead on a narrow range of stabilomer[21] PACs made up of pericondensed, six-membered carbon rings.

Such simplistic approaches offer insufficient insight into the PACs' chemical space as they fail to fully characterize PACs' properties and effects. Numerous works have indicated that PACs' size and aliphatic branching along with the flame environment determine whether PACs will nucleate into larger combustion nanoparticles[22–25]. Recently, Elvati *et al.* presented a nuanced view of molecular descriptors, showing how the complex interplay of properties such as mass, presence of rotatable bonds, and oxygen content affect the propensity of PACs to dimerize and form clusters[26,27]. Beyond nanoparticle formation, PACs' structure is relevant in health studies where specific properties have been linked to cellular uptake[28,29], toxicity[30], and bacterial

interactions[31]. Thus, an accurate depiction of PACs' diversity is paramount to properly capture the effects specific combustion environments have on health and the environment.

There is a pressing need to link specific combustion environments (*i.e.* temperature and composition of gas-phase species) to the properties and growth pathways of these molecules. A comprehensive analysis of the PACs' chemical landscape is made difficult by the complexity, in terms of number, variety, and reactivity, of species within the system. Deterministic methods simulate PACs with a series of rate equations as part of a larger combustion model[15,32]. Not only do these methods require *a priori* knowledge of all species and reactions involved, but are also only applicable when continuous deterministic approximations are valid. As such, these methods are strongly limited in their accuracy by the narrow validity of the approximation they are built upon.

To remedy this, studies have found success predicting the formation of PACs' growth in flames through stochastic modeling. Frenklach and coworkers developed a kinetic Monte Carlo (*kMC*) model[33] focused on identifying growth pathways and has since produced updated *kMC* models including hydrogen-abstraction-acetylene-addition, five-membered ring migration, and ring oxidation[34,35]. Violi *et al.* created the AMPI code[36] that combines *kMC* and molecular dynamics to simulate polycyclic aromatic hydrocarbons' (PAHs') growth and has successfully reproduced many PACs observed experimentally. Raj *et al.* have developed a *kMC* model named *kMC*-ARS[37], to determine PAHs' surface growth based on edge sites[38]. This model has been applied to study the role of PAHs in a larger soot population balance[39], however, it focuses primarily on aromatic carbon rings and thus is not best suited for predicting other compounds (*e.g.* oxygenated PACs) that might be seen in flame systems. Lai *et al.* developed the SNapS code which has achieved excellent quantitative agreement with experiment[40]. SNapS is a *kMC*/molecular mechanics model that simulates the

growth of seed molecules into larger PACs through a series of general site reactions. This code has been repeatedly modified to include new sets of reactions[41] and has been validated under different experimental conditions, correctly predicting oxygenated PACs[20] and the presence of five-membered rings[42].

Recently, Wang *et al.* fully revised the SNapS code and created *SNapS2* to improve handling of time-steps and to remove the molecular mechanics calculations, which drastically improved the computational performances. Such improvements have allowed this software to analyze PACs' growth along selected streamlines within a 2-D counter-flow ethylene diffusion flame where a diverse compositional array of PACs were identified[19]. Given the important implications this diversity has on PAC growth and soot formation, it is imperative to further investigate practical fuels and flame configurations more representative of real systems.

In this chapter, I study the chemical interactions of these PACs to provide more information on this crucial step in the formation of combustion nanoparticles. My work in this section seeks to address three areas. **(1)** How atomistic simulations capture the chemical growth of these PACs. **(2)** How to use descriptors to characterize the properties of PACs which are produced during simulations. **(3)** How the simulations and descriptors can be combined with machine learning to quantitatively predict the formation of combustion nanoparticles. To address the issue of simulating these interactions, I apply *kMC* simulations using the *SNapS2* model on a number of different flame geometries and compositions. These flames include a coflow diffusion Jet A-1 surrogate aviation fuel flame as well as ethylene and ethylene-ethanol premixed flames. The *SNapS2* simulations provide a spatial dependence of PAC growth through an ensemble of PACs formed in the flame. However, direct interpretation of these results is challenging due to the diversity of species observed. Due to a combinatorial explosion of possible compounds it is common to see millions of unique structures[1,43]. It is not feasible to manually interpret such a large quantity of PACs

individually and as such numerical descriptors must be applied to characterize the PAC chemical space within these flame systems. These descriptors can be applied to validate against experiment, measure the development of certain nano-interactions and properties, and serve as inputs to machine learning models to predict the growth of combustion nanoparticles.

2.3 Methodology

The methodology of studying PACs can be broken down broadly into pre-processing, *SNapS2 kMC*, and post-processing. Pre-processing consists of obtaining small molecule gas-phase data to be used as input into *SNapS2*. This data consists of small molecules, typically one ring or smaller, formed in the flame with concentrations sufficient to be accurately modelled deterministically. This data is then input into *SNapS2* and the growth of PACs are simulated. Finally, *SNapS2* results are post-processed with numerical descriptors to gain additional insights into nano-interactions and the properties of the PACs.

2.3.1 Flame Systems

Four different flame systems are considered in this work. The first is a coflow diffusion laminar flame with a Jet A-1 surrogate fuel (69 % *n*-decane, 11 % *n*-propylcyclohexane, and 20 % *n*-propylbenzene)[44]. For this flame, the reaction conditions vary radially and axially. The remaining three systems are all 1-D pre-mixed flames which vary axially and are radially homogeneous. For the ethylene-ethanol study[3], six different ethylene flames are considered based on the work of Gerasimov *et al.*[45] with ethanol doping mole fractions of 0, 0.2, and 0.4 and equivalence ratios of 2.34 and 2.64. For the ethylene sooting flames used in the machine learning case study, three different pre-mixed ethylene-air flames from Xu *et al.*[46] were used with equivalence ratios of 2.34, 2.64, and 2.94. Finally for the comparisons

between *SNapS2* and atomic force microscopy, a pre-mixed ethylene-air flame from Commodo *et al.*[47] was used with an equivalence ratio of 2.03.

2.3.2 Pre-processing

In this step, the small-molecule gas-phase chemistry profile in each flame is obtained either from simulations or adapted from existing literature and is used to create inputs which can be used for *kMC* simulations. The specific pre-processing requirements of each flame varies slightly based on the flame configuration and study objectives.

For the Jet A-1 flame, the gas-phase data is adapted from an existing computational fluid dynamics (CFD) study. The details of the CFD simulations can be found in Saffaripour *et al.*[48] while additional experimental details used in validation can be found in the original experimental works[49,50]. In order to describe the environment to use in the *SNapS2* simulations, I used the CFD data to build two dimensional profiles (radial and axial) of species mole fractions, temperatures, and velocities for the flame. The streamlines were determined by numerical integration of the velocities starting at a radial distance of 0.5, 2.5, 5, 7.5, 10, and 12.5 mm from the flame central axis (assuming cylindrical symmetry). Of note, I have defined the center-line as the streamline originating at a radius of 0.5 mm. This small discrepancy is needed to avoid discontinuities and numerical artifacts arising from the boundaries in the CFD simulations and this choice does not affect the gas phase appreciably. Figure 2.1 shows the streamlines in relation to flame temperature. For the ethylene sooting flames used in the machine learning case study, the CFD was performed using the CoFlame package[51] by others with details being given in the corresponding publication[5]. Given that it was a 1-D flame, integration was performed only along the axial coordinates to obtain a spatio-temporal relationship. The other two sets of flames were simulated using the PREMIX code implemented in CHEMKIN[52] using the KM2 mechanism[32].

Specifically for the ethylene/ethanol flames, the KM2 mechanism was merged with an ethanol oxidation mechanism from Sarathy *et al.*[53]. Since the purpose of this study was to assess the change in PACs due to ethanol combustion chemistry separate from the effect of temperature, temperature related differences were minimized between all flames by keeping temperature profiles nearly constant by solving the energy equations in CHEMKIN and adjusting the cold gas velocity. This is similar to an approach employed for by Golea *et al.* to study benzene/ethanol flames[54]. For the ethylene atomic force microscopy comparison[4], an experimental temperature profile was used[47].

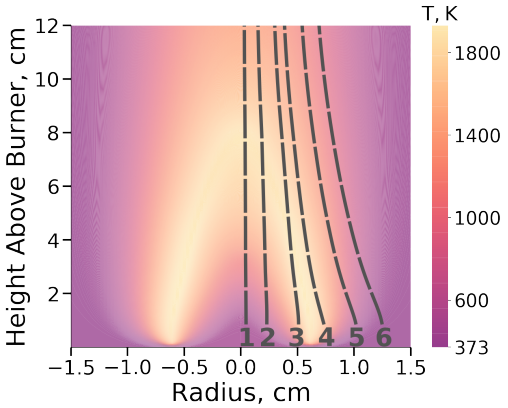

Figure 2.1: Temperature profile of the Jet-A1 coflow diffusion flame from CFD data published by Saffaripour and coworkers[48]. Streamlines created in this work are shown as dark lines.

2.3.3 *SNapS2* Simulations

After obtaining the flame temperatures and gas-phase species, chemical interactions with the PAC nanostructures were stochastically simulated with the *SNapS2* software with a methodology similar to the one used recently by Wang *et al.*[19]. In these simulations, a starting seed molecule was grown through a series of kMC reaction steps. At each step, one of approximately 400 general reactions[19] (small molecule addition, small molecule removal, or PAC isomerization) was stochastically selected proportional to the reaction rate. This was performed until the end of the streamline

was reached or the mass exceeded a mass threshold (specified for each flame below) to match the experimentally sampled range.

For all flames, a pool of small aromatic seed molecules were periodically introduced at intervals in the flame in approximate proportion to their concentration in the gas-phase mechanism. Based on the chemistry, flame, and study objective, the specific pool of seed molecules and intervals varied slightly.

For the coflow Jet A-1 flame[1,2], seed molecules were selected among a pool composed of cyclopentadiene, cyclopentadienyl, benzene, phenyl, toluene, napthalene, phenol, phenolate, phenanthrene, and acenaphthylene which includes a set of small aromatics diverse in carbon number, mass, and oxygenation and likely to be present in high concentrations. For computational cost reasons only seed molecules which reached, at any point in the streamline, a concentration greater than 5% compared to the maximum concentration of any seed molecule were simulated. In order to capture early growth, I started my simulations at a height above the burner (HAB) of 0 mm and at intervals of 1 ms or 5 ms depending if the concentration of the molecules used as a seed has already reached (or not) the 5 % concentration threshold. At each time interval, I ran 100 simulations of the seed molecule with the highest peak concentration (C_6H_6 in this flame) while the number of simulations for the remaining seed molecules was determined relative to each molecule's maximum concentration. For efficiency reasons, I stopped my simulations when the PAC either reached 600 u (where other growth phenomena become dominant[26,55]) or reached the end of my flame system. For the ethylene-air flames simulated in CoFlame[5,51], a similar procedure was adopted with a pool of cyclopentadiene, cyclopentadienyl, benzene, phenyl, phenol, toluene, napthalene, phenanthrene, and acenaphthylene and interval of 5 ms. For the pre-mixed flames used in the ethylene-ethanol study[43], benzene, cyclopentadiene, and phenol were used as seed molecules. For the pre-mixed ethlyene flame used to compare against atomic force microscopy[4], the goal was to identify specific struc-

tural motifs from *SNapS2* rather than matching overall PAC mass spectra intensities and therefore 37000 benzene seeds were used and an upper mass limit of 1000 u was considered. In all cases, seed contributions to final properties were normalized by the number of simulations performed at each time step so that the number of simulations did not affect a molecule's relative intensity.

2.3.4 Descriptor Computation

In order to describe PAC chemistry, I compute a large set of descriptors for each molecule. All descriptors can be derived from the molecular SMILES[56] of the PAC with negligible computational cost. These descriptors capture properties such as mass, atomic ratios such as carbon-hydrogen ratio, and counts of specific subgroups such as aromatic rings. In addition to these descriptors, I aggregate a large set of chemical descriptors from multiple other sources including WHIM descriptors[57], CPSA descriptors[58], VSA descriptors[59], and tessellation descriptors[60]. Additional details of these descriptors are provided in the methodology section of chapter III. For descriptors which require atomic positions, I optimize the geometry using the MMFF94 force field[61] in RDKIT[62].

In addition to the computation of descriptors for each individual molecule, it is also important to extend these descriptors to an ensemble of PACs. Due to the large number of different molecules which might be present at one time, this is important because the properties observed in the flame are often a function of the ensemble of PACs rather than an individual molecule. To this end, when computing a descriptor at a certain area in the flame each molecule is assigned a weight w^T according to equation 2.1 where w_c is the concentration of the trajectory's seed molecule at the start of the simulation, w_l is the fraction of time within the area of measurement which the molecule exists before and after reacting, and n_s is a normalization term equal to the number of replicate simulations run for that seed used. This final term is

included such that the choice of replicate simulations influences simulation diversity but not absolute intensity. A graphical depiction of these terms are given in figure 2.2

$$w^T = \frac{w_c w_l}{n_s} \tag{2.1}$$

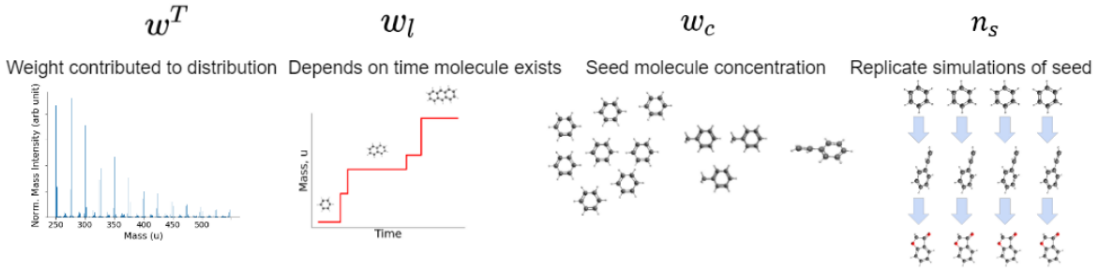

Figure 2.2: Graphical depiction of how nanoparticle property distributions are weighted in equation 2.1. The total weight (w_T) is proportional to the concentration of the seed molecule from which it comes (w_C) and the lifetime that molecule exists before its subsequent reaction (w_l) and normalized by the number of replicate simulations (n_s) for that seed.

With these weights, distributional information such as mean, median, minimum, maximum, and quartiles are then extracted to provide a comprehensive view of the PAC properties. These descriptors can then be compared to the appropriate experimental results to ensure the accuracy of the simulations. For example, average mass can be compared to mass spectrometry data as has been done previously[19]. Comparing distributional properties, especially those promoting aggregation, between the PACs of different flame systems will show how specific conditions and environments promote or inhibit the interactions leading to the growth of combustion nanoparticles.

2.4 PAC Growth in a Jet A-1 Surrogate Flame

The following results are adapted from two of my publications studying the formation of PACs in a Jet A-1 surrogate co-flow diffusion flame[1,2]. The key findings of this work demonstrate that the chemical interactions of PACs can be modeled in a complex fuel and flame geometry with my *kMC* simulations. Descriptors can be

used to validate against experiment[50] and to provide a comprehensive spatial characterization of properties in order to gain spatial insights into PACs' properties of interest.

2.4.1 Descriptors for Polycyclic Growth Validation

I first characterize the simulation results of the Jet A-1 flame by using descriptors to compare a number of different PAC properties from *SNapS2* against experimental measurements. Previous experimental electron ionization mass spectrometry measurements of this flame identified a large range of oxygenated compounds between 150 u and 600 u[50], indicating that oxygenated species are an important contributor to the diversity of PACs.

Figure 2.3 shows the comparison in the oxygen-carbon ratio (O/C) between experimental and simulated compounds along the center streamline at different HABs. To provide a clearer comparison, a small number of PACs with more than 10 oxygen atoms and a larger number PACs without oxygen were omitted from the figure. This choice reflects the experimental difficulty of detecting signals from these compounds. On one hand, highly oxygenated species, consisting of molecules with 11-13 oxygen atoms, contained a large number of hydroperoxides (and derived groups) that are difficult to detect experimentally[63,64]. On the other hand, apolar PACs, like most of the species with O/C of zero, were not measured in the original experiment due to the ionization spray technique used.

With the exception of a few low mass, highly oxygenated species, the range of oxygenated species in mass and number of oxygen atoms is generally matched between 150 u and 450 u at HABs of 40, 50, and 60 mm. There are, however, some revealing discrepancies. The most noteworthy deviation is the under prediction of oxygen in my simulations at lower HABs. At an HAB of 30 mm (shown in supporting information of original work[1]) experiments observe a large number of diverse PACs

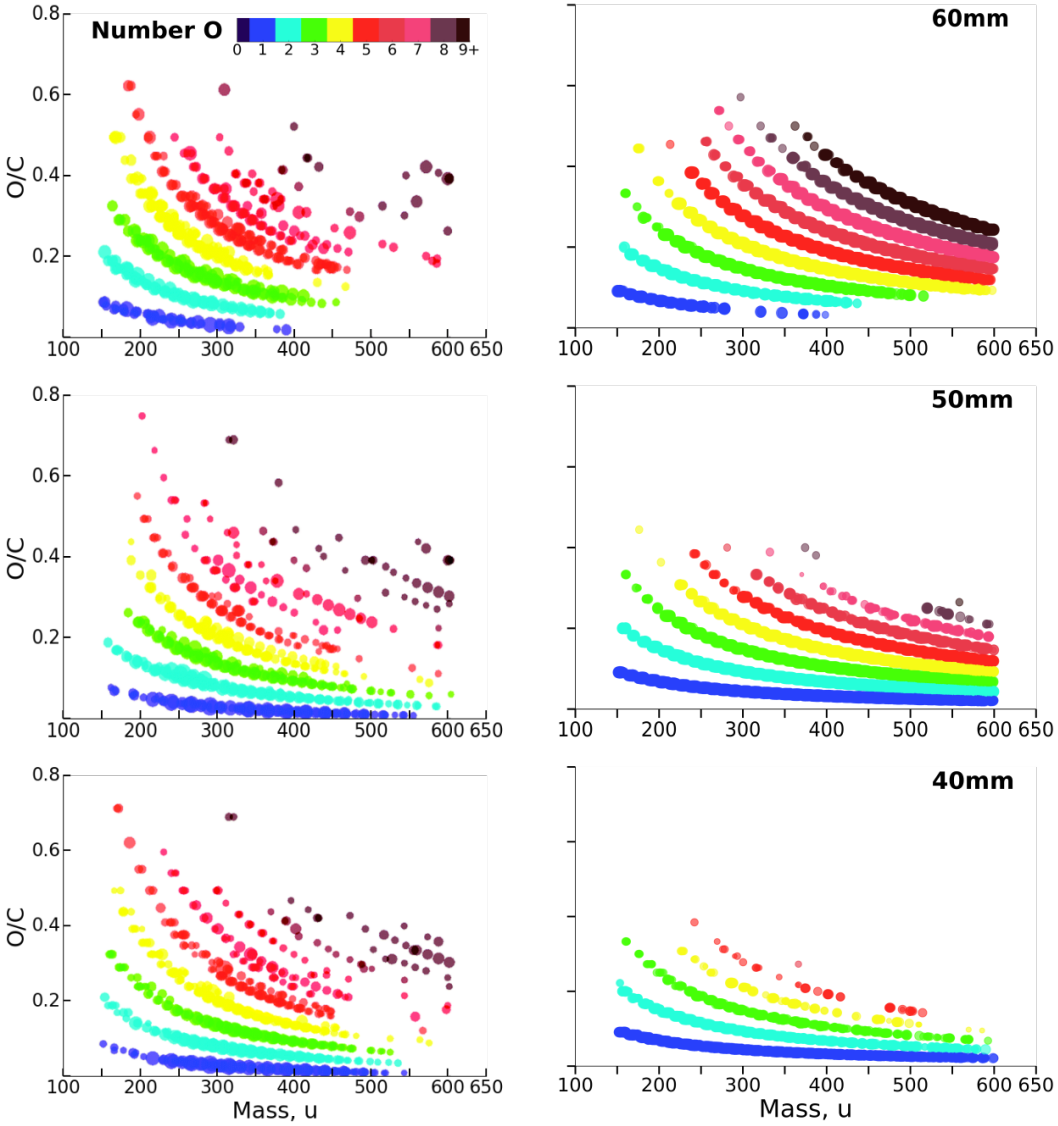

Figure 2.3: Oxygen-carbon ratio plotted against mass for different points along the center streamline. The experimental results are shown on the left and the simulations' results on the right. The size of the circles in each plot is proportional to the logarithm of species relative concentration. The figure is reproduced from Saldinger *et al.*[1] in which experimental results are reproduced/adapted from ref. Cain *et al.*[50] with permission from the PCCP Owner Societies.

up to 600 u, little growth is observed in the simulations. One possible reason for this discrepancy could be that there exists some low-temperature PAC oxygen chemistry[65]

not captured in the *SNapS2* kinetic mechanism. Prior to this region, the temperature along the centerline is relatively low and as such my simulations show only a few reactions mostly involving PACs with a single or no oxygen atoms. An alternate explanation is that the experimental measurements of the soot precursors do not accurately represent the true undisrupted gas-phase environment. The rapidly increased mass growth and oxygenation of up to 600 u and 6 oxygens seen between 30 mm and 40 mm demonstrates that the PACs' formation in this area is very spatially sensitive. Minor streamline disruptions during thermopheretic sampling[66] may be significant enough to bring about this discrepancy. Furthermore, although not as pronounced as in other flame configurations, studies of thermophoresis in coflow diffusion flames have identified complex two dimensional migration patterns which cause larger soot precursors to travel at a different rate and direction than the bulk gas-phase[67]. Thus, PACs thermally diffusing from higher HABs[68] or nearby streamlines may account for this difference. In addition, electron-spray ionization tends to produce results that are biased towards more oxygenated compounds when samples with lower O/C ratios are present[69] and as such it could be possible that these compounds are present in very small amounts but appear at high concentrations. Finally, while my results suggest the majority of these oxygenated compounds grow within the flame, it is not possible to rule out, as the authors of the experimental study acknowledge[50], that some growth may occur after experimental sampling[70].

Other deviations in my simulations are likely due to the fact *SNapS2* is designed to focus on PAC chemical growth. First, simulations fail to capture some low mass, highly oxygenated compounds. Unlike other differences, this issue is persistent at all heights in the flame. Such low masses and high O/C ratios suggest these molecules are most likely small acyclic or highly saturated molecules, in short not PACs. Figure 2.4 further supports that these molecules are likely not products of aromatic growth as the majority of species observed in experiment with an O/C ratio of greater than 0.5

have less than 15 carbons and an H/C ratio between 1 and 2.25 which falls above the PAC cata-condensed[71] limit. While *SNapS2* does include the addition of aliphatic chains and non-aromatic rings in the PAC growth mechanism, these low mass, non-aromatic molecules are not captured as my seed selection and mechanism emphasizes polycyclic aromatic growth. Second, for HABs of 50 mm and 60 mm, the model over-predicts high mass compounds (between 450 u and 600 u), especially with less than 5 oxygen atoms, that are not observed in experiment. This is most likely due to the removal from the gas-phase of heavier species through physical growth[72] or radical-radical combination reactions with other PACs[73]. The importance of radical-radical recombinations of PACs relative to other growth mechanisms has been observed[26] to become a significant factor in predicting mass growth at higher masses although eventually this is offset by the low concentrations of large PACs. Although the role of these reactions in this flame may differ, stochastic modeling of other flames[19] have found good agreement with experimental results up to approximately 500 u without including these radical-radical reactions. Physical growth is thought to significantly contribute to mass growth in coflow diffusion flames as masses approach 600 u[72]. This hypothesis is also sustained also by the observation that this discrepancy occurs primarily for PACs with low oxygen content, which is inline with the findings by Elvati *et al.*[27,74] suggesting that oxygen can inhibit physical dimerization.

The same agreement and discrepancies discussed above can be seen when looking at the degree of aromaticity, as illustrated in figure 2.4. Here, however, it is easier to see how the simulations capture well the trends pertaining to the polycondensed aromatic molecules. The simulations reproduce the ranges of H/C below 0.75 and O/C up to approximately 0.5 in the bottom cluster. As H/C increases, however, molecules become much more aliphatic in character and fall outside the scope of the *SNapS2* simulations. Also, as discussed above, at higher HABs my simulations over-predicts heavier PACs which accounts for the over prediction of high carbon number

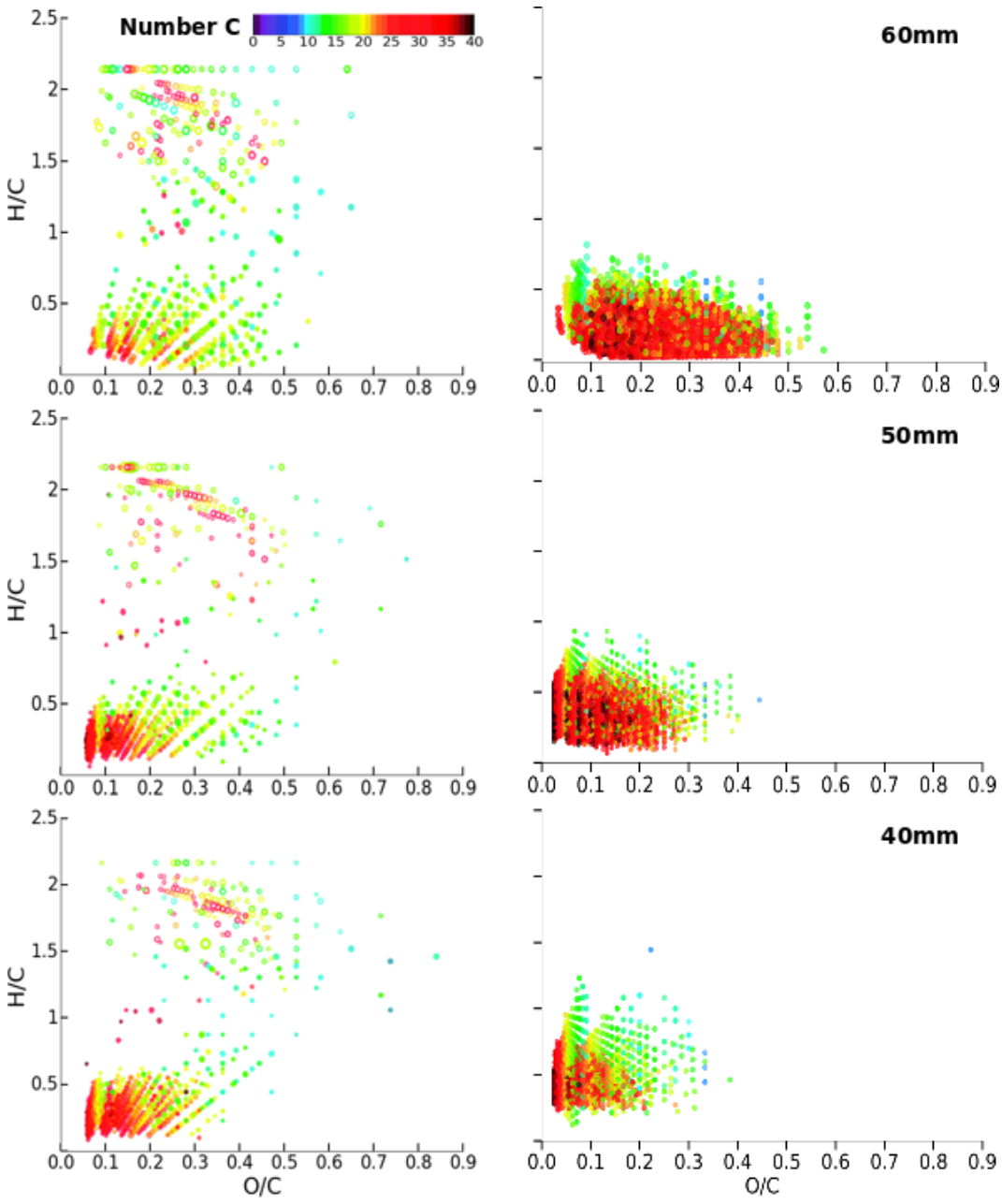

Figure 2.4: Hydrogen-carbon ratio plotted against oxygen-carbon ration for different points along the streamline. The experimental results are shown on the left and the simulations' results on the right. The size of the circles in each plot is proportional to the logarithm of species relative concentration. In the experimental plot, unfilled circles are non-aromatic hydrocarbons. The figure is reproduced from Saldinger *et al.*[1] in which experimental results are reproduced/adapted from ref. Cain *et al.*[50] with permission from the PCCP Owner Societies.

PACs seen at HABs of 50 mm and 60 mm.

Overall, despite the over-prediction of heavier PACs at higher HABs, the simulations reproduce the diversity of the experimentally observed oxygenated polyaromatic molecules. This, demonstrates how descriptors can be applied to characterize the PAC products of these chemical interactions.

2.4.2 Composition and size spatial dependence

After performing validation of the *SNapS2* simulations, I then applied a large number of descriptors at multiple different streamlines to provide a comprehensive spatial characterization of how PAC properties of interest develop in the flame. Of the six streamlines I considered, PAC growth was only observed on the inner three (labeled 1,2, and 3 in Fig. 2.1). This result can be explained by considering the acetylene concentration because only the three streamlines closest to the center pass through an acetylene rich region. Of note, the streamline number 4 goes through a region with a high concentration of radicals (mostly H, O, OH and CH_3) but I do not observe any relevant growth, showing again that acetylene is one of the most critical species for PAC growth[75,76]. Since only the three internal streamlines (1 to 3 or r_o=0.5, 2.5, and 5 mm) show significant growth, in the following I will focus only on those streamlines.

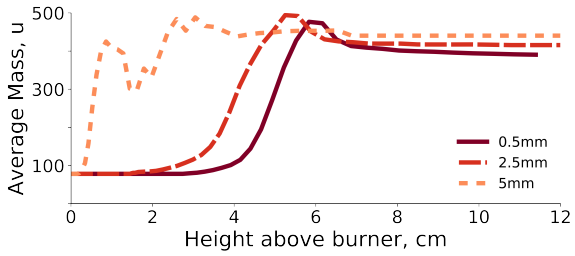

Figure 2.5: Average mass vs. HAB along the three streamlines.

Figure 2.5 reports the evolution of the average PAC mass along streamlines 1–3. As expected based on the lack of growth in the outer streamlines, PACs experience

their fastest growth when they pass through the high temperature, radical rich section in the flame. After traveling through this region, the gas-phase environment is less conducive to PAC growth as the average PAC mass remains the same or slightly decreases. As my model does not consider the aggregation of PACs, it is likely that at higher HABs the average mass of the PACs may be lower than the one predicted in Figure 2.5 as heavier PACs may aggregate into larger soot particles[50]. While mass provides a first characterization of PAC growth, in order to understand the diversity of the PACs formed in distinct regions, I subsequently study the molecular chemical characteristics, *i.e.* oxygen content, presence of five-membered rings, and aliphatic chains.

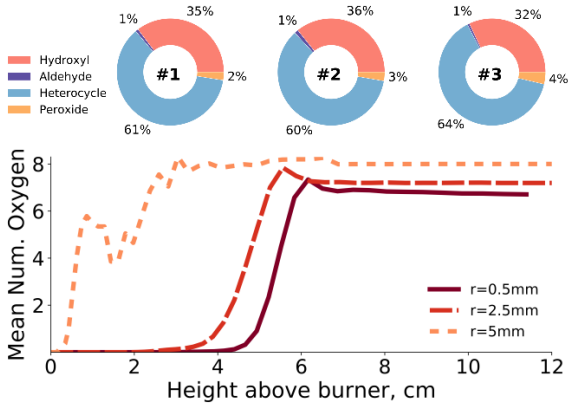

Figure 2.6: Oxygen atoms in the PACs' structures in each of the 3 inner streamlines: mean number of oxygen atoms per PAC at different HABs (bottom) and analysis of the distribution of oxygen containing functional groups near the peak oxygen content of each streamline (top). Functional groups were sampled over a range of 5 mm centered at 62 mm, 56 mm, and 30 mm for streamlines number 1, 2, and 3, respectively.

I first analyzed how oxygen content varies between and along different streamlines in Fig. 2.6. Streamline number 3 (r=5 mm) starts in a highly reactive region in the flame wing and begins growing immediately, while the other two exhibit minimal oxygenation prior to reaching a similarly reactive environment.

The most interesting aspect however, is the fact that the trend in the average oxygen content among the streamlines matches the one observed in the mass growth.

This is one additional proof of previous findings[26,43] which observed that the oxygenated species feature prominently in the early growth of PACs. In contrast to scenarios where PACs undergo oxygen addition reactions[43], here I do not observe separate regions favoring PAC growth through oxygen addition reactions and carbon addition reactions. In this flame, the gas-phase conditions where reactions involving the addition of carbon atoms and the addition of oxygen atoms are favorable, are the same.

The analysis of the types of oxygenated chemical groups shows a large prevalence of furans and hydroxyls, followed by small amounts of peroxides formed by the reactions with HO_2 radicals. Aldehydes make up an almost negligible percentage of the groups, which is consistent with another work[77] that suggests they play a larger role in the post-flame region. I observe only trace amounts of ethers. The comparison between the streamlines shows a slight increase of furan and peroxide groups when moving away from the centerline. I attribute the latter to a 10-times higher peak concentration of HO_2 and a lower temperature in the areas of high HO_2 concentration, which slows the peroxide decomposition rate[78].

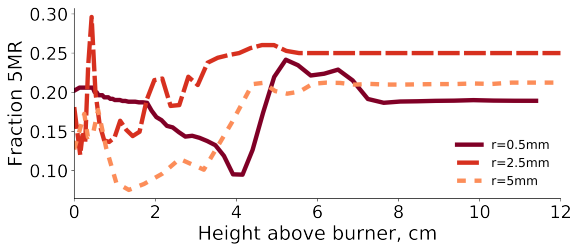

Figure 2.7: Fraction of five-membered carbocycles as a function of HAB.

Studies of many different flames[79–85] show that five-membered aromatic carbon rings are a common functional group in the PAC chemical space. Here, I study how the fraction of five-membered carbon rings, defined as the total number of five-membered carbon rings divided by the total number of five and six-membered carbocycles in the PAC ensemble, varies along and between streamlines. The analysis of

these streamlines confirms that five-membered carbon rings constitute an important structural feature of PACs in this flame as I observe a 10% to 30% fraction across all streamlines (Fig. 2.7). While in some measure, particularly at very low heights above the burner, the concentration of seed molecules affects the value (*e.g.* benzene *vs.* cyclopentadiene), the evolution of this fraction provides useful insights into the mechanisms of growth.

Early in all streamlines, the fraction of five-membered rings decreases as the PACs begin to grow. While the number of five-membered carbocycles is still increasing in this region (albeit slowly), the six-membered ring growth is occurring much more rapidly. This is consistent with known chemistry as six-membered rings can grow from a variety of pathways and active sites involving the edges of aromatic rings[14]. Later in the flame however, large PACs possess more zigzag sites[33] which have been shown to be particularly amenable[34] to the growth of five-membered carbon rings. Thus, while edge growth is always favorable for six-membered rings, five-membered carbon ring growth does not become favorable until larger PACs exist in the system and more of these zig-zag sites are available. Examining the effect of these five-membered rings on shape, figure 2.8 shows that compared to those with only six-membered rings, structures with five-membered rings can exhibit significant curvature although this is not necessarily the case.

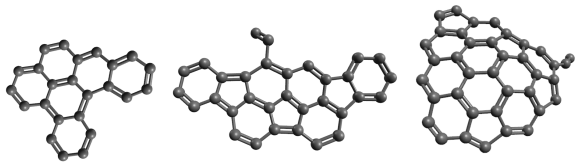

Figure 2.8: Examples of different ring motifs. Left: A planar PAC with only six-membered rings. Center: A planar PAC with five and six-membered rings. Right: A curved PAC with five and six-membered rings.

Numerous studies[47,86] suggest carbon chain growth occurs on PACs in flames. Thus, I conclude my PAC structural analysis describing the aliphatic components of

PACs (not to be confused with purely aliphatic molecules). The majority of carbons within my simulated PACs are members of rings or sp^2 hybridized, indicating there is very little non-aromatic branching. Still, notable exceptions do exist.

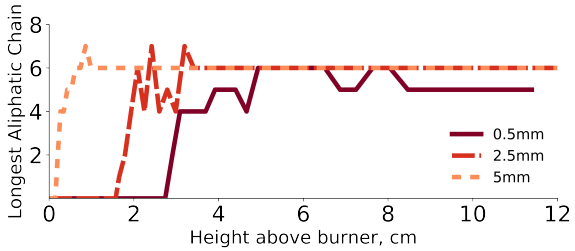

Figure 2.9: Length of longest aliphatic chain along each streamline.

Figure 2.9 shows the longest observed aliphatic chain on each streamline at each HAB. These carbon chains begin to be observed just as each streamline starts to enter the high temperature region of the flame. In these sections of the flame, the temperature is high enough for PAC growth reactions to occur, however, the streamline has yet to reach its maximum temperature. This behavior supports past work[87], which has hypothesized that aliphatic growth is most prevalent at elevated temperatures but not at the highest temperature section of the flame. As indicated in figure 2.9, the maximum chain length does not appreciably change after its initial growth even as the streamlines pass through the highest temperature, highest radical concentration section of the flame. This suggests that these side chains have a degree of stability and do not necessarily break down even in the most reactive environments of this system. The presence of aliphatic chains can be also dictated by some characteristics of the PACs' structures that are conducive to the growth of these chains. To test this hypothesis, I identified only those growth histories in which an aliphatic chain of at least five carbons is formed and analyzed the structure of the first PAC molecule in the history that had this aliphatic chain. Although I observe no unifying feature (two examples of the variety of structures can be seen in Fig. 2.10), I most commonly observe PACs between 30 and 35 carbons, including the aliphatic chain. Based on

some of the observed structures (such as Fig. 2.10) it appears many of the aliphatic chains form when a smaller carbon chain is unable to close into a ring. Thus, it follows that more aliphatic chains grow on PACs in this size range as they require a particular growth site which is more likely to exist on larger PACs.

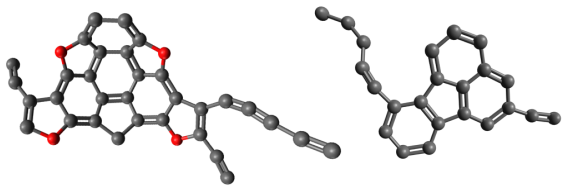

Figure 2.10: Two examples of observed molecules with aliphatic chains.

While these long hydrocarbon chains are not observed in high concentrations, numerous studies[23,24,88] have suggested these compounds can play a role in the physical growth of PACs into combustion nanoparticles. The early growth of these branched PACs and their long lifetimes within the flame suggest that despite being minor species they can still contribute to PAC aggregation, by stabilizing the formation of PACs aggregates long enough for them to form a chemical bond for example by radical-radical reactions[26,73].

Overall, these findings show how *kMC* can model PAC growth in a flame with both complex fuel and geometry to both reproduce experimental observations as well as provide a comprehensive radial and axial characterization of a number of different PAC descriptors including mass, oxygenation, five-membered rings, and aliphatic chains.

2.5 Distributional Descriptors Elucidate the Effect of Ethanol Doping on PAC Growth

The following results are adapted from my publication studying the effect of ethanol doping on the growth of PACs[3]. There are many studies on how the doping of fuels with oxygenates (*e.g.* methanol, ethanol, and methyl tertiary-butyl ether)

alters the chemistry of the gas-phase[45,54,89–91] and the characteristics of the resulting soot particles[92–95]. Here, I investigate the the degree to which this reduction of soot is caused by differences in the PAC chemical space by applying descriptors to distinguish the different contributions to PAC growth. I consider pure ethylene flames as well as ethylene-ethanol flames with different ethanol doping percentages. Notably, I find that insights are able to be gained not just by considering the mean values of these descriptors, but by considering how distributions of oxygenation descriptors evolve through the flame.

2.5.1 Differences Observed in PAC Chemical Space

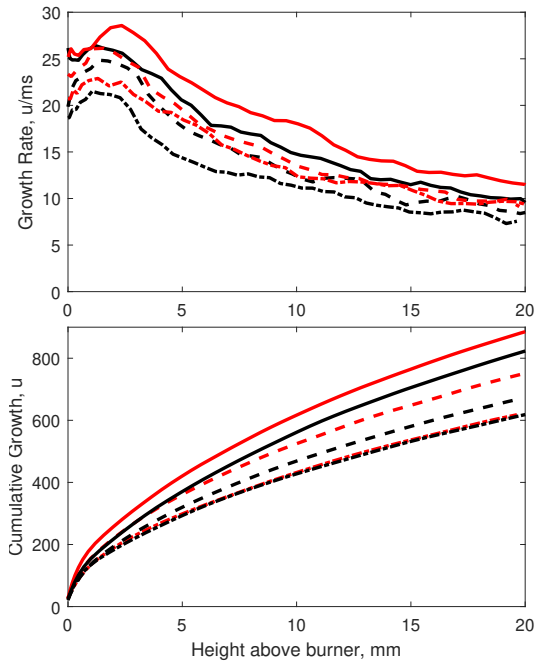

Figure 2.11: Average (upper panel) and cumulative (lower panel) chemical growth of PACs as a function of height above the burner from *SNapS2* simulations. Solid lines represent pure flames, dashed lines 20% doped flames, dash-dot line 40% doped flames; equivalence ratio is indicated by the color: black for 2.34 and red for 2.64.

First, the differences in PAC growth observed in ethylene flames with different levels of ethanol doping are considered in order to distinguish the effects of ethanol.

To this end, I compare six ethylene pre-mixed flames with three different ethanol doping fractions between 0-40 % and equivalence ratios of 2.34 and 2.64. PACs' chemical growth rate for the flames were taken by averaging the molecular growth of different traces starting at the same location, and the cumulative chemical growth, obtained by integrating the chemical growth rate from a HAB of 0 mm. The results for all the flames are shown in Fig. 2.11. Of note, different time intervals should be considered when integrating to compensate for differences in the gas flow rate. The plot shows a noticeable reduction in PACs' chemical growth when increasing the ethanol doping percentage and when decreasing the equivalence ratio. This is in agreement with the results of deterministic CHEMKIN simulations of PACs which show the same trend. However, *SNapS2* simulations indicate that PAC growth starts earlier than the deterministic gas-phase simulations due to the inclusion of reactions for the formation of PACs with oxygenated groups, as discussed further below.

To understand the contribution of oxygenated species to the growth of PACs, the percentage of PACs that contained at least one oxygen at different HABs was computed. To avoid biasing the results by including a large number of unstable species, all the PACs were weighted by their lifetime. Oxy-PACs were largely observed between 2 mm and 4 mm, which correspond to the region with the maximum growth rate and around the location where both O and H concentrations peak, as shown in figure 2.12. Collectively, these results indicate that a large portion of the initial growth is due to oxygenated PACs, and that is the reason why less early oxygen growth was observed in the deterministic simulations.

2.5.2 Distributional Descriptors Characterize PAC Growth

Statistical analysis of the reactions happening at different flame locations helps further elucidate the effects of oxygen on PACs' growth phenomena. Analysis of reactions suggests oxygen chemistry dominates around an HAB of 2 mm, while the

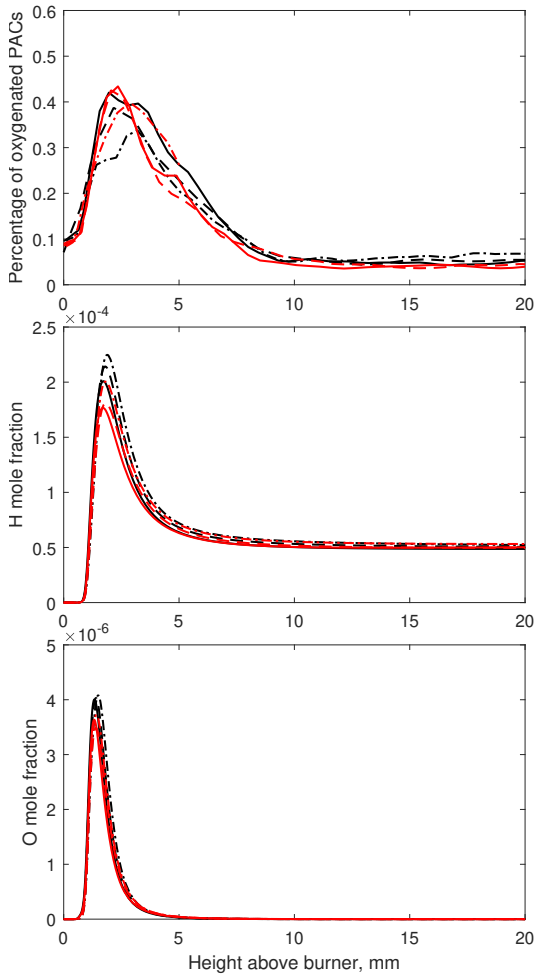

Figure 2.12: Comparison of PACs and gas-phase environment between different flames. Upper panel: percentage of oxy-PACs in different flames as a function of HAB from *SNapS2* simulations; Middle and bottom panel: the mole fraction profile for atomic hydrogen and atomic oxygen from gas-phase simulations correspondingly. Solid lines represent pure flames, dashed lines 20% doped flames, dash-dot line 40% doped flames; equivalence ratio is indicated by the color: black for 2.34 and red for 2.64.

major formation route for HACA is more relevant at higher heights such as 10 mm. At the same time, the rapid decomposition of the oxy-PACs indicate that these reactions provide only an initial increase in PAC size and that they are not responsible for the sustained growth of PACs. This effect can be quantified by examining the evolution of the mass distribution and its correlation with the oxygen content in PACs, particularly in the region where oxy-PACs are common (HAB up to $\sim$ 8 mm) since in this region

the PACs' masses begin to diverge.

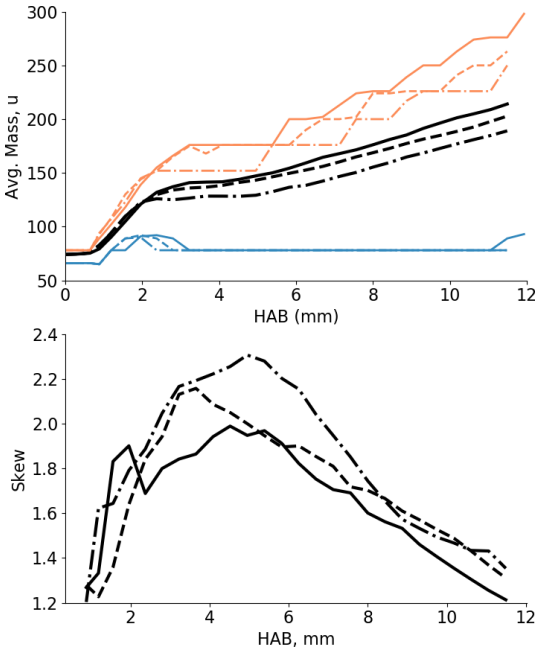

Figure 2.13: Evolution of PACs' mass as function of HAB. Only flames with $\phi = 2.34$ are shown for clarity although the trends are representative of both cases. Black solid line is used for the pure flame, dashed for the 20% doped flame, and dotted-dashed for the 40% doped flame. (Upper panel) Black lines show average mass; blue lines show the first mass quartile and the orange lines show the fourth mass quartile. (Lower panel) shows the skewness of mass distribution.

As shown in Fig. 2.13, the average mass starts increasing at the same time independently of the doping (1 mm to 2 mm) but it then plateaus differently. The mass increase corresponds to a broadening of the mass distribution, without a substantial change in the presence of species with lower masses, as indicated by a near constant first mass quartile (0th to 25th percentile). The positive skew in the mass distribution indicates that a few compounds create a leading tail, and interestingly these compounds have a higher oxygen content than the average compound in the same region as shown in Fig. 2.14.

This result suggests that as the percentage of oxy-PACs increases, early growth is sustained by oxygenated compounds in all flames. However, carbon growth mechanisms also play a role during and after this region. The concentrations of C_2H_2

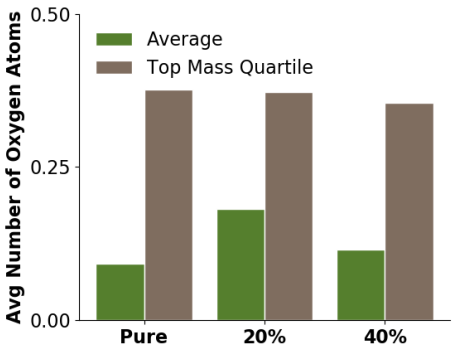

Figure 2.14: Average oxygen content of all PACs compared with high mass (fourth mass quartile) PACs. Values are calculated for flames with $\phi = 2.34$ ($\phi = 2.64$ exhibits similar behavior) and are weighted by species lifetime and concentration up to an HAB of 8 mm.

and C_2H_3 in the doped flames are significantly lower as hydrogen abstraction from C_2H_4 decreases due to competing H and OH radical reactions with ethanol[96]. These species are all important for HACA activity and their low concentration in the doped flames decrease C_2H_2 additions relative to the pure flames[46,97]. Thus, C_2H_2 additions contribute more prominently in undoped flames. Once the percentage of oxygenated PACs begins to decrease (2 mm to 8 mm) this disparity in C_2H_2 activity becomes more visible as C_2H_2 activity takes a more leading role relative to oxygen growth. Overall, these results indicate that the formation of oxy-PACs leads to a rapid increase in mass early in the flame, followed by a slower growth rate where PACs diverge in size as they are de-oxygenated and grow primarily through HACA. The distribution of the types of oxygenated groups (*i.e.* 15% furans, small fractions of ethers, and remaining hydoxyls) are similar among all the flames, but interestingly, the high mass PACs in Fig. 2.14 show a slightly larger percentage of heteroaromatic rings (approximately 20%-25%) than the total group of oxy-PACs at the expense of hydoxyl groups.

This analysis of the PACs' distributional properties highlights the contribution of oxygenated species in PAC growth both in ethylene and ethylene-ethanol flames. It suggests that subtle differences in these oxygenation profiles may explain the differing

mass trends observed at different levels of ethanol doping. Moreover, these findings underscore how considering the distribution of descriptors across the ensemble of PACs provides additional insights that may not necessarily be gained from an analysis of the mean properties.

2.6 Novels Descriptors of Curvature in Ethylene Pre-mixed Flame

The following results are adapted from my publication[4] comparing *SNapS2* results with experimentally observed AFM structures[47] in an ethylene pre-mixed flame. Among the features observed within both the simulated and experimental PACs are a number of different kinds of five-membered rings. In this section, I emphasize my contribution explaining why certain features were observed in *SNapS2* simulations but not experimentally. I first present a comparison between five-membered rings observed in *SNapS2* simulations and AFM results[47] and discuss how the observed agreement and disagreements are evidence of curved PACs being present in the flame. Curved PACs are an important property in combustion systems[79–85], however, providing a descriptor of this property is difficult. Thus, I develop a definition of curvature based on WHIM descriptors[57] and apply it to measure curvature in this flame.

2.6.1 Results Overview

First, comparisons are made between *SNapS2* simulations and AFM measurements of PACs from a previous study of an ethylene pre-mixed flame[47]. This previous study used AFM to gain molecular insights into the types of PACs which grow in a pre-mixed ethylene flame and emphasized that a wide range of nanostructures exist[47]. Similar to the AFM findings, *SNapS2* simulations are also able to model a large di-

versity of PACs in this flame. Figure 2.15 shows 32 examples of molecules sampled at an HAB of 8 mm. These molecules (which represent less then 0.005% of all the *SNapS2*-generated unique structures at this height) show the presence of oxy-PACs, condensed aromatics, curved molecules, different types of five-membered rings, PACs with only six-membered rings, as well as aliphatic structures. For comparison, the set of AFM observed PACs which show similar diversity are provided in the original experimental publication[47].

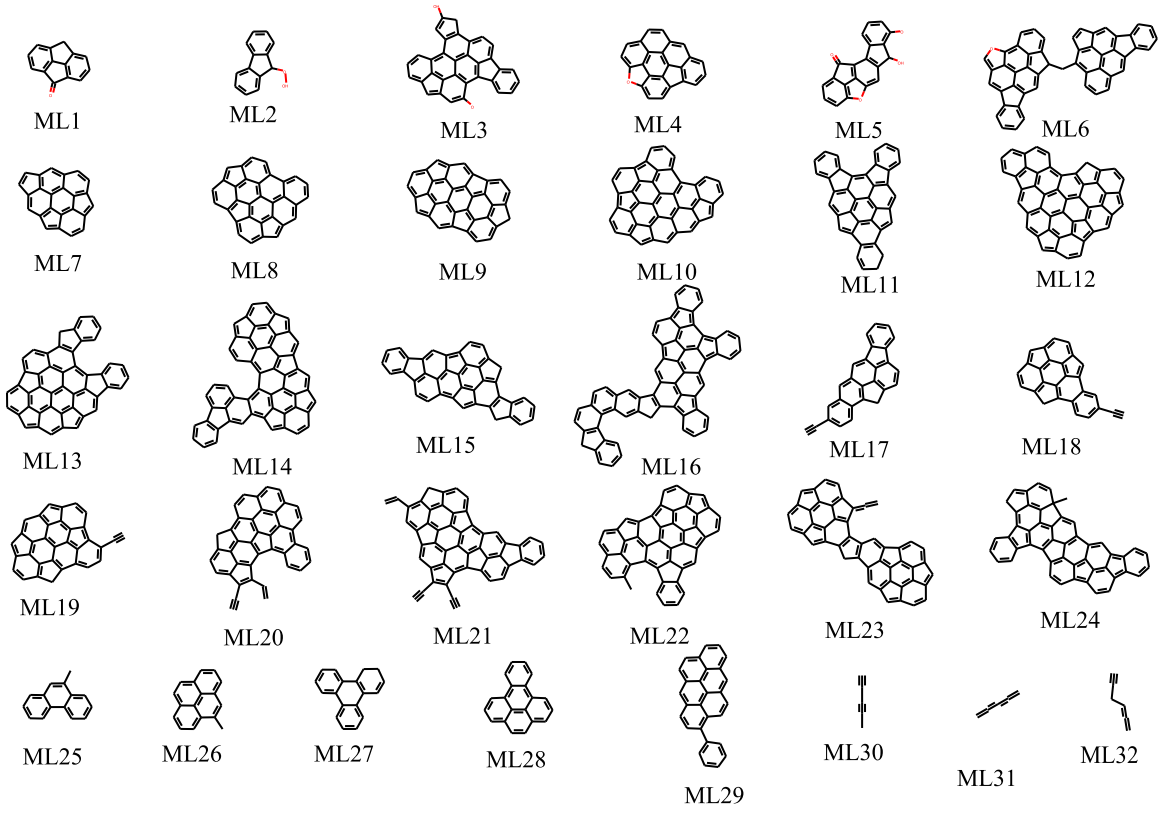

Figure 2.15: Example molecules simulated by *SNapS2* code at an HAB of 8 mm.

2.6.2 Ring Structures

One notable area of comparison are the ring structures which are observed in both the simulated and experimental results, particularly the five-membered rings. While a few example molecules containing only six-membered rings were observed (**ML25** to

ML29), most of the *SNapS2*-generated molecules contain at least one five-membered ring. For *SNapS2*-generated PACs, the number of five-membered rings is markedly lower than the number of six-membered rings, and the difference becomes larger at a higher number of carbon atoms. At the same time, the results indicate that at least a small number of five-membered rings are a general feature of the PACs. The *SNapS2* results show more five-membered ring structures than the AFM images, but this difference can partially arise from the experimental difficulty to sample or analyze curved molecules. Therefore, in the first comparison, molecules with embedded five-membered rings were removed from the data and only the number of non-embedded five-membered rings were evaluated from the remaining structures (Fig. 2.16).

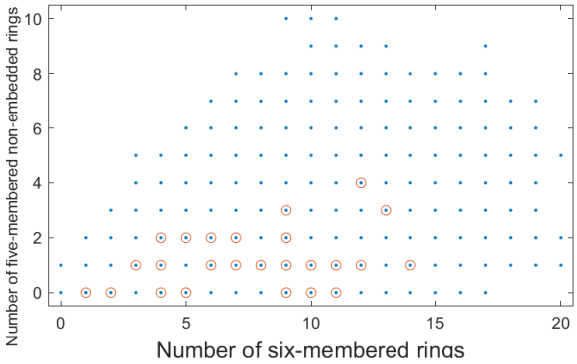

Figure 2.16: The number of non-embedded five-membered rings with respect to the number of six-membered rings for *SNapS2*-generated molecules (blue dots) and assigned PACs from AFM images (red circles)[47].

The comparison indicates that even considering only the non-embedded five-membered rings, *SNapS2*-generated molecules may still have more five-membered rings compared to experimental data. The most likely reason for the overestimation is a lack of five-member ring oxidation or migration pathways in the current available literature, which are included in *SNapS2*[35,98].

There are four types of five-membered rings identified by the AFM study, namely acenaphthylene-type, acenaphthene-type, fluorene-type, and indane-type. *SNapS2* simulations show examples of the first three types (**ML7** to **ML24**) but the *SNapS2*

kinetic mechanism currently does not contain pathways that will lead to the formation of indane-type five-membered rings. Notably, however, this comparison does not explain the embedded five-membered rings observed in *SNapS2* simulations but missing from the AFM results.

2.6.3 WHIM Descriptors

Since one of the main discrepancies between the previously discussed AFM and *SNapS2* results is the presence of embedded five-membered rings, in this section I introduce a descriptor of geometric curvature to assess how much of this difference can be attributed to molecular curvature. The WHIM descriptors[99] are a class of descriptor that provide a measure of size, shape, and atomic distribution of the molecule. To begin, a principal component analysis is performed on a covariance matrix based on the atomic coordinates. This can occur either with unweighted coordinates or coordinates weighted by an atomic property such as mass, polarizability, electronegativity, or Van der Waal's volume. For example, the mass-weighted WHIM descriptor corresponds roughly to the principal axes of inertia.

Given atomic coordinates $q_{i=1...N}$ and atomic property weights $w_{i=1...N}$, the covariance matrix S can be derived such that each element s_{jk} is defined by equation 2.2.

$$s_{jk} = \frac{\sum_{i=1}^{N} w_i (q_{ij} - \bar{q_j})(q_{ik} - \bar{q_k})}{\sum_{i=1}^{N} w_i} \tag{2.2}$$

A principal component analysis of this covariance matrix provides three principal components, $\lambda_{1,2,3}$, corresponding to the longest, second longest, and shortest principal component respectively. Based on these three principal components, a number of properties can be calculated such as length, area, volume, and skew. In the context of PACs, the relative length of the third mass-weighted WHIM axis (equation 2.3) is used as a measure of curvature. This third axis corresponds to the relative out-of-plane displacement. Mass weighting is used because otherwise hydrogen, which

has only minor effects on the curvature, would disproportionately contribute to the calculation. A completely flat circular molecule would have relative lengths of the first 2 axes equal to 0.5 and the relative length of the third axis (the aforementioned descriptor) equal to 0. A perfectly spherical molecule would have all three relative axis lengths equal to approximately 0.33.

$$curvature = \frac{\lambda_3}{\sum_{i=1}^{3} \lambda_i} \tag{2.3}$$

2.6.4 Curvatures of PACs

I apply this descriptor of curvature to the PACs observed in the *SNapS2* simulatuions. As mentioned above, I observe a large number of molecules, for example, **ML8**, **ML9**, **ML10**, **ML12**, **ML19**, **ML20**, **ML21**, and **ML22** which contain curvature. Since curved structures were difficult to sample or characterize in the cited AFM work, here I focus my discussion on the amount and frequency of the curvature I observe in *SNapS2* simulations. I quantify curvature using the previously discussed descriptor based around the shortest principal axis of inertia which is relatively insensitive to the presence of hydrogen and size independent. The results are shown in Fig. 2.17.

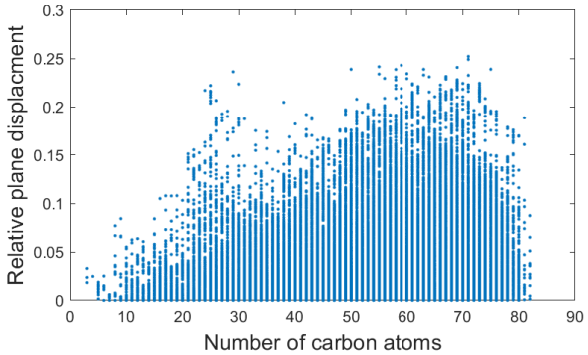

Figure 2.17: The relative plane displacement with respect to the number of carbon atoms.

The plot shows a large number of molecules with little to no curvature, namely less

than 0.05 (which is approximately the value caused by the presence of the aliphatic chain in ethylbenzene). A general increase in the curvature occurs with higher carbon numbers as more five-membered rings are added and additional growth occurs around previously embedded five-membered rings, while the drop above approximately 70 carbons is simply related to the simulation's upper mass limit. Interestingly, I can observe a peak for a set of molecules between 20 and 30 carbons, which is consistent with the number of carbons needed for a structure with embedded five-membered rings (such as **ML18**).

Overall this descriptor suggests that a likely reason for the discrepancy in the simulated five-membered rings and experimental observations is due to the curvature and that these embedded five-membered rings are a key contributor to the curvature of these PACs. Not only does this descriptor provide a novel and flexible means to quantify the curvature of nanoparticles, but these findings demonstrate how descriptors can be applied in tandem with experimental validation in order to explain discrepancies and understand growth phenomena.

2.7 Machine Learning Soot Inception Rate

The following results are adapted from my publication using machine learning (ML) and PAC chemical features to predict soot inception rates[5]. In previous sections, I have shown how to use chemical descriptors to characterize the development of PAC properties and interpret the chemical growth and nano-interactions of PACs in the flame. In this section, I discuss how numerical descriptors can quantitatively relate these PAC properties to nanoparticle formation rate by inputting these features into a machine learning model to predict the formation of soot and larger combustion nanoparticles. Since PACs are a critical component of soot formation, it is logical that predictions of soot formation can be improved by more accurately and comprehensively characterizing the chemical growth of PACs in flames and using chemical

descriptors relevant to their transition into larger nanoparticles.

2.7.1 Machine Learning Details

In this work, a ML framework was developed to predict the soot inception rate along the centerlines of premixed laminar ethylene flames in $g_{soot}/cm^3_{gas} \cdot s$. The inputs to the model are the temporal rates of change of the *SNapS2*-determined molecular properties which are thought to be relevant to PAC growth and to PAC-PAC interactions. The molecular properties considered were the molecular weight (MW), number of 6-membered rings (N6MR), number of 5-membered rings (N5MR), number of radical sites (NR), length of aliphatic chains (LC), number of rotatable bonds (NRB), number of OH groups (NOH), topological polar surface area (TPSA)[100], minimum partial charge (MINPC), and the maximum partial charge (MAXPC). Also, the PAC lifetime was considered as an input. Since lifetime is already a temporal variable, it was taken as-is (*i.e.* not converted into a temporal rate of change). *SNapS2*-predicted molecules were binned according to their position along the centerline, and the arithmetic mean (henceforth "average") and geometric mean of each molecular property was computed across each bin. Then, the temporal rates of change of the average and geometric mean of the properties were computed. Thus, the full list of inputs to the ML models is $\frac{d}{dt}$ of the arithmetic and geometric mean of each of the above properties plus the average and geometric mean of the PAC lifetime.

When testing and tuning the ML models, recursive feature elimination (RFE) was performed by a co-author Luke Di Liddo to determine the most accurate subset of input variables. RFE starts with all the input features, fits the data to a given ML algorithm, and successively removes the least important features (where "important" features are determined by importance scores that are returned by the ML algorithm) until a specified number of features (N) is reached. RFE was performed for values of N from the maximum down to 1 and the accuracy of each regression algorithm

(with each value of N) was evaluated using repeated 3-fold cross validation, which is a way to measure the robustness of the algorithm. Repeated 3-fold cross validation randomly splits the data into 3 folds, uses each fold as the testing data once, reports the model's accuracy for each test fold, and then repeats the procedure for different randomly selected folds.

A variety of ML algorithms were tested, including XGBoost regression[101], kernel ridge regression (with linear and polynomial kernels), support vector regression (also with linear and polynomial kernels), ordinary least squares linear regression, multivariate adaptive regression splines[102], and random forest regression.

2.7.2 Descriptors and Relationship to Sooting Rate

When considering the rates of change of the average and geometric mean of the molecular properties, and the average and geometric mean of PAC lifetime, there are 22 possible input features to a ML model. However, different subsets of input features may give different model predictions and accuracy, and not every input is strongly related to the output. To navigate these considerations, RFE was used to find the optimal set of input features for the different regression algorithms tested (listed above). The model with the lowest mean squared error and absolute error used the following four inputs: $d(TPSA_{avg})/dt$, $d(LC_{avg})/dt$, $d(MW_{g.mean})/dt$ and $lifetime_{g.mean}$. As seen in Tab. 2.1, each of these inputs have strong individual correlations with the inception rate. Also, each property is physically distinct from the others, implying there may be little multicollinearity among these 4 inputs.

In addition to RFE, the Pearson and Spearman correlation coefficients were computed between each input feature and the output in order to discern the strength of the individual relationships between each input and the output. The Pearson coefficient is a measure of linear correlation between two variables, and the Spearman coefficient describes the strength of the monotonic relationship between two variables. Table 2.1

shows the correlation coefficients for the most highly correlated input features (Pearson coefficient > 0.6). Table 2.1 shows that a number of the input variables have strong individual correlations with the inception rate (Spearman coefficients $>\sim 0.7$). In particular, the TPSA and NRB are interesting features. The TSPA is a measure of the oxygen content related to the PAC shape (total surface over all polar atoms), and NRB describes both the aliphatic nature of the PAC and how it may distribute rotational energy upon collision. Both are examples of molecular properties that are only available from detailed modelling methods like *SNapS2*. Furthermore, the strong correlations show that these types of detailed *SNapS2*-determined molecular properties may be relevant to various soot processes and that further exploration between stochastic molecular modelling, CFD, and AI may be fruitful. Finally, while NOH and NR also have high correlation coefficients, they were not part of the subset of inputs that resulted in the best performing model, potentially due to multicollinearity.

Table 2.1: Pearson and Spearman correlation coefficients between inputs and the soot inception rate for the most strongly correlated inputs. Bold entries represent the inputs that were used in the final model (as selected by RFE).

Input Feature	Pearson	Spearman
$\boldsymbol{d(TPSA_{avg})/dt}$	**0.82**	**0.78**
$d(NOH_{avg})/dt$	0.72	0.73
$\boldsymbol{d(LC_{avg})/dt}$	**0.69**	**0.61**
$\boldsymbol{d(MW_{g.mean})/dt}$	**0.68**	**0.69**
$d(N5MR_{avg})/dt$	0.63	0.61
$d(NRB_{avg})/dt$	0.60	0.49
$\boldsymbol{Lifetime_{g.mean}}$	**-0.70**	**-0.70**
$d(NR_{avg})/dt$	-0.78	-0.82

Kernel ridge regression with a linear kernel and an α value (regularization strength) of 0.85 was the model that resulted in the lowest errors and is henceforth the selected model. The accuracy metrics for that model are reported in Tab. 2.2, where $RMSE$ is the root mean squared error, MAE is the mean absolute error, and R^2 is the coefficient of determination (the proportion of variance in the dependent variable that can be predicted by the independent variables). Repeated 3-fold cross validation was used

to compute these accuracy metrics. Since that technique repeatedly re-samples and re-trains the model, there are different values for $RMSE$, MAE, and R^2 when each fold is used as the testing data, and as such, both the mean and standard deviation are reported. The Supplementary Material of the corresponding publication[5] show the mean MAE and $RMSE$ of the various other other ML algorithms tested.

Table 2.2: Error metrics for the final soot inception prediction model computed using repeated 3-fold cross validation.

Metric	Mean	Standard Deviation
$RMSE$	6.73×10^{-12}	1.52×10^{-12}
MAE	5.77×10^{-12}	1.64×10^{-12}
R^2	0.71	0.12

The final model has a MAE of 5.77×10^{-12} and a $RMSE$ of 6.73×10^{-12}. For reference, the mean value of the targets (soot inception rate in $g/cm^3 \cdot s$) is 2.26×10^{-11}, signifying that the absolute error in the model predictions is approximately one quarter of the magnitude of the target values.

Figure 2.18 shows how the final model's predictions align with the target soot inception rates[103]. The predicted values from the final ML model follow the experimental trend closely and without significant deviation from the dashed line, and almost every predicted value is within the experimental uncertainty of the target values. Future improvements to this procedure may come in the expansion of the data set, *e.g.* applying Equation S1.1 originally from[103] to other flames that measured the diameter of primary particles, d_p (from which n_p and, subsequently, the inception rate are derived[103]). Additional improvements may come through the consideration of more molecular properties or by more extensive feature engineering.

2.7.3 Comparisons to other potential methods

To assess the effectiveness of the current procedure, the predictions from the final *SNapS2*-informed ML model were compared to the predictions from existing state-

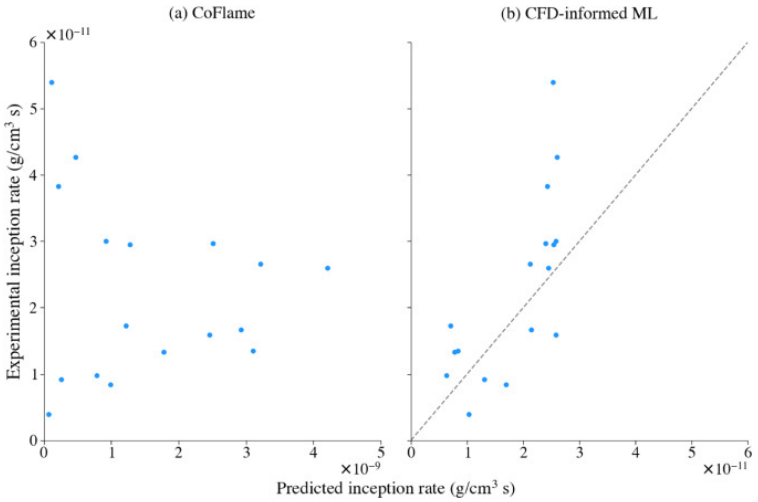

Figure 2.18: The experimentally-derived soot inception rates[103] plotted against the final ML model's predicted soot inception rates. The dashed line represents the line where the model's predicted values equal the target values.

of-the-art methods. The first method was the physics inspired CoFlame inception predictions and the second was a ML model built using CFD data (temperature, concentrations of key stabilomer PACs) as inputs (and with the same target data from Xu *et al.*[103]).

CoFlame simulations were performed by co-author Luke Di Liddo for each of the three flames and validated against the experimental data for f_v, d_p, and n_p. CoFlame's predicted inception rates were approximately 1 order of magnitude higher than the experimentally-derived rates. Besides being an order of magnitude too high, CoFlame's predicted inception rates are also more scattered compared to the final model in the present work. Despite the inaccurate inception rate predictions, f_v was still predicted accurately due to compensation from the other soot sub-models (*e.g.* surface growth, condensation, coagulation, or oxidation). While still a relatively advanced inception model, the inaccurate inception rate predictions from CoFlame, viewed in light of the accurate predictions from the *SNapS2*-informed ML model, suggest that a source of potential improvement may be the consideration of a more diverse PAC landscape.

The ML model was trained using the same approach as discussed above. However, this time, instead of using *SNapS2* PAC descriptors, the method used CFD-determined properties, namely temperature, average concentration of groups of peri-condensed PACs (A1 to A3, A4 to A5, A6 to A7), PACs with 5-membered rings (A2R5, A3R5, A4R5), groups of radical PACs (A1- to A3-, A4- to A5-, and A6-), and all radical PACs (A1- to A6-). The best model (using ordinary least squares linear regression) has an MAE and $RMSE$ approximately 57% and 65% higher, respectively, than the model trained on *SNapS2* data, suggesting that the nuanced molecular information may be an important aspect to modelling and understanding soot inception.

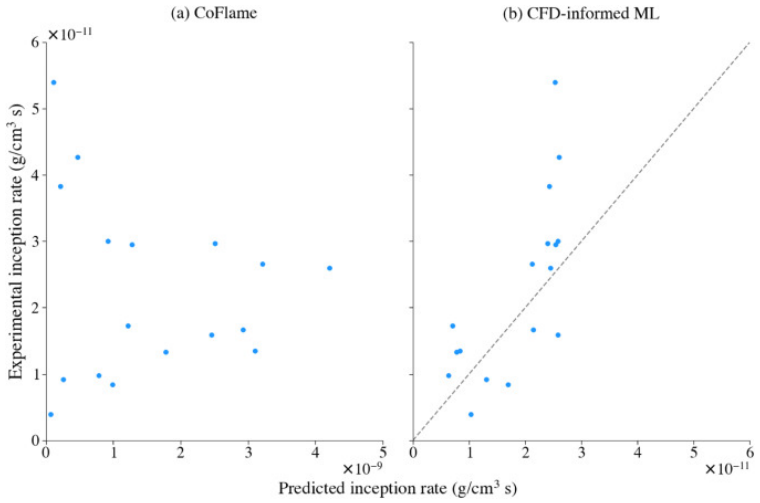

Figure 2.19: Predictions of the soot inception rate from the ML model using CFD (CoFlame[51]) and CFD-inputs compared to the experimentally-derived soot inception rates of[103]. Note, parity line not shown in (a) due to predicted values being 2 orders of magnitude higher than experimental ones.

Overall, these findings emphasize that properly representing PAC chemical space is important to modeling the growth of larger nanoparticles. A machine learning approach incorporating *SNapS2* molecules and a diverse set of molecular descriptors significantly outperforms existing methods which oversimplify the PAC chemical space. Moreover this work suggests that when applying the proper molecular descrip-

tors, machine learning provides a means to quantitatively relate PAC properties to more complex multi-scale phenomena which are difficult to directly model.

2.8 References

[1] Jacob C. Saldinger, Qi Wang, Paolo Elvati, and Angela Violi. Characterizing the diversity of aromatics in a coflow diffusion Jet A-1 surrogate flame. *Fuel*, 268:117198, 2020.

[2] Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Stochastic and network analysis of polycyclic aromatic growth in a coflow diffusion flame. *Phys. Chem. Chem. Phys.*, 23:4326–4333, 2021.

[3] Qi Wang, Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Insights on the effect of ethanol on the formation of aromatics. *Fuel*, 264:116773, 2020.

[4] Qi Wang, Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Molecular structures in flames: A comparison between snaps2 and recent afm results. *Proc. Combust. Inst.*, 38(1):1133–1141, 2021.

[5] Luke Di Liddo, Jacob C. Saldinger, Mehdi Jadidi, Paolo Elvati, Angela Violi, and Seth B. Dworkin. Exploring soot inception rate with stochastic modelling and machine learning. *Combustion and Flame (in press)*, page 112375, 2022.

[6] Charles Bauschlicher and Alessandra Ricca. Mechanisms for polycyclic aromatic hydrocarbon (PAH) growth. *Chemical Physics Letters*, 326(3):283–287, 2000.

[7] Irvin Glassman. Soot formation in combustion processes. *Symp. Combust.*, 22(1):295–311, 1989.

[8] Yanxu Zhang, Shu Tao, Huizhong Shen, and Jianmin Ma. Inhalation exposure to ambient polycyclic aromatic hydrocarbons and lung cancer risk of Chinese population. *Proc. Natl. Acad. Sci.*, 106(50):21063–21067, 2009.

[9] Andrea D'Anna. Combustion-formed nanoparticles. *Proceedings of the Combustion Institute*, 32(1):593–613, 2009.

[10] Rajesh Koirala, Sotiris E. Pratsinis, and Alfons Baiker. Synthesis of catalytic materials in flames: opportunities and challenges. *Chem. Soc. Rev.*, 45:3053–3068, 2016.

[11] Raul Serrano-Bayona, Carson Chu, Peng Liu, and William L. Roberts. Flame synthesis of carbon and metal-oxide nanoparticles: Flame types, effects of combustion parameters on properties and measurement methods. *Materials*, 16(3):1192, 2023.

[12] Changran Liu, Ajay V. Singh, Chiara Saggese, Quanxi Tang, Dongping Chen, Kevin Wan, Marianna Vinciguerra, Mario Commodo, Gianluigi De Falco, Patrizia Minutolo, Andrea D'Anna, and Hai Wang. Flame-formed carbon nanoparticles exhibit quantum dot behaviors. *Proceedings of the National Academy of Sciences*, 116(26):12692–12697, 2019.

[13] B.D. Crittenden and Ronald Long. Formation of Polycyclic Aromatics in Rich Premixed Acetylene and Ethylene. *Combust. Flame*, 20:359–368, 1973.

[14] Henning Richter and Jack B. Howard. Formation of polycyclic aromatic hydrocarbons and their growth to soot–a review of chemical reaction pathways. *Prog. Energ. Combust.*, 26(4-6):565–608, 2000.

[15] Jorg Appel, Henning Bockhorn, and Michael Frenklach. Kinetic Modeling of Soot Formation with Detailed Chemistry and Physics: Laminar Premixed Flames of C2 Hydrocarbons. *Combust. Flame*, 121(1–2):121–136, 2000.

[16] Angela Violi, Gregory A. Voth, and Adel F. Sarofim. The relative roles of acetylene and aromatic precursors during soot particle inception. *Proc. Combust. Inst.*, 30(1):1343–1351, 2005.

[17] Nazly E. Sánchez, Alicia Callejas, Angela Millera, Rafael Bilbao, and Maria U. Alzueta. Polycyclic Aromatic Hydrocarbon (PAH) and Soot Formation in the Pyrolysis of Acetylene and Ethylene: Effect of the Reaction Temperature. *Energy & Fuels*, 26(8):4823–4829, 2012.

[18] Peng Liu, Zepeng Li, Anthony Bennett, and et. al. The site effect on PAHs formation in HACA-based mass growth process. *Combust. Flame*, 199:54–68, 2019.

[19] Qi Wang, Paolo Elvati, Doohyun Kim, K. Olaf Johansson, Paul E. Schrader, Hope A. Michelsen, and Anegla Violi. Spatial dependence of the growth of polycyclic aromatic compounds in an ethylene counterflow flame. *Carbon*, 149:328–335, 2019.

[20] K. Olof Johansson, T. Dillstrom, M. Monti, F. El Gabaly, M.F. Campbell, P.E. Schrader, D.M. Popolan-Vaida, N.K. Richards-Henderson, K.R. Wilson, A. Violi, and H.A. Michelsen. Formation and emission of large furans and oxygenated hydrocarbons from flames. *Proc. Natl. Acad. Sci.*, 113(30):8374–8379, 2016.

[21] Stephen E. Stein and Askar Fahr. High-temperature stabilities of hydrocarbons. *J. Phys. Chem.*, 89(17):3714–3725, 1985.

[22] Seung Hyun Chung and Angela Violi. Nucleation of fullerenes as a model for examining the formation of soot. *The Journal of Chemical Physics*, 132(17):174502–174502–9, 2010.

[23] Seung Hyun Chung and Angela Violi. Peri-condensed aromatics with aliphatic chains as key intermediates for the nucleation of aromatic hydrocarbons. *Proc. Combust. Inst.*, 33(1):693–700, 2011.

[24] Paolo Elvati and Angela Violi. Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains. *Proceedings of the Combustion Institute*, 34(1):1837–1843, 2013.

[25] Qian Mao, Adri C.T. van Duin, and Kai H. Luo. Formation of incipient soot particles from polycyclic aromatic hydrocarbons: A ReaxFF molecular dynamics study. *Carbon*, 121:380–388, 2017.

[26] Paolo Elvati, V. Tyler Dillstrom, and Angela Violi. Oxygen driven soot formation. *Proceedings of the Combustion Institute*, 36(1):825–832, 2017.

[27] Paolo Elvati, Kirk Turrentine, and Angela Violi. The role of molecular properties on the dimerization of aromatic compounds. *Proceedings of the Combustion Institute*, 37(1):1099–1105, 2019.

[28] C. Achten and J.T. Andersson. Overview of Polycyclic Aromatic Compounds (PAC). *Polycyclic Aromatic Compounds*, 35(2-4):177–186, 2015.

[29] Kristen A. Russ, Paolo Elvati, Tina L. Parsonage, Alyssa Dews, James A. Jarvis, M. Ray, B. Schneider, P.J.S. Smith, P.T.F. Williamson, Angela Violi, and Martin A. Philbert. C60 fullerene localization and membrane interactions in raw 264.7 immortalized mouse macrophages. *Nanoscale*, 8:4134–4144, 2016.

[30] Wentao Wang, Narumol Jariyasopit, Jill Schrlau, Yuling Jia, Shu Tao, Tian-Wei. Yu, Roderick H. Dashwood, Wei Zhang, Xuejun Wang, and Staci L.M. Simonich. Concentration and Photochemistry of PAHs, NPAHs, and OPAHs and Toxicity of pm2.5 during the Beijing Olympic Games. *Environ. Sci. Technol.*, 45(16):6887–6895, 2011.

[31] Yichun Wang, Usha Kadiyala, Qu Zhibei, Paolo Elvati, Christopher Altheim, Nicholas A. Kotov, Angela Violi, and J. Scott VanEpps. Anti-biofilm activity of graphene quantum dots via self-assembly with bacterial amyloid proteins. *J. Phys. Chem. A*, 13(4):4278–4289, 2019.

[32] Yu Wang, Abhijeet Raj, and Suk Ho Chung. A PAH growth mechanism and synergistic effect on PAH formation in counterflow diffusion flames. *Combustion and Flame*, 160(9):1667–1676, 2013.

[33] Michael Frenklach. On surface growth mechanism of soot particles. *Symp. Combust.*, 26(2):2285–2293, 1996.

[34] Russell Whitesides and Michael Frenklach. Detailed Kinetic Monte Carlo Simulations of Graphene-Edge Growth. *The Journal of Physical Chemistry A*, 114(2):689–703, 2010.

[35] Michael Frenklach, Zhenyuan Liu, Ravi I. Singh, Galiya R. Galimova, Valeriy N. Azyazov, and Alexander M. Mebel. Detailed, sterically-resolved modeling of soot oxidation: Role of O atoms, interplay with particle nanostructure, and emergence of inner particle burning. *Combust. Flame*, 188:284–306, 2018.

[36] Angela Violi. Modeling of soot particle inception in aromatic and aliphatic premixed flames. *Combust. Flame*, 139(4):279–287, 2004.

[37] Abhijeet Raj, Matthew Celnik, Raphael Shirley, Markus Sander, Robert Patterson, Richard West, and Markus Kraft. A statistical approach to develop a detailed soot growth model using PAH characteristics. *Combust. Flame*, 156(4):896–913, 2009.

[38] Matthew Celnik, Abhijeet Raj, Richard West, Robert Patterson, and Markus Kraft. Aromatic site description of soot particles. *Combust. Flame*, 155(1-2):161–180, 2008.

[39] Markus Sander, Robert Patterson, Andreas Braumann, Abhijeet Raj, and Markus Kraft. Developing the PAH-PP soot particle model using process informatics and uncertainty propagation. *Proc. Combust. Inst.*, 33(1):675–683, 2011.

[40] Jason Y. W. Lai, Paolo Elvati, and Angela Violi. Stochastic atomistic simulation of polycyclic aromatic hydrocarbon growth in combustion. *Phys. Chem. Chem. Phys.*, 16(17):7969, 2014.

[41] Tyler Dillstrom and Angela Violi. The effect of reaction mechanisms on the formation of soot precursors in flames. *Combust. Theor. Model.*, 21(1):23–34, 2017.

[42] K. Olof Johansson, Jason Y.W. Lai, S.A. Skeen, D.M. Popolan-Vaida, K.R. Wilson, N. Hansen, A. Violi, and H.A. Michelsen. Soot precursor formation and limitations of the stabilomer grid. *Proc. Combust. Inst.*, 35(2):1819–1826, 2015.

[43] Qi Wang, Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Insights on the effect of ethanol on the formation of aromatics. *Fuel*, 264:116774, 2019.

[44] Philippe Dagaut and Sandro Gaïl. Chemical Kinetic Study of the Effect of a Biofuel Additive on Jet-A1 Combustion. *J. Phys. Chem. A*, 111(19):3992–4000, 2007.

[45] Ilya E Gerasimov, Denis A Knyazkov, Sergey A Yakimov, Tatyana A Bolshova, Andrey G Shmakov, and Oleg P Korobeinichev. Structure of atmospheric-pressure fuel-rich premixed ethylene flame with and without ethanol. *Combust. Flame*, 159(5):1840–1850, 2012.

[46] Fusheng Xu, Peter B. Sunderland, and Gerard M. Faeth. Soot formation in laminar premixed ethylene/air flames at atmospheric pressure. *Combustion and Flame*, 108(4):471 – 493, 1997.

[47] Mario Commodo, Katharina Kaiser, Gianluigi De Falco, Patrizia Minutolo, Fabian Schulz, Andrea D'Anna, and Leo Gross. On the early stages of soot formation: Molecular structure elucidation by high-resolution atomic force microscopy. *Combust. Flame*, 205:154–164, 2019.

[48] Meghdad Saffaripour, Mohammadreza R. Kholghy, Seth B. Dworkin, and Murray J. Thomson. A numerical and experimental study of soot formation in a laminar coflow diffusion flame of a Jet A-1 surrogate. *Proc. Combust. Inst.*, 34(1):1057–1065, 2013.

[49] Mohammadreza R. Kholghy, Meghdad Saffaripour, Christopher Yip, and Murray J. Thomson. The evolution of soot morphology in a laminar coflow diffusion flame of a surrogate for Jet A-1. *Combust. Flame*, 160(10):2119–2130, 2013.

[50] Jeremy Cain, Alexander Laskin, Mohammad Reza Kholghy, Murray J Thomson, and Hai Wang. Molecular characterization of organic content of soot along the centerline of a coflow diffusion flame. *Phys. Chem. Chem. Phys.*, 16(47):25862–25875, 2014.

[51] Nick A Eaves, Qingan Zhang, Fengshan Liu, Hongsheng Guo, Seth B Dworkin, and Murray J Thomson. CoFlame: A refined and validated numerical algorithm for modeling sooting laminar coflow diffusion flames. *Computer Physics Communications*, 207:464–477, 2016.

[52] Reaction Design. *CHEMKIN-PRO 15112*. San Diego, 2011.

[53] S. Mani Sarathy, Goutham Kukkadapu, Marco Mehl, Tamour Javed, Ahfaz Ahmed, Nimal Naser, Aniket Tekawade, Graham Kosiba, Mohammed AlAbbad, Eshan Singh, Sungwoo Park, Mariam Al Rashidi, Suk Ho Chung, William L Roberts, Matthew A Oehlschlaeger, Chih-Jen Sung, and Aamir Farooq. Compositional effects on the ignition of FACE gasolines. *Combust. Flame*, 169:171–193, 2016.

[54] Djemaa Golea, Yacine Rezgui, Miloud Guemini, and Soumia Hamdane. Reduction of PAH and soot precursors in benzene flames by addition of ethanol. *J. Phys. Chem. A*, 116(14):3625–3642, 2012.

[55] Jennifer D Herdman and J Houston Miller. Intermolecular Potential Calculations for Polynuclear Aromatic Hydrocarbon Clusters. *J. Phys. Chem. A*, 112(28):6249–6256, 2008.

[56] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988.

[57] Roberto Todeschini and Paola Gramatica. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR and QSAR in Environmental Research*, 7(1-4):89–115, 1997.

[58] David T. Stanton and Peter C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, 1990.

[59] Paul Labute. A widely applicable set of descriptors. *J. Mol. Graph*, 18(4-5):464–477, 2000.

[60] Xiliang Yan, Alexander Sedykh, Wenyi Wang, Xiaoli Zhao, Bing Yan, and Hao Zhu. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale*, 11(17):8352–8362, 2019.

[61] Thomas A. Halgren. Merck molecular force field. ii. mmff94 van der waals and electrostatic parameters for intermolecular interactions. *Journal of Computational Chemistry*, 17(5-6):520–552, 1996.

[62] Greg Landrum. Rdkit: Open-source cheminformatics.

[63] Oliver Welz, John D. Savee, David L. Osborn, Subith S. Vasu, Carl J. Percival, Dudley E. Shallcross, and Craig A. Taatjes. Direct Kinetic Measurements of Criegee Intermediate (CH2oo) Formed by Reaction of CH2i with O2. *Science*, 335(6065):204–207, 2012.

[64] Judit Zádor, Haifeng Huang, Oliver Welz, Johan Zetterberg, David L. Osborn, and Craig A. Taatjes. Directly measuring reaction kinetics of ˙QOOH – a crucial but elusive intermediate in hydrocarbon autoignition. *Phys. Chem. Chem. Phys.*, 15(26):10753–10760, 2013.

[65] Zhandong Wang, Oliver Herbinet, Nils Hansen, and Frederique Battin-Leclerc. Exploring hydroperoxides in combustion: History, recent advances and perspectives. *Progress in Energy and Combustion Science*, 73:132–181, 2019.

[66] Chiara Saggese, Alberto Cuoci, Alessio Frassoldati, Sara Ferrario, Joaquin Camacho, Hai Wang, and Tiziano Faravelli. Probe effects in soot sampling from a burner-stabilized stagnation flame. *Combust. Flame*, 167:184–197, 2016.

[67] K.T. Kang, J.Y. Hwang, S.H. Chung, and W. Lee. Soot zone structure and sooting limit in diffusion flames: Comparison of counterflow and co-flow flames. *Combustion and Flame*, 109(1):266 – 281, 1997.

[68] Hongsheng Guo, Fengshan Liu, Gregory L. Smallwood, and Omer L. Gulder. A numerical investigation of thermal diffusion influence on soot formation in ethylene/air diffusion flames. *International Journal of Computational Fluid Dynamics*, 18(2):139–151, 2004.

[69] Adam P. Bateman, Julia Laskin, Alexander Laskin, and Sergey A. Nizkorodov. Applications of high-resolution electrospray ionization mass spectrometry to measurements of average oxygen to carbon ratios in secondary organic aerosols. *Environ. Sci. Technol.*, 46(15):8315–8324, 2012.

[70] Charles S. Davis, Phil Fellin, and R. Otson. A review of sampling methods for polyaromatic hydrocarbons in air. *JAPCA*, 37(12):1397–1408, 1987.

[71] Konstantin Siegmann and Klaus Sattler. Formation mechanism for polycyclic aromatic hydrocarbons in methane flames. *Journal of Chemical Physics*, 112:698–709, 2000.

[72] Erin M. Adkins and J. Houston Miller. Extinction measurements for optical band gap determination of soot in a series of nitrogen-diluted ethylene/air nonpremixed flames. *Phys. Chem. Chem. Phys.*, 17(4):2686–2695, 2015.

[73] K. Olaf Johansson, Tyler Dillstrom, Paolo Elvati, Matthew F. Campbell, Paul E. Schrader, Denisia M. Popolan-Vaida, Nicole K. Richards-Henderson, Kevin R. Wilson, Angela Violi, and Hope A. Michelsen. Radical–radical reactions, pyrene nucleation, and incipient soot formation in combustion. *Proc. Combust. Inst.*, 36(1):799–806, 2017.

[74] Paolo Elvati and Angela Violi. Homo-dimerization of oxygenated polycyclic aromatic hydrocarbons under flame conditions. *Fuel*, 222:307–311, 2018.

[75] Michael Frenklach, David W. Clary, William C. Gardiner, and Stephen E. Stein. Detailed kinetic modeling of soot formation in shock-tube pyrolysis of acetylene. *Symp. Combust.*, 20(1):887–901, 1985.

[76] Michael Frenklach. Reaction mechanism of soot formation in flames. *Phys. Chem. Chem. Phys.*, 4(11):2028–2037, 2002.

[77] Jennifer A. Giaccai and J. Houston Miller. Examination of the electronic structure of oxygen-containing PAH dimers and trimers. *Proc. Combust. Inst.*, 37(1):903–910, 2019.

[78] F. Battin-Leclerc. Detailed chemical kinetic models for the low-temperature combustion of hydrocarbons with application to gasoline and diesel fuel surrogates. *Prog. Energ. Combust.*, 34(4):440–498, 2008.

[79] Russell Whitesides, Dominik Domin, Romelia Salomón-Ferrer, William A. Lester, and Michael Frenklach. Graphene Layer Growth Chemistry: Five- and Six-Member Ring Flip Reaction. *J. Phys. Chem. A*, 112(10):2125–2130, 2008.

[80] Edward K.Y. Yapp, Clive G. Wells, Jethro Akroyd, Sebastian Mosbach, Rong Xu, and Markus Kraft. Modelling PAH curvature in laminar premixed flames using a detailed population balance model. *Combust. Flame*, 176:172–180, 2017.

[81] Klaus-Heinrich Homann. Fullerenes and soot formation— new pathways to large particles in flames. *Angewandte Chemie International Edition*, 37(18):2434–2451, 1998.

[82] Ph. Gerhardt, Silke Löffler, and Klaus H. Homann. Polyhedral carbon ions in hydrocarbon flames. *Chemical Physics Letters*, 137(4):306–310, 1987.

[83] Randy L. Vander Wal, Andrea Strzelec, Todd J. Toops, C. Stuart Daw, and Caroline L. Genzale. Forensics of soot: C5-related nanostructure as a diagnostic of in-cylinder chemistry. *Fuel*, 113:522–526, 2013.

[84] Jacob W. Martin, Kimberly Bowal, Angiras Menon, Radomir I. Slavchov, Jethro Akroyd, Sebastian Mosbach, and Markus Kraft. Polar curved polycyclic aromatic hydrocarbons in soot formation. *Proc. Combust. Inst.*, 37(1):1117–1123, 2019.

[85] Arthur L. Lafleur, Jack B. Howard, Joseph A. Marr, and Tapesh Yadav. Proposed fullerene precursor corannulene identified in flames both in the presence and absence of fullerene production. *The Journal of Physical Chemistry*, 97(51):13539–13543, 1993.

[86] Klaus H. Homann and Heinz.Gg. Wagner. Some new aspects of the mechanism of carbon formation in premixed flames. *Symp. Combust.*, 11(1):371–379, 1967.

[87] Hong-Bo B. Zhang, Dingyu Hou, Chung K. Law, and Xiaoqing You. Role of Carbon-Addition and Hydrogen-Migration Reactions in Soot Surface Growth. *J. Phys. Chem. A*, 120(5):683–689, 2016.

[88] Kim C. Le, Christophe Lefumeux, Per-Erik Bengtsson, and Thomas Pino. Direct observation of aliphatic structures in soot particles produced in low-pressure premixed ethylene flames via online Raman spectroscopy. *Proc. Combust. Inst.*, 37(1):869–876, 2019.

[89] Fikret Inal and Selim M Senkan. Effects of oxygenate additives on polycyclic aromatic hydrocarbons(pahs) and soot formation. *Combust. Sci. Technol.*, 174(9):1–19, 2002.

[90] Oleg P. Korobeinichev, Sergei A. Yakimov, Denis A. Knyazkov, T. A. Bolshova, Andrey G. Shmakov, Jiuzhong Yang, and Fei Qi. A study of low-pressure premixed ethylene flame with and without ethanol using photoionization mass spectrometry and modeling. *Proc. Combust. Inst.*, 33(1):569 – 576, 2011.

[91] Nicolas Leplat, Philippe Dagaut, Casimir Togbé, and Jacques Vandooren. Numerical and experimental study of ethanol combustion and oxidation in laminar premixed flames and in jet-stirred reactor. *Combust. Flame*, 158(4):705–725, 2011.

[92] Maurin Salamanca, Mariano Sirignano, Mario Commodo, Patrizia Minutolo, and Andrea D'Anna. The effect of ethanol on the particle size distributions in ethylene premixed flames. *Exp. Therm. Fluid Sci.*, 43:71–75, 2012.

[93] Mario Commodo, Gabriella Tessitore, Gianluigi De Falco, Patriza Minutolo, and Andrea D'Anna. Photoionization Study of Soot Precursor Nanoparticles in Laminar Premixed Ethylene/Ethanol Flames. *Combust. Sci. Technol.*, 186(4-5):621–633, 2014.

[94] Mariano Sirignano, Anna Ciajolo, Andrea D'Anna, and Carmela Russo. Chemical Features of Particles Generated in an Ethylene/Ethanol Premixed Flame. *Energy & Fuels*, 31(3):2370–2377, 2017.

[95] Elizabeth A. Griffin, Moah Christensen, and Ömer L. Gülder. Effect of ethanol addition on soot formation in laminar methane diffusion flames at pressures above atmospheric. *Combust. Flame*, 193:306 – 312, 2018.

[96] Nadezhda Slavinskaya, Victor Chernov, Ryan Whitside, Jan Hendrik Starcke, Uwe Riedel, Oleg Korobeinichev, and Murray J. Thomson. Kinetic study of the effect of ethanol addition on pah and soot formation in ethylene flames. In *29th International Seminar On Flame Structure*, 2017.

[97] Michael Frenklach, Charles A. Schuetz, and Jonathan Ping. Migration mechanism of aromatic-edge growth. *Proc. Combust. Inst.*, 30(1):1389–1396, 2005.

[98] Galiya R. Galimova, Valeriy N. Azyazov, and Alexander M. Mebel. Reaction mechanism, rate constants, and product yields for the oxidation of Cyclopentadienyl and embedded five-member ring radicals with hydroxyl. *Combust. Flame*, 187:147–164, 2018.

[99] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry. Wiley, 1 edition, 2000.

[100] S Prasanna and RJ Doerksen. Topological polar surface area: a useful descriptor in 2D-QSAR. *Current medicinal chemistry*, 16(1):21–41, 2009.

[101] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[102] Jerome H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

[103] Fusheng Xu, Peter B Sunderland, and Gerard M Faeth. Soot formation in laminar premixed ethylene/air flames at atmospheric pressure. *Combustion and Flame*, 108(4):471–493, 1997.

CHAPTER III

Physical Interactions of Gas-Phase Nanoparticles

3.1 Summary

In this chapter, I discuss computational strategies to characterize the physical interactions of PACs as they transition into larger nanoparticles. Central to this process is physical aggregation or dimerization, where PACs form clusters, held together by Van der Waal's and electrostatic forces. Traditionally, this has been captured by modeling the physical interactions of a small number of symmetrical *stabilomer* PACs, however, this approach is inadequate as it neglects the complex PAC chemical space discussed in the previous chapter. This chapter is based on two of my works[1,2] where I show how computational methods can be applied to simultaneously capture the diversity of a large number of PACs and quantify how they will physically interact. Using enhanced molecular dynamics, I simulate a number of dimerization free energy surfaces (FES) which describe the physical interactions of these PACs. Then, using this as training data, I fit an interpretable machine learning model which simultaneously allows me to predict the free energies of the dimerization process and gain insights into the chemical descriptors which control this process. This chapter is broken up into three sections. In the first section, I apply this approach to the free energy of dimerization which governs the thermodynamic equilibrium of the aggregated and disaggregated PACs. In the second section, I apply this approach to the free

energy barrier which governs the kinetics of this process. In the final section, I show how thermodynamic relationships and transition state theory can convert these free energies into useful model values such as the equilibrium constant and rate constant.

3.2 Introduction

Central to modeling soot formation in combustion environments is understanding how polycyclic aromatic compounds (PACs) transition into larger nanoparticles. These multi-ringed aromatic structures are believed to transform into soot through chemical and physical pathways[3]. Unpaired electrons on PACs have been observed to react with other radical species to form three-dimensional structures[4–8], while PACs can stack into larger clusters held together by electrostatic and dispersion forces[9–11]. This physical interaction process is believed to be an important component of soot formation as it provides the initial nucleation step or can hold PACs in proximity with each other, so other chemical growth mechanisms can occur[6,12,13]. For this reason, a proper understanding of the physical interactions of the inception process is a necessary step towards creating a comprehensive model of combustion nanoparticle growth.

Insight into this process can be gained by considering free energies of the dimerization process. Various free energy differences can be applied in models either as the equilibrium constant between aggregated and disaggregated states, or as the kinetic rate constant of the associated reactions (Fig. 3.1). Transition state theory relates the free energy barrier, $\Delta A^{\ddagger}$, with the reaction rate constant through equation 3.1 where κ is the transmission coefficient and h is Planck's constant. Meanwhile the thermodynamic equilibrium constant can be obtained from equation 3.2 with ΔA being the difference in free energies between states. A visual depiction of these relationships is given in fig. 3.1.

$$k_{reaction} = \kappa \frac{k_B T}{h} e^{-\frac{\Delta A^{\ddagger}}{k_B T}} \tag{3.1}$$

$$K_{eq} = \exp \frac{-\Delta A}{k_B T} \tag{3.2}$$

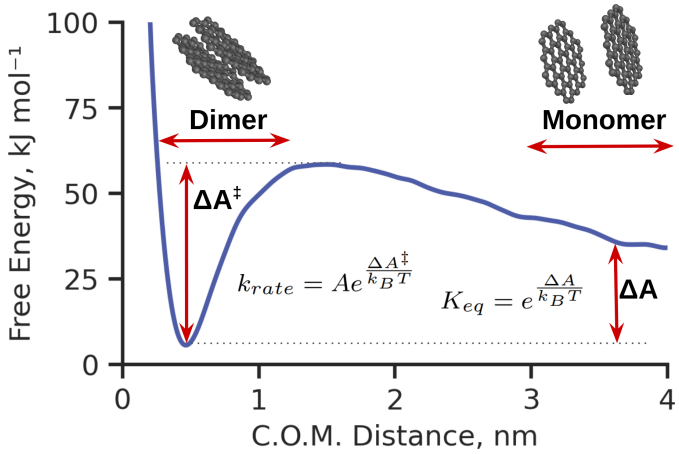

Figure 3.1: Free energy surface of circumcoronene illustrating the free energy difference (ΔA) between stable states and free energy barrier ($\Delta A^{\ddagger}$) for the dimerization process. Exact values depend on the definition of state, as qualitatively shown for monomer and dimer using horizontal arrows, exact extents given in main text. Relationship to equilibrium constant (K_{eq}) and rate constant (k_{rate}) are provided on the plot.

A number of experimental and computational studies have sought to characterize how the process of physical aggregation occurs[14,15]. Miller developed a model that showed the importance of mass in hydrocarbon aggregation and determined that only PACs larger than 800 Da would exist long enough at 1500 K to play a significant role in physical growth[9]. Other studies have concluded that aggregation can occur at lower masses[16]. While many of these earlier studies were based on PACs within the Stein-Fahr stabilomer grid[17], more recent studies of polycyclic aromatic formation have suggested that PACs occupy a much more diverse chemical space with properties such as oxygenation, aliphatic branching, and five-membered rings[18–23]. A number of works have assessed the effects of these properties on the propensity of

these molecules to form physical dimers. Molecular dynamics studies have found that physical dimerization is promoted by aliphatic chains, and thus mass alone is not a sufficient descriptor of the process[24,25]. Moreover, the presence of oxygen[6,26] and molecular curvature[27] have been shown to affect how these molecules dimerize.

In addition to characterizing the properties that promote dimerization, researchers have looked to use these properties to make quantitative relationships about the thermodynamic tendencies of molecules to aggregate[11,28]. As the size of the PAC is understood to be an important property, Herdman and Miller developed a linear relationship between the reduced mass of PAC monomers and their propensity to dimerize[10]. Raj *et al.* showed that the collision efficiency is an important factor in representing dimerization and can be predicted from the mass and shape of constituent PAC molecules[29]. A predictive model for dimer stability has also been developed by fitting the free energy (FE) of aggregation to molecular properties, such as size, number of carbons, and solvent accessible surface area[30].

Although these studies have shown some predictive capacity, they are unable to account for the diverse PAC feature space that has been observed both experimentally[21] and computationally[31] in flames. Recently, Elvati *et al.* examined a number of these properties including size, oxygenation, radius of gyration, and presence of rotatable bonds, finding that all these features affect the aggregation FE landscape[32]. This result suggests that models that do not account for these properties are incomplete and highlights the need for a prediction scheme that can identify the complex relationships these molecular properties have on the physical growth process.

While much of the aforementioned work focuses on the thermodynamic free energy differences of aggregation, kinetic studies of aggregation similarly do not account for the complex PAC landscape. A good deal of soot formation models have used collision theory to determine a physical dimerization rate[6,29,33] where the formation rate constant of the dimer is represented, among a number of other phenomena, with

a single semi-empirical collision efficiency parameter[29]. This collision efficiency can differ by up to five orders of magnitude, and simulations using these efficiencies to implicitly capture reversibility can significantly deviate from expected inception behavior[11]. As such, other models have found success explicitly including the kinetic reversibility. Mao *et al.* found that experimental measurements[34] of pyrene dimerization differed significantly from theoretical values unless the reverse dissociation rate was included[35]. Eaves *et al.* demonstrated that incorporating the reversibility of the physical dimerization process significantly improved predictions of soot volume fraction in flames[28,36] while others have achieved similar results[37]. Recently, a model explicitly considering the dissociation kinetics of five condensed hydrocarbons produced a more accurate characterization of physical dimerization than a model using purely collision efficiencies[38]. These current models of the dimer dissociation kinetics, however, do not properly account for the complexity of PACs in flame systems. State-of-the-art methods[37,39] are still confined to a few stabilomer PACs and compute the kinetic dissociation barriers of these dimers primarily using parameters derived from the PAC mass[4,10,40]. Simulations have suggested that the dimerization is heavily influenced by other properties in the PAC chemical space such as shape and presence of functional groups[26,41] and mass alone provides a poor descriptor of PAC dimerization. Other works have shown significantly more accurate soot formation rates can be derived when including information on these PACs' properties[42]. Thus, accurate modelling of soot inception requires a methodology to quantify the barriers of dimerization accounting for all types of PACs observed in flames.

While molecular dynamics and chemical master equation simulations can provide this information for specific dimer pairs[15,24,34,39,43], it is not feasible to extend these methods to all possible dimers and even less so for larger aggregates. With millions of different PACs observed in even simple flames[44] it is too costly to compute this information for all possible PACs, especially when accounting for all the hetero-dimer

pairs.

To overcome these challenges, in this chapter I combine molecular dynamics, numerical descriptors, and machine learning to quantitatively predict the physical interaction free energies of PACs. Using molecular dynamics simulations, I create a dataset with hundreds of free energy profiles in order to have FE as a function of molecular distance for PACs with different functional groups such as aliphatic chains, five-membered rings, oxygenated groups, and aliphatic linkages. Then labeling each molecular pair with a large number of chemical descriptors, I trained supervised machine learning models (Lasso method[45]) in order to both predict the FE difference (ΔA) related to the equilibrium constant and the FE barrier ($\Delta A^{\ddagger}$) related to the rate constant. The results underscore how machine learning can be used to process a large chemical space with millions of unique PACs with complex properties in order to create more accurate models. As a final discussion, I show how the free energies produced in this work can be extended to practical parameters such as the rate and equilibrium constants which can be applied in nanoparticle growth models.

3.3 Methodology

3.3.1 Datasets

A number of different PACs are considered in this work. In all cases, attempts have been made to sample a wide range of properties including size, shape, five-membered rings, and oxygen groups. For the free energy of dimerization, 14 PACs are considered given in fig. 3.2. For the free energy barrier calculations, a slightly modified set of PACs (A-N) are considered in fig. 3.3. This modification is made since some of the PACs have complex FE surfaces that are not able to properly represent the barrier in a single dimension and therefore would require a different simulation procedure. For the free energy barrier calculations, a separate set of stabilomer PACs[17] is also

considered (A4-A10).

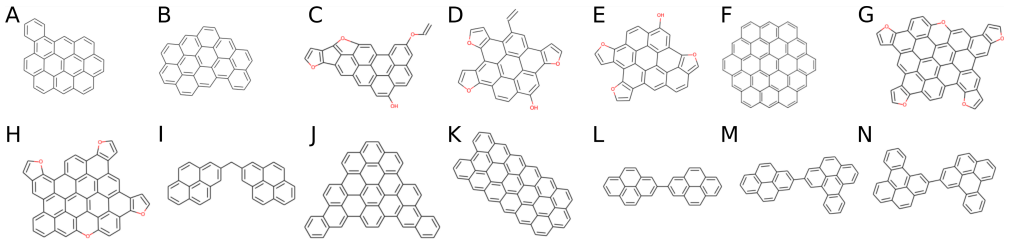

Figure 3.2: PACs used in this work for free energy of dimerization.

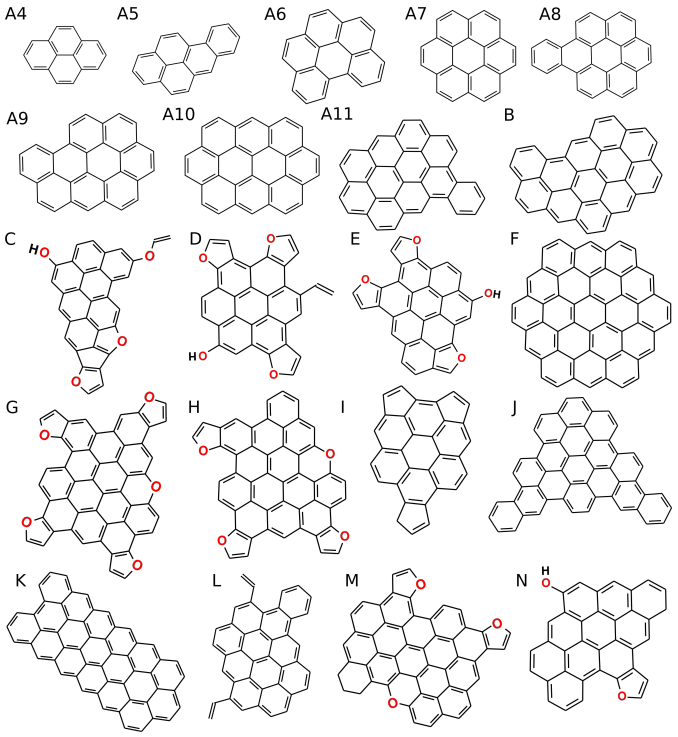

Figure 3.3: PACs used in this work for free energy barriers.

3.3.2 Molecular Dynamics

To generate data for the machine learning model, I considered the FE profiles of all the possible combinations of PACs as discussed in section 3.3.1. This class of molecules contains a diverse set of chemical features observed in both experimental[21] and computational[31] studies, and is an extension of the one used in previous works[6,26,32].

When available, I used previously computed FE profiles[32]; otherwise the missing FE profiles were obtained with the same procedure of that work, briefly reported below. FE profiles were reconstructed by using the well-tempered Metadynamics (WTMD) technique; all simulations were carried out in the NAMD program[46] using the PLUMED plugin[47]. Each simulation consisted of a 1 ns minimization and equilibration starting from two PAC molecules spaced 1 nm apart. This step was followed by a 100 ns simulation biased on the distance between the molecules' center-of-mass (COM). Specific WTMD parameters and temperatures are available in table 3.1.

Table 3.1: Parameters used in Metadynamics simulations.

Parameter	Value
Gaussian Sigma	0.04 nm
Gaussian Height	0.418 kJ mol^{-1}
Gaussian Interval	0.1 ps
Bias Factor at 750 K	20
Bias Factor at 1000 K	15
Bias Factor at 1250 K	12.5

3.3.3 Computation of Free Energy Surfaces

The curves of three independent runs were used to compute two different FE values related to the aggregation process. The free energy of dimerization (ΔA) is computed by considering the difference in free energy between the monomer and dimer state and is related to thermodynamic equilibrium. The free energy barrier ($\Delta A^{\ddagger}$) is the FE difference between the dimer state and the transition state and is related to the kinetic rate constant. While this latter term is related to the dissociation rate constant of the dimer, it can be combined with the free energy of dimerization to get the association rate constant as well. The free energy of each state was defined as the difference between the exponential weighted average for the dimer (approximately 0.35-0.75 nm) and monomer state (3.75-4.0 nm). The transition state was defined as the local maxima between the monomer and dimer state. For the convention of the

free energy of dimerization, positive values indicate that the molecules are more likely to be found not aggregated, while for negative values the aggregate state is preferred.

For computation of the free energy barrier, a geometrical correctional term introduced by Bal *et al.*[48] was applied to ensure independence from the choice of the reaction coordinate. Briefly, the standard free energy surface F^S can be related to the collective variable of COM distance d, by considering the probability distribution of the collective variable, $p(d)$ through equation 3.3:

$$F^S(d) = -\frac{1}{\beta} \ln p(d) \tag{3.3}$$

This formulation is sufficient when computing the thermodynamic free energy difference between the monomer and dimer states, as any dependence on the collective variable is removed during integration. However, the transition state remains highly dependent on the choice of collective variable, which can be corrected by adding a gauge correction ensuring that the kinetic barrier is invariant to the collective variable selection. The resulting geometric free energy surface F^G is given by equation 3.4:

$$F^G(s) = F^S(d) - \frac{1}{\beta} \ln \langle \lambda | \nabla d | \rangle_d \tag{3.4}$$

where λ is a length scale unit conversion and $\langle |\nabla d| \rangle_d$ is the ensemble average of gradient d with respect to center of mass distance coordinates.

Then, in order to obtain the free energy barrier $\Delta A^\ddagger$ in a form consistent with Eyring's equation, equation 3.5 is applied which considers the difference between the transition state $F^G(TS)$, the dimer state $F^G(D)$, collective variable units λ, Planck constant h, and reduced mass m.

$$F^\ddagger(s) = F^G(TS) - F^G(D) + \frac{1}{\beta} ln(\frac{\lambda}{h}\sqrt{\frac{2\pi m}{\beta}}) \tag{3.5}$$

A more detailed derivation of this procedure is given in the original work[48].

3.4 Molecular Descriptors

For each PAC, I computed 312 molecular features describing size, shape, composition, and chemistry with an in-house code. These features were compiled from previous quantitative structure-property relationship studies and have shown success predicting a wide range of organic molecule properties[49–53]. All features included in this study can be easily derived from a molecule's composition, connectivity, and atomic positions without any requirement for atomistic simulations or electron structure calculations. A brief description of features is given as follows, with an exact number of each descriptor given in tab. 3.2:

Basic descriptors: These descriptors are zero dimensional descriptors that can be derived entirely from the molecular formula such as mass and C/H ratio or descriptors based on substructures such as aliphatic chains, oxygen groups, and number of aromatic rings. **CPSA descriptors[50]:** Charged partial surface area descriptors consider the surface area weighted by charge. Examples include total positive surface area and total negative surface area. **Crippen descriptors[54]:** These are the calculated molar refractivity and logP. **SASA descriptors:** Solvent accessible surface area descriptors. Includes total sasa, sasa of all carbon atoms, and average depth from solvent accessible surface. The probe radius is based on a water solvent. **Tessellation descriptors[53]**: A delaunay tessellation is performed with each atom represented as a point. Tessellations find groups of four atoms within proximity of each other. Each tessellation descriptor represents the number of times one atom combination (e.g. 4 carbons, 3 carbons and 1 hydrogen, ect.) is observed. Unlike the original publication, only the unweighted tessellations are computed since property weighted tessellations on this dataset will be linearly correlated. **VSA descriptors[52]:** Van der Waal's surface area descriptors consider 3 properties: atomic charge, atomic logP, and atomic molar refractivity. Each of these properties are binned by their value. Each property-bin combination represents a single descriptor described as the total Van der Waal's

surface area of all atoms whose property values fall within that bin. **WHIM descriptors[55]:** These descriptors consider the 3D coordinates of the molecule and derive a set of descriptors from the the principal axes. WHIM descriptors are weighted in this work by five properties: equal weights, mass, electronegativity, polarizability, and Van der Waal's volume although we note that in this work WHIM descriptors with different property weightings are highly correlated. **3D Descriptors:** While many of the above descriptors include 3D information, this encapsulates the remaining descriptors derived from 3D coordinate data including the radius of gyration, asphericity, and eccentricity.

Table 3.2: Description of feature classes used in the model and number of features within each class.

Feature	Count
Basic descriptors	45
CPSA descriptors	29
Crippen descriptors	2
SASA descriptors	3
Tesselation descriptors	126
VSA descriptors	34
WHIM descriptors	70
3D descriptors	3

Since each FE of aggregation depends on two molecules, I combined individual molecular features by computing the harmonic average, discussed more in the Results section.

3.4.1 Machine Learning

Before training my machine learning model, I eliminated similar features by removing any feature with a variance of zero and any feature with a Pearson correlation greater than 0.95. To build a predictive model for the FE of aggregation, I applied the Lasso method (Scikit-learn implementation[56]), as it allows for high accuracy and often interpretable predictions[45]. Lasso, which uses the least absolute shrinkage and selection operator, has been applied successfully to make interpretable predictions in

chemical problems[57] as it eliminates extraneous features and only selects a smaller subset of properties needed to make the predictions. Specifically, it is a supervised machine learning regression model that minimizes a loss function given by:

$$L(\beta, \lambda) = \sum_{i=1}^{n} (Y_i - \beta X_i) + \lambda \sum_{j=1}^{p} \beta_j \tag{3.6}$$

In this equation, Y is the true value, X are the input features, β is a set of feature weights learned by the model, and λ (different from the λ in equation 3.5) is a regularization parameter set manually. In other words, Lasso minimizes the sum of the squares of the residuals with a regularization term proportional to the l_1 norm, creating a penalty on feature weights, which results in a more concise model. The further regularization of the l_0 norm, which would produce the simplest model by finding the smallest subset of features that fits the data, is the ideal next step, but it is computationally intractable. The l_1 norm employed by Lasso provides an approximation that can be efficiently solved as a convex optimization problem[57].

To avoid artifacts due to the different magnitudes of features affecting my results, training data was first scaled using a standard scaler fit[56] that centers each feature at its mean value and normalizes by the standard deviation. I then used the training data to select the optimal hyperparameters and to train a Lasso model. The Lasso algorithm has only one hyperparameter which needs to be tuned. This parameter is λ, which describes the relative penalization of the l_1 norm. I consider 21 possible values of 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1.0, 2.5, 5, 7.5, 10.0, 25.0, 50.0, 75.0 and 100.0. For each fold of leave-one-out cross-validation, I withhold the testing sample and perform 5-fold cross validation on my training data and select the model using the hyperparameter value with the lowest mean absolute error across the cross validation set. For other machine learning methods used in comparison, I optimized hyperparameters with a similar procedure. Kernel

ridge regression and support vector regression used the same regularization or margin hyperparameter candidate set as Lasso λ and a radial basis function kernel with a gamma value chosen from 0.001 and 1 equally spaced on a log scale. For random forest, either 10, 25, 50, 100, or 250 distinct estimators were used. For the neural network, a network was chosen with either one or two layers using either 25, 50, or 100 neurons in each layer.

I tested the model using leave-one-out cross validation (*i.e.* withholding one FE for each fold and using the remaining as training data) and considered the selected features as those with non-zero coefficients.

3.4.2 Rate Constants and Equilibrium Constants

The free energy values can be converted into model parameters describing the physical interactions of polycyclic aromatic compounds. The equilibrium constant can be derived using the thermodynamic relationship described in equation 3.2. The kinetic rate constant can be derived using the transition state theory from equation 3.1 with a κ value approximated at 0.66. Note that transition state theory can be applied to both the association or dissociation depending on which barrier is used.

One consideration for these calculations, however, is that while the dimer state exists across a clear range of distances (as defined by the local minima of the free energy well), the precise definition of the monomer state is more complicated and depends on the system temperature and pressure. Due to the entropic contribution to the free energy, as two molecules separate, the most stable monomer state will maximize the inter-particle distance. The thermodynamic basis for this is given in the monotonically decreasing equation 3.7. It can be seen in fig. 3.1, that the free energy surface decreases with increasing C.O.M. distances according to this principle. While this correction shows that the free energy of the monomer state will continually decrease with increasing C.O.M. distance, there is a practical upper limit set by

the partial pressure of each PAC. Given a partial pressure and temperature, above a certain C.O.M. distance a PAC monomer will be closer to other monomers in the system. Therefore, while free energy dictates that higher C.O.M. distances are preferred, nearby molecules place a practical constraint on the maximum distance. For these gas systems, however, there is also a statistical distribution of distances as not all particles will necessarily be in the most stable state but rather occupy an ensemble of distances.

$$\Delta A(r) = -2k_B T ln(r) \tag{3.7}$$

Therefore to calculate the free energy of the monomer state, I numerically simulated a distribution of minimum distances r between 2500 particles randomly sampled in a box. The box was sized with the ideal gas law for a specific temperature and pressure and distances were calculated with periodic boundary conditions. I then used the free energy surface, the histogram weight of each distance $w(r)$, and the entropic correction from equation 3.7 to get a free energy of the monomer state $A(r)$ as a function of distance. For this, atomistic simulations or machine learning was used to obtain the monomer state at 4 nm and approximated with a proportional relationship with temperature. Then, equation 3.7 was used to adjust the free energy surface to the requisite COM distance in order to obtain the free energy of the monomer state. I then obtained the free energy of the monomer state S with equation 3.8. Unless otherwise mentioned, the monomer and dimer state were considered at a standard pressure of 1 bar.

$$A_{monomer} = \int_S w(r)A(r)dr \tag{3.8}$$

3.5 Free Energy of Dimerization Results

In this section, I consider the free energy of dimerization. The subsequent section is based on one of my publications[1]. This free energy describes the difference between the monomer and dimer state and is related to the thermodynamic equilibrium of the physical interaction process. I show how simulation and machine learning can be combined to calculate these free energy differences and discuss various aspects of the model.

3.5.1 Machine Learning Predictions

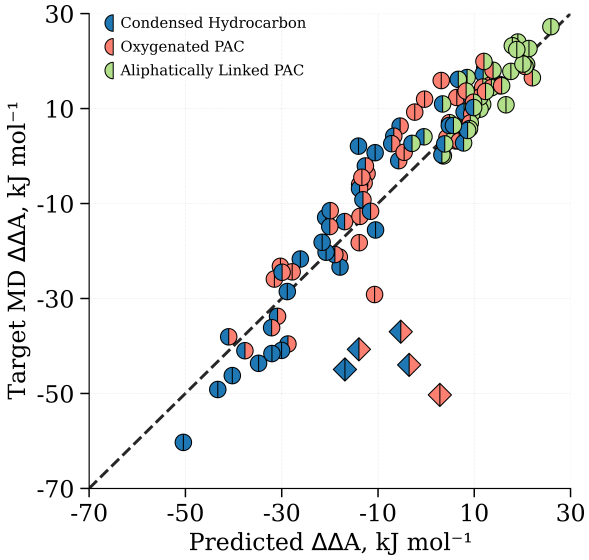

Figure 3.4: Comparison between calculated and predicted FE of aggregation at 1000 K. The dashed line provides reference of correct predictions. Color represents dimer component type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics. Diamonds represent dimer pairs with errors at least twice the RMSE (10.2 kJ mol^{-1}).

After running 105 simulations at 1000 K using the structures in fig. 3.2, I consider the accuracy to which my machine learning procedure can predict the free energy of dimerization. My predictive model (Fig. 3.4) performs well, with a mean absolute error (MAE) of 6.4 kJ mol^{-1}, only slightly higher than the average uncertainty of my

MD simulations ($3.5\,\mathrm{kJ\,mol^{-1}}$) and the average energy for one translational degree of freedom ($4.184\,\mathrm{kJ\,mol^{-1}}$) at 1000 K (see "This work" in Fig. 3.5). My training MAE is only slightly lower at $5.8\,\mathrm{kJ\,mol^{-1}}$ which suggests minimal over-fitting of the model. To test for information leakage, or in other words that the model is learning from general molecular properties and motifs and not the presence of the same monomer in the training set, I repeated the leave-one-out cross validation while omitting from the training data all samples that share a monomer with the testing sample. As expected, since I am training with fewer data ($\sim$13% for homo-aggregation and $\sim$26% for hetero-aggregation, smaller dataset), the prediction slightly worsens ("This work, restrict" in Fig. 3.5). However, with a root mean squared error (RMSE) of $11.3\,\mathrm{kJ\,mol^{-1}}$ and MAE of $7.6\,\mathrm{kJ\,mol^{-1}}$, the model still performs better than existing models (see Fig. 3.5).

Interestingly, the RMSE of my predictions is higher than the MAE, which suggests that a few interactions are not predicted as well as the rest of the data. The analysis of the data highlights that there are 5 pairs (*i.e.* AD, AJ, BD, CF, and DE) for which the error in the predicted FE value is twice the RMSE value. For all of them, the predicted aggregate is less stable than the MD simulations would indicate, and four of them involve molecule C or D, which are the only ones in the dataset that have both aliphatic chains and oxygenated groups. It has been observed that oxygenation destabilizes the physical aggregations of PACs[6,26] while aliphatic chains show the opposite trend[24,25], and when multiple competing features of similar magnitude affect the FE, the outcome is not easy to predict[32]. Moreover, since my dataset lacks molecules that have only aliphatic chains or more cases of similar molecules, it is possible that the model is not able to properly learn the interplay of these particular features.

My model outperforms existing physical dimerization models present in the literature, as shown in Fig. 3.5. I compared my results, including the test on a restricted

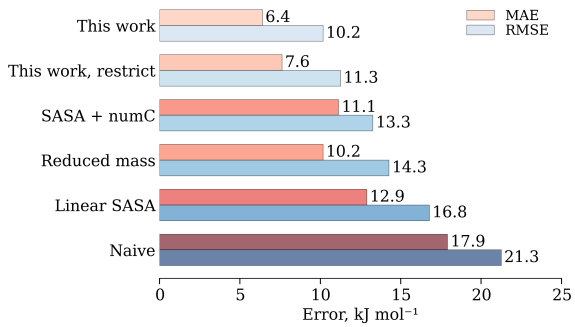

Figure 3.5: Comparison between proposed predictive model and published ones applied to same dataset at 1000 K. Red shows the MAE and blue the RMSE. For reference, the average standard error of the input data (MD simulations) is $\sim$3.5 kJ mol^{-1}.

dataset with no data leakage illustrated above (labeled "This work" and "This work, restrict", respectively) with three additional models. For all methods, I performed a leave-one-out validation procedure, fitting each model's parameters and functions to the molecules and FE in my dataset. First, I compared my results with the widely used model introduced by Herdman and Miller[10], which assumes a linear correlation between the reduced mass and the binding energy (labeled "Reduced mass"). More recently, Lowe *et al.*[30] characterized a number of polycyclic aromatic dimers and developed a predictive model for the change in FE between the monomer and dimer states based around the solvent accessible surface area and number of carbons. For my second comparison, I used the linear fit from the original publication that relates the average carbon surface area and the FE (labeled linear SASA). Next, instead of using the published linear fit, I instead used the molecular descriptors (number of carbons and solvent accessible surface area) as input features into a Lasso model (labeled as SASA + numC). In addition to these, I consider the naive case in which all values were predicted as the average free energy of the dataset. In all cases, the predictive model presented in this work performs better than the previous models, showing that a more comprehensive feature set, such as the one employed here, can better capture molecular properties responsible for dimerization.

Finally, I compared the performance of Lasso against other five machine learning methods, namely, linear regression, support vector machine, kernel ridge regression, random forest, and multilayer perceptron methods. The results (table 3.3), show that all methods have a similar predictive error, except linear regression, which has markedly lower performances. These results confirm the validity of my choice, as no method outperforms Lasso in both MAE and RMSE, and, unlike the other methods, Lasso creates a sparse feature space adding some interpretability by identifying the features that are used in the predictions.

Table 3.3: A comparison of prediction errors for different machine learning methods for free energy of dimerization. All units are in $\mathrm{kJ\,mol^{-1}}$.

Model	MAE	RMSE
Lasso	6.4	10.2
Linear Regressor	10.2	13.9
Support Vector Machine	5.7	10.5
Kernel Ridge Regression	6.8	10.5
Random Forest	6.5	10.1
MLP Neural Network	7.5	11.1

3.5.2 Molecular Feature Selection

As discussed above, one advantages of Lasso is its ability to provide a degree of interpretability towards the aspects that control the prediction, as it sets coefficients of unused features to zero[45]. Thus, by analyzing which features the Lasso model retains, I can gain a sense of which molecular properties are important for predicting the FE of dimer aggregation.

Overall, across all 105 folds of cross validations, the model selects a nearly identical set of 10 features. If I exclude these top features, no other feature is selected in more than four folds and as such I will not discuss them. Broadly, the top features can be divided into three groups of properties that are important for PAC dimerization: size, shape, and presence of specific chemical groups.

3.5.2.1 Size

The first class of properties are extrinsic properties that are broadly related to the size of the molecule. Specifically, the algorithm selected the number of aromatic rings, the number of carbons not connected to a hydrogen, the number of tessellations containing four carbons, the number of tessellations with three carbons and a hydrogen, and the number of six-membered rings.

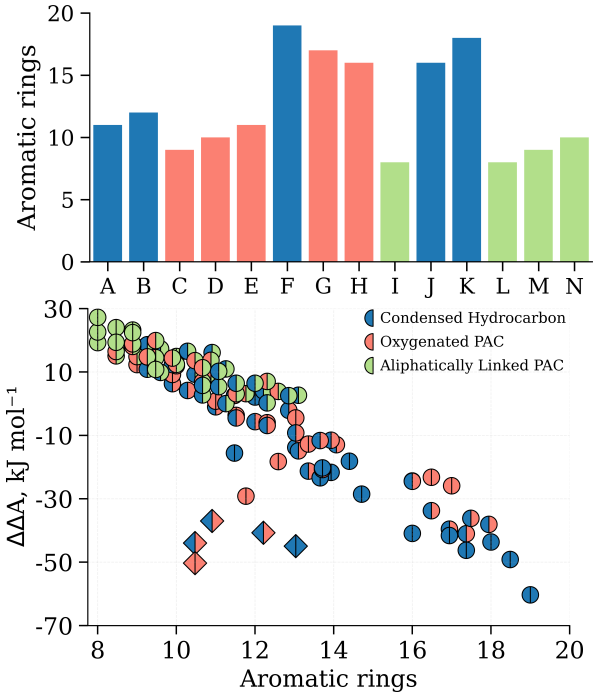

Figure 3.6: Relationship between the number of aromatic rings and dimerization FE. Top: The number of aromatic rings associated with each dimer. Bottom: Aggregation propensity compared to average number of aromatic rings in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Colors represent the dimer's component type: green indicates an aliphatically linked PAC, red an oxygenated PAC, and blue a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

Figure 3.6 shows that the FE of dimerization is strongly related to the (harmonic) average number of aromatic rings in the dimer (Pearson coefficient of -0.8397 and Spearman coefficient of 0.8719). This result agrees with the general observation that PACs will often cluster in lateral stacks, and the interaction strength between PACs

is closely related to their number of aromatic rings[58,59]. Moreover, at least for soot precursors, the number of aromatic rings is closely correlated with mass, hence the use of the latter as a descriptor for the aggregation strength in other works[10].

Among the molecular descriptors in this class, the number of aromatic rings is the feature that has the highest correlation with the FE (more than the number of six membered rings, for example), but it is crucial to note that, by itself, it is not sufficient to fully capture the physical dimerization. A linear fit of the FE as a function of the total number of aromatic rings produces a predictive model with an RMSE of $15.6\,\mathrm{kJ\,mol^{-1}}$ and a MAE of $11.3\,\mathrm{kJ\,mol^{-1}}$, which has a significantly larger error than the Lasso model and is (not coincidentally) comparable to using only the mass as a descriptor (see Fig. 3.5)

Some features in this group encode size with molecular shape information. One such example is the number of internal carbon atoms, defined as the aromatic carbon atoms that are not bonded to H atoms. As, many of the molecules in the dataset are highly pericondensed hydrocarbons, these PACs will have a greater percentage of internal carbons than catacondensed PACs.

The plot of the number of internal carbons against the dimerization propensity, presented in Fig. 3.7, shows three somewhat distinct groupings: molecules with less than 10 internal carbons, which represent aliphatically linked hydrocarbons, pericondensed molecules with approximately 20 internal carbons, and larger pericondensed molecules with 30 or more carbons. When ignoring the outliers discussed in the previous section, these groupings generally correspond to the stability of the dimer, where aliphatically linked hydrocarbons are less stable than smaller pericondensed molecules and larger pericondensed molecules are the most stable, inline with previous works on the importance of shape of PACs[29] and on the lower dimerization speed and shorter lifetimes of linked PACs[25].

Finally, tessellation descriptors contain similar information of size and shape as

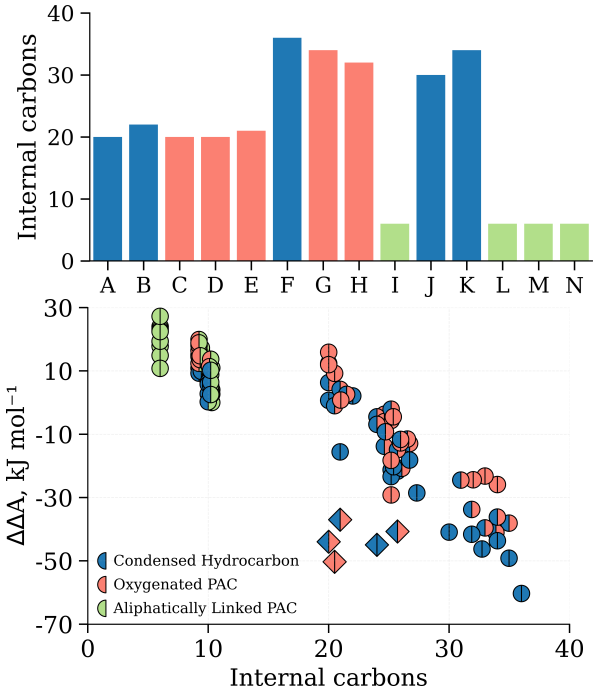

Figure 3.7: Relationship between the number of internal carbons and dimerization FE. Top: The number of internal carbons associated with each monomer. Bottom: Aggregation propensity compared to average number of internal carbons in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

they count the number of times four carbons are in proximity with each other (mostly internal carbons) and the number of times three carbons are in proximity with a hydrogen (mostly edge carbons).

3.5.2.2 Shape

The second group of properties corresponds to quantities that purely describe the shape of the molecules, such as the relative lengths of the first and second principal axis of inertia (WHIM[49] mass axis 1 and 2), which are both size independent.

Figure 3.8 shows the ratio between the second and first principal axes of inertia (*i.e.* aspect ratio), along with its relationship with the propensity of these molecules

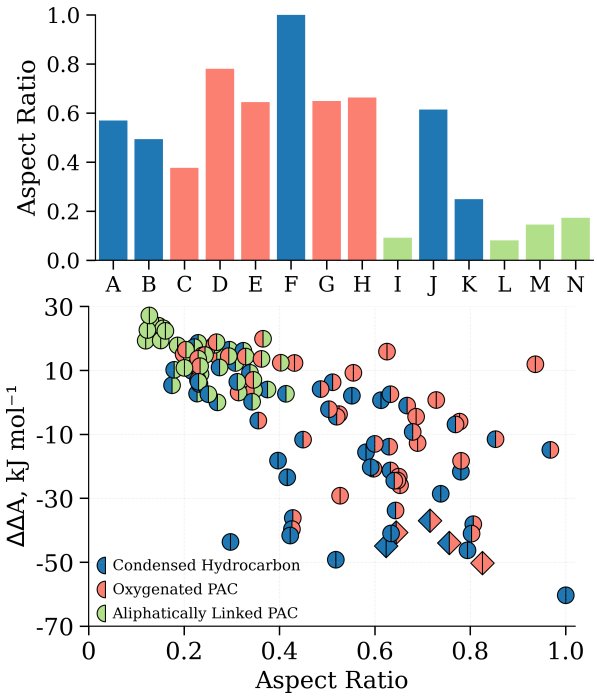

Figure 3.8: Relationship between aspect ratio and dimerization FE. Top: The aspect ratio associated with each monomer. Bottom: Aggregation propensity compared to average aspect ratio in the dimer. The five outliers discussed in the previous section are denoted as diamonds. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. Points with two colors share all the corresponding characteristics.

to dimerize. While a clear separation exists for the less stable dimers containing an aliphatically linked hydrocarbon, it is difficult to identify a trend for the remaining dataset. This suggests that size independent descriptors of shape are likely being used by the model only to identify aliphatically linked PACs and not other compounds. This phenomenon does not imply that shape descriptors do not have a clear relationship with the free energy, as they may play an important role for curved PACs[31]. However, due to the complexity of the FE landscape of curved molecules and the presence of multiple distinct configurations at short distances, I did not include any in this work.

3.5.2.3 Specific chemical groups

The third class of properties groups descriptors that are a metric for the presence of specific chemical groups, like the number of tessellations with three carbons and an oxygen atom, the total Van der Waals surface area of all atoms with a partial charge between -0.05 and 0 (known as the vsa[52] charge 7), and length of the longest aliphatic chain. The tessellation descriptor[53] considers each atom as a point in space and computes a Delaunay triangulation, counting the number of times each combination of each element appears in a tessellation. While the property does not necessarily correspond directly to the number of oxygen atoms (an atom can appear in multiple tessellations), it accounts for the presence of oxygen by counting the number of times an oxygen is in proximity with three other carbons. The vsa charge 7 property encodes information about surface area but also implicitly captures information about oxygenated groups: most carbons that are located near oxygen functional groups are slightly positively charged and thus are not included in the surface area computation. Therefore, for equivalent sizes and geometries, the vsa charge 7 will be lower for molecules with electrophilic groups.

The last property in this group is the length of the longest aliphatic chain, which accounts for the presence of both rotatable bonds and side chains. In combination with the aspect ratio, this feature can distinguish between aliphatically linked chains and side chains, which have the ability to stabilize PAC clusters and make aggregation more favorable[24,25].

3.5.3 Hetero-aggregation

Based on existing models, to predict the aggregation propensity for the hetero-moelcular pairs, I computed the harmonic mean of the two monomers' molecular features. To test if this choice is optimal, I compared the performance of the model

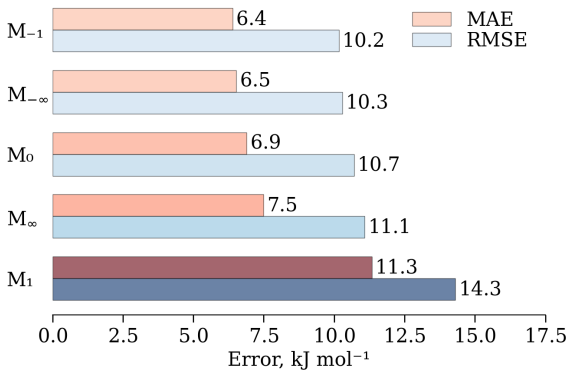

Figure 3.9: Comparison of the predictive performance for different methods of combining monomer features for heterodimerization. $M_{-\infty}$ is the minimum value, M_{-1} is the harmonic mean, M_0 is the geometric mean, M_1 is the arithmetic mean, and M_{∞} is the maximum value. For reference, the RMSE of the input data (MD simulations) is $\sim$3.5 kJ mol^{-1}.

with five different combination rules. Using the definition of generalized mean,

$$M_p(x_1, \ldots, x_n) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^p \right)^{\frac{1}{p}} \tag{3.9}$$

where x_i are the n averaged values, I defined the combination rules as $M_{-\infty}$ (minimum value), M_{-1} (harmonic mean), M_0 (geometric mean), M_1 (arithmetic mean), and M_{∞} (maximum value).

The results, illustrated in Fig. 3.9, show that the harmonic mean outperforms the other metrics, even though the minimum value and geometric mean yield relatively similar results. This trend suggests that between two constituent molecules, the smaller properties tend to have a greater influence on the final stability. Interestingly, however, the error (as a function of p) has a minimum, since $M_{-\infty} \leq M_{-1} \leq M_0$, but the difference is small enough for the current dataset that no further optimization is relevant. While in some cases (*e.g.* charge or shape features) the magnitude of the property does not correspond to the size of the molecule, eight of the top ten features selected (see previous subsections) are extrinsic properties, suggesting that the characteristics of the smaller monomer plays a disproportionately larger role in the

stability and the lifetime of the aggregate. This conclusion provides some empirical foundation to similar observations present in the literature[10,30].

3.5.4 The Effects of Temperature

Up to this point, I considered only data at 1000 K. Here, I test the generality of the selected features at different temperatures. I do this in two ways. First I show that the above method can predict the free energy of dimerization at new temperatures by altering the training set to include only data at a different temperature. Second I show that multiple temperatures can be predicted at once by using a dataset with multiple temperatures and adding temperature as a feature into the machine learning model.

First, I used the previously published homodimerization FE obtained at 500 K and 1680 K[32] to train and test (at each temperature) a Lasso model using only the 10 features selected at 1000 K. While the dataset covers a quite smaller subset of the data used at 1000 K, at very different temperatures the balance of the entropic and enthalpic contributions differs, which can result in the aggregation giving more weight to different molecular characteristics. The prediction results at these two temperatures are shown in Fig. 3.10, with both temperatures, showing an RMSE and MAE lower than the one for the model trained on FE at 1000 K, likely due to the smaller error associated with the prediction of homodimerization.

Overall, the results show that the selected features are valid in a large temperature range. Of note, the error for the model trained with data at 1680 K is significantly greater than the one trained at 500 K, potentially, because physical dimerization is a much less important process at this elevated temperature[15,32] and the system tends towards the ideal gas behavior, for which many of the descriptors become meaningless.

As a second test of the effects of temperature, I predicted the free energy of dimerization for the union of structures in fig. 3.2 and fig. 3.3 at 750 K, 1000 K, and

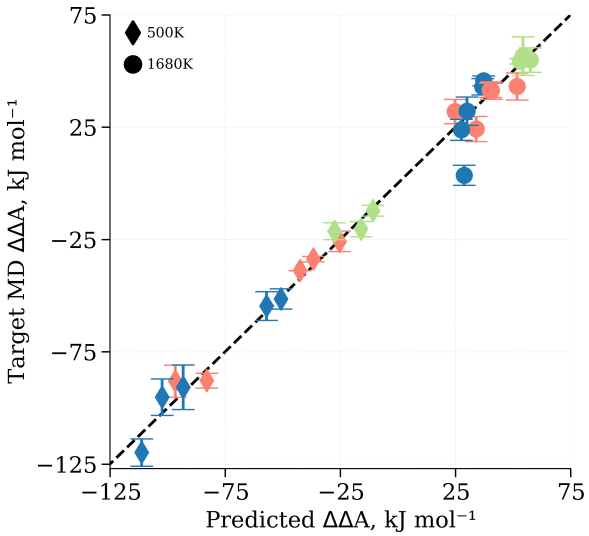

Figure 3.10: Comparison of calculated (MD) and predicted FE of aggregation at 500 K (diamonds) and 1680 K (circles) using only the 10 features selected at 1000 K. Color represents PAC type: green is an aliphatically linked PAC, red is an oxygenated PAC, and blue is a condensed hydrocarbon. The dashed line provides reference of correct predictions. At 500 K, RMSE is 4.9 kJ mol^{-1} and MAE is 4.1 kJ mol^{-1}. At 1680 K RMSE is 8.8 kJ mol^{-1} and MAE is 6.1 kJ mol^{-1}.

1250 K, 165 samples total. In contrast to the above method predicting each dataset at discrete temperatures, this predicted the free energy of dimerizations with all three temperatures in the dataset using temperature as an additional feature. The results of my predictions are shown in Fig. 3.11. Across all temperatures, the Lasso model performs well with a mean average error (MAE) of 5.5 kJ mol^{-1} and root mean squared error (RMSE) of 7.0 kJ mol^{-1}. For comparison, the average experimental uncertainty of the simulations are only slightly lower at 3.0 kJ mol^{-1}. Particularly notably, this method out-performs a linear fit based on the reduced mass and temperature which is the basis of current dimerization energetic predictions[4,10] with a mean average error (MAE) of 8.1 kJ mol^{-1} and root mean squared error (RMSE) of 10.2 kJ mol^{-1}. This is not surprising since a large body of work has shown that properties such as oxygenation, aliphatic linkages, and shape are highly relevant to the dimerization process[30,32,60].

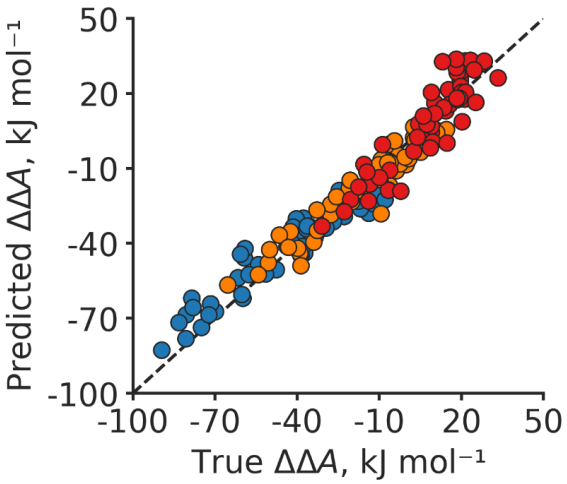

Figure 3.11: Comparison between calculated and Lasso predicted dimerization FE. The dashed line represents the correct predictions. Color represent temperatures: blue is 750 K, orange is 1000 K, and red is 1250 K.

3.6 Free Energy Barriers

In the previous section, I showed how molecular dynamics and machine learning could predict the free energy difference between states. In this section, based on one of my works[2], I consider the free energy barrier. This value is the free energy difference between the dimer state and and transition state and is related to the kinetics of the process. Beginning with the overall simulation results and then extending a machine learning framework, I show that the specific findings and nuances of these barriers are different than the free energy differences. However these barriers can also be predicted accurately with simulation and machine learning.

3.6.1 Simulation Results

From the molecular dynamics simulations, I computed 315 unique free energy dissociation barriers, as shown in Fig. 3.12, with an average uncertainty for the simulations (standard error) at each temperature of 0.9 kJ mol^{-1}, 0.7 kJ mol^{-1}, and 0.5 kJ mol^{-1} at 750 K, 1000 K, and 1250 K respectively. Generally, I observe the same qualitative relationships with size and temperature that I have identified previously

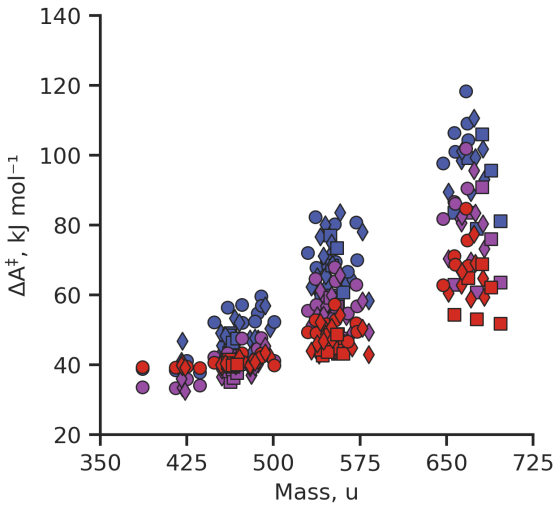

Figure 3.12: Comparison between FE barrier (from molecular dynamics) vs. reduced mass at 750 K (blue), 1000 K (purple), and 1250 K (red). Marker shape represents dimer component type: circles are condensed hydrocarbons, squares are oxygenated PACs, and diamonds are condensed hydrocarbon and oxygenated PAC heterodimers.

for the free energy of dimerization in the previous section. All else being equal, for these molecules and this process, the free energy barriers tend to increase with mass as Van der Waal's and electrostatic interactions become stronger and decrease with temperature as entropic effects increase in importance. It is important to note, however, that these parameters are insufficient to describe all differences observed in the energy barrier and a quantitative trend cannot accurately be derived solely from these two values.

The COM distance of the transition state tends to increase slightly at lower temperatures and is on average 1.4 nm across all simulations. Any distances closer than the transition state could reasonably be referred to as the dimer state for physical aggregation.

To assess the effect of the gauge-corrected free energies[48], I compared the barrier computed in this work with the standard free energy surface computed previously[1,32]. By using the free energy surface from equation 3.4, I obtain a mean difference of $2.9 \pm 0.2\,\mathrm{kJ\,mol^{-1}}$ which is greater than three times the average uncertainty of the

Table 3.4: A comparison of mean absolute errors for different machine learning methods for free energy barrier predictions. All quantities are expressed in $\mathrm{kJ\,mol^{-1}}$ and lowest error(s) at each temperature are shown in bold.

	Temperature (K)		
Model	750	1000	1250
Linear Regressor	3.9	2.6	1.5
Lasso	**3.2**	**2.0**	1.3
Support Vector Machine	3.3	2.2	**1.2**
Kernel Ridge Regression	3.4	2.3	**1.2**
Random Forest	4.0	3.0	1.8
MLP Neural Network	8.6	6.8	5.3
Naive	16.7	13.1	7.7

simulations. Applying the additional gauge correction in equation 3.5 further increases the barriers by a mean value of $31.1 \pm 0.4\,\mathrm{kJ\,mol^{-1}}$. It is worth repeating that this gauge correction is important only for the calculation of the barrier, and therefore this correction does not affect previous works that use the free energy curve to determine the stability of the states.

3.6.2 Machine Learning Predictions

Next, I tested different machine learning models to understand if the dissociation free energy could be extrapolated from the data I computed with molecular dynamics. The results, including a naive predictor, which estimates a new barrier as the average value of the dataset, are reported in tables 3.4 and 3.5. Not only do all the methods significantly outperform this naive predictor, suggesting they all have predictive power, but they have similar performance, except for the neural network, which likely suffers from overfitting. Therefore, due to the relatively high performance and interpretability[1,57] of Lasso, I select this method and use it for my analyses unless otherwise mentioned.

At temperatures of 750 K, 1000 K, and 1250 K, the free energies are predicted with a MAE of $3.2\,\mathrm{kJ\,mol^{-1}}$, $2.0\,\mathrm{kJ\,mol^{-1}}$, and $1.3\,\mathrm{kJ\,mol^{-1}}$ and an RMSE of $4.1\,\mathrm{kJ\,mol^{-1}}$,

Table 3.5: A comparison of root mean squared error for different machine learning methods for free energy barrier predictions. All quantities are expressed in kJ mol^{-1} and lowest error(s) at each temperature are shown in bold.

Model	750 K	1000 K	1250 K
Linear Regressor	4.9	3.5	2.0
Lasso	**4.1**	**2.7**	**1.8**
Support Vector Machine	4.3	3.2	**1.8**
Kernel Ridge Regression	4.5	3.3	**1.8**
Random Forest	5.3	4.0	2.6
MLP Neural Network	13.8	9.8	8.1
Naive	20.3	16.3	10.0

2.7 kJ mol^{-1}, and 1.8 kJ mol^{-1}. The training errors at these temperatures are similar with an MAE of 2.5 kJ mol^{-1}, 1.4 kJ mol^{-1}, and 0.9 kJ mol^{-1}, respectively, suggesting minimal over-fitting. I note that the errors tend to increase at lower temperatures, but that largely corresponds to the greater absolute values in free energy barriers rather than an increased difficulty in prediction of these energies. This idea is supported by the naive predictions that follow the same trend.

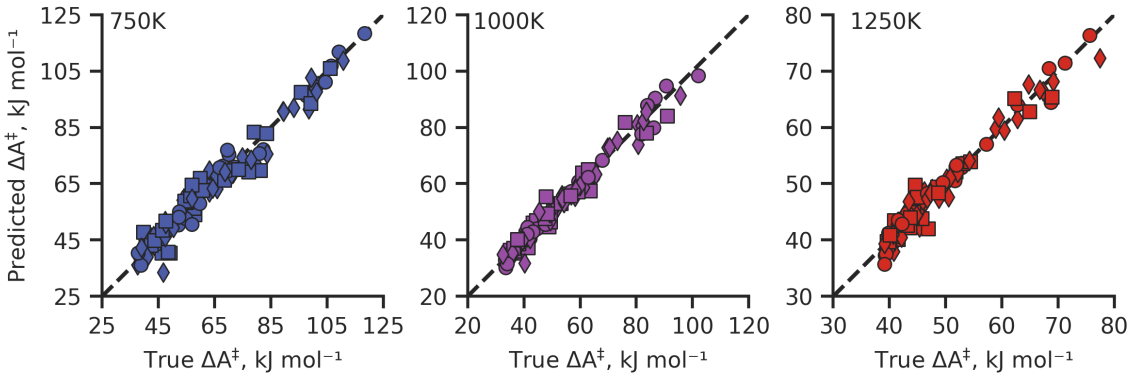

Figure 3.13: Comparison between calculated and predicted FE barrier of aggregation for leave-one-dimer-out at 750 K (blue), 1000 K (purple), 1250 K (red). The dashed line provides reference of correct predictions. Marker shape represents dimer component type: circles are condensed hydrocarbons, squares are oxygenated PACs, and diamonds are condensed hydrocarbon and oxygenated PAC heterodimers.

To demonstrate the improvement of these results over previous methods, I compare

these prediction errors with predictions derived from a linear fit of the reduced mass in Fig. 3.14. The correlation with reduced mass is commonly used[4,10] to derive the dimer dissociation rates in many state-of-the-art methods[36,38]. At all three temperatures, the reduced mass model produces barrier predictions with errors greater than twice my machine learning errors. The significantly higher error of these methods suggest mass is an inadequate descriptor and is not able by itself to capture the effects different PAC features have on the dimerization process. The correlations between mass and dimerization barrier, such as those presented in Fig. 3.12, are not strong enough to capture all variation in free energy with a mean Pearson correlation of 0.40. This results in significant under-fitting, resulting in the higher errors. Machine learning, by contrast, is able to consider a large set of features (including mass) and learn a more complicated relationship between molecular properties and dimer free energy barriers, which results in significantly improved performance.

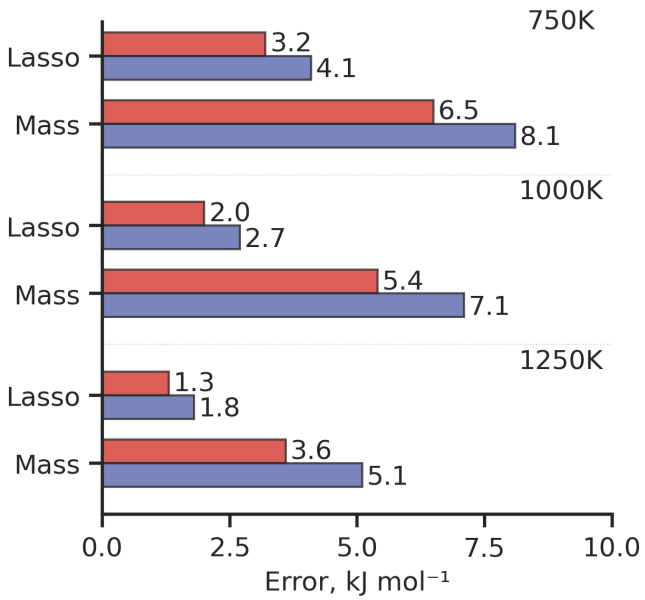

Figure 3.14: Comparison between machine learning (Lasso) and reduced mass (Mass) methods for computing FE barrier of aggregation at 750 K, 1000 K, 1250 K. Red shows the MAE and blue the RMSE. All units are in kJ mol^{-1}.

In addition to producing high accuracy predictions on a diverse set of PACs, I

also demonstrate that my machine learning approach can readily be extended to improve predictions of energy barriers for more simple pericondensed hydrocarbon PAC structures than have traditionally been studied for this problem[10]. To this end, I consider the homo-dimerization and hetero-dimerization for the seven stabilomers between four and ten rings studied by Lowe *et al.*[30] and predict 28 unique barriers in fig. 3.15. At 750 K, I predict these values with an MAE of 0.5 kJ mol^{-1} and RMSE of 0.7 kJ mol^{-1} which is approximately equal to the uncertainty in my simulations. Even in this more simple case, machine learning still outperforms a mass-based fit which has more than twice the error with am MAE of 1.5 kJ mol^{-1} and RMSE of 2.0 kJ mol^{-1}. This suggests that even when predicting the barriers of simple, condensed hydrocarbon PACs, machine learning is still able to capture nuances in the molecular representation and provide more accurate predictions than than a linear mass fit.

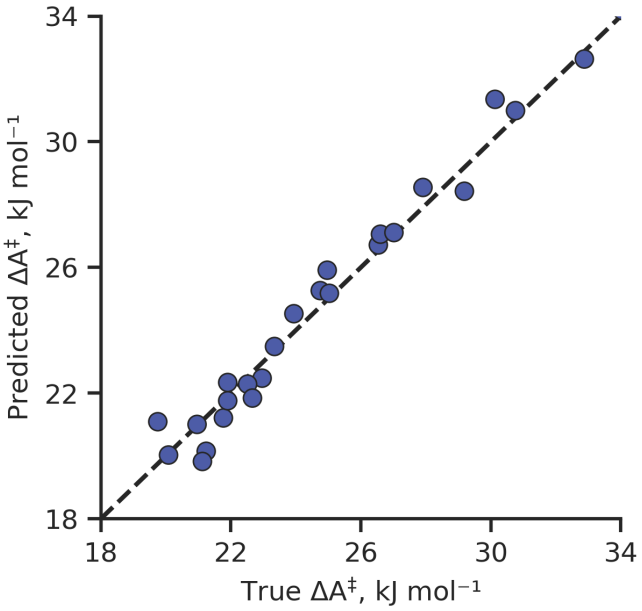

Figure 3.15: Comparison between calculated and predicted FE barrier of aggregation for leave-one-dimer-out at 750 K using a stabilomer-only dataset. Dashed line provides reference of correct predictions.

3.6.3 Influence of Molecular Features

In this section, I consider which descriptors are most important to the prediction of the energy barrier. As discussed above, one of the advantages of Lasso is that it can offer interpretability towards predictions by selecting a smaller subset of features which are used to make the prediction. In the previous section, this approach was able to select 10 PAC features which were necessary for the prediction of thermodynamic energy differences between the monomer and dimer state[1]. The selection of these features is important on two levels. First, selected features can be related to specific chemical properties of the PACs to further enhance our understanding of the dimerization process. Secondly, these descriptors provide a quantitative way to express important properties which can be directly related to the dimerization energetics.

First, I consider an example of how multiple numerical features can capture the dimerization barrier in Fig. 3.16. Here, I consider the PACs' mass and charge weighted fraction of negative surface area, which is a feature selected at all three temperatures. One can see that the mass can describe much of the free energy barrier, however, this single descriptor is too simple to capture all combinations of PACs. At an average mass between 650 u and 700 u, the possible barriers in this narrow mass range differ by as much as $39.4\,\mathrm{kJ\,mol^{-1}}$ resulting in a maximum deviation from the mass fit of $22.7\,\mathrm{kJ\,mol^{-1}}$. With a single descriptor, the model is under-fit and unable to properly capture all the molecular contributions to dimerization in my dataset.

A second descriptor capturing additional properties can significantly improve predictions. Here I consider the ratio of negatively charged to overall solvent accessible surface area weighted by the absolute value of the atomic charge[50]. Due to the weighting by charge, this value is significantly higher for oxygenated PACs whose oxygen atoms exhibit a highly negative character. It can be seen that at higher fractions of negative surface area, the dimer exhibits a lower barrier at equivalent masses. Thus, this descriptor appears to validate the two hypotheses proposed elsewhere[32] (1): that

oxygenated PACs exhibit a slightly destabilizing effect on the PAC dimer and (2): that these oxygen effects are typically secondary to the role of size on dimer stability. It is important to note that this quantitative relationship is specific to this descriptor and is not observed to the same degree when selecting a more heuristic feature to represent oxygenation, such as the number of oxygen atoms. Rather, Lasso automatically selects a feature which implicitly captures the oxygenation while encoding information about the molecular surface area which is also important to PAC dimerization[30] and in turn results in much higher predictive performance.

Next, I consider the totality of features selected by the machine learning model. First I note that a significantly larger number of features are needed to predict the energetic transition state barrier compared to the 10 features needed to predict the thermodynamic dimer stability[1]. There are many commonalities between features selected in predicting the kinetic barrier and thermodynamic stability such as the number of aromatic rings which is highly correlated to the size and mass of the molecule, a number of features capturing aliphatic chains which have been observed to strengthen interactions[25], and features implicitly representing the oxygenation which weaken interactions[26] are also present. I also note and herein discuss a number of features unique to predicting the barrier pertaining to presence of functional groups, surface area, and shape.

When considering features only selected when predicting the barrier, Lasso selects the number of five-membered rings at all three temperatures. Beyond their contribution to curvature[61], which likely falls outside the scope of my dataset, five-membered carbocycles[30] and heterocycles[26] have been observed to slightly destabilize the dimer cluster. It is interesting that this particular feature is selected since the model has the option to explicitly select the number of furans, which is does select, and the number of five-membered carbocycles which it omits. Since the contributions between a furan group and a carbocycle are expected to be different, this suggests that the additive

nature of the Lasso model is able to capture the effect of the five-membered ring with this feature then separately with additional features, such as the number of furans and one shown in Fig. 3.16, control for the destabilizing role of oxygen.

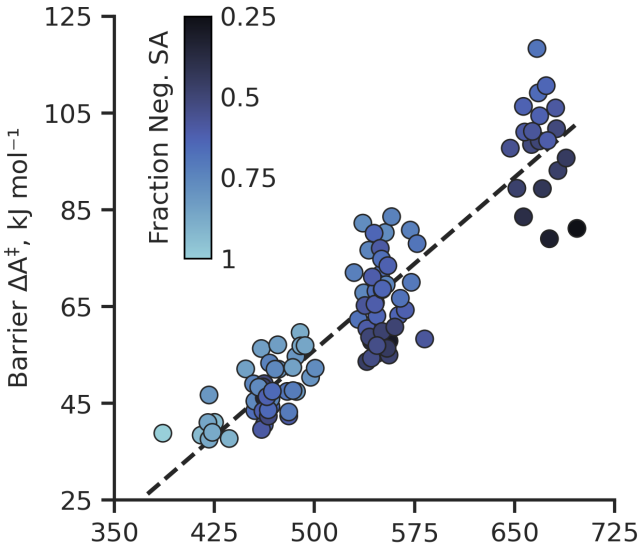

Figure 3.16: Relationship between free energy barrier and mass, charge, and surface area descriptors. Scatter plot shows FE barrier of aggregation vs. reduced mass at 750 K. Dotted line shows the best linear fit between mass and FE barrier. Colors represent the charge weighted fraction of negative surface area.

The surface area is also an important parameter, with descriptors selected including property-weighted measures of solvent accessible and Van der Waal's surface area, such as the one discussed in Fig. 3.16. These descriptors capture information about the interacting surface area which is an essential component of collision theory rate calculations[33] and when normalized by size can numerically represent the degree to which the PAC is catacondensed or pericondensed. Interestingly, unlike other works which focus on number of carbons and mass to describe the size of the PACs, my machine learning framework shows a preference for surface area descriptors. While all these size descriptors are highly correlated, these surface area descriptors encode more information about the actual positions of atoms in the molecule and available interaction area and thus may be more useful when developing quantitative dimerization

relationships.

Another class of descriptors selected are the WHIM descriptors, which encode information about atomic distribution of the molecule along three principal axes. These descriptors encode shape while providing additional information about the size of the molecule. In addition to providing a measure of effective area of the molecule based on its principal axes of inertia, these descriptors can measure the linearity and symmetry of a PAC[62]. This allows it to distinguish between catacondensed PACs (such as dimer KK) that have a lower dissociation barrier than other PACs with equivalent masses and pericondensed symmetric stabilomers (such as dimer FF) which are more stable.

Overall, the model is able to automatically select a number of descriptors encapsulating size, surface area, shape, and presence of specific chemical groups which align with previous studies suggesting that these properties are important[1,4,6,10,29,30,32]. While these properties are helpful for understanding the dimerization chemistry, they are not interchangable with the descriptors selected by Lasso. For example, Lasso does not use atomic counts of carbon or oxygen and selects other descriptors in place of the commonly used reduced mass. Despite the role these properties may play in the dimerization process, Lasso suggests there are more suitable descriptors to implicitly capture these properties in the molecular representation. Thus, the descriptors selected by Lasso are not simply a direct count of properties which have previously been known to influence aggregation, but rather are a set of numerical descriptors which best represent multiple properties important to the dimerization process and can quantitatively be related to the dimer dissociation barrier.

3.6.4 Free Energy Barrier: Predicting Temperature Effects

While up until this point I have considered all barrier predictions independently at three separate temperatures, practical application of this machine learning to soot

inception models requires that predictions can incorporate temperature effects and make predictions across multiple temperatures. Thus, to assess how well my model can be extended to predicting new temperatures, I aggregate data from all three temperatures and perform leave-one-out validation of all dimer-temperature combinations, including the temperature as an additional numerical feature. Using Lasso in Fig. 3.17 to predict these free energies across all three temperatures, I obtain a MAE of $4.7\,\mathrm{kJ\,mol^{-1}}$ and RMSE of $5.9\,\mathrm{kJ\,mol^{-1}}$. Once again, this error is considerably lower than a fit with reduced mass and temperature, which has an MAE and RMSE of $6.4\,\mathrm{kJ\,mol^{-1}}$ and $8.1\,\mathrm{kJ\,mol^{-1}}$ respectively.

From the data, it can be observed that the relationship between temperature and the dimer free energy barrier is often non-linear. A potential drawback of the Lasso method, is that it is only capable of identifying linear relationships between features and a target value. Therefore, while the method performs well at individual temperatures and still outperforms mass-based fits across multiple temperatures, improvement can likely be realized in predicting temperature effects by incorporating a non-linear machine learning method. Therefore, in Fig. 3.17 I also include predictions from support vector regression (SVR) with a radial basis function kernel, which is among the highest performing non-linear method in Tab. 3.4. This method further improves upon the Lasso performance with an MAE and RMSE of $2.4\,\mathrm{kJ\,mol^{-1}}$ and 3.2, demonstrating that predictions specifically across multiple temperatures may benefit from incorporating a non-linear relationship with temperature. It is observed that samples at temperatures of 750 K and 1250 K have higher errors than those at 1000 K. This is to be expected since these intermediate temperature values are interpolated between two known free energies, while the other temperatures require some extrapolation. Based on these results, the range of temperatures used in my simulations should provide a temperature domain across which this prediction methodology should be valid. While simulations at additional temperatures may be added in fu-

ture works, at lower temperatures I would expect minimal PAC growth[63] and higher temperatures other phenomena such as chemical bond formation likely become much more prevalent[5,13].

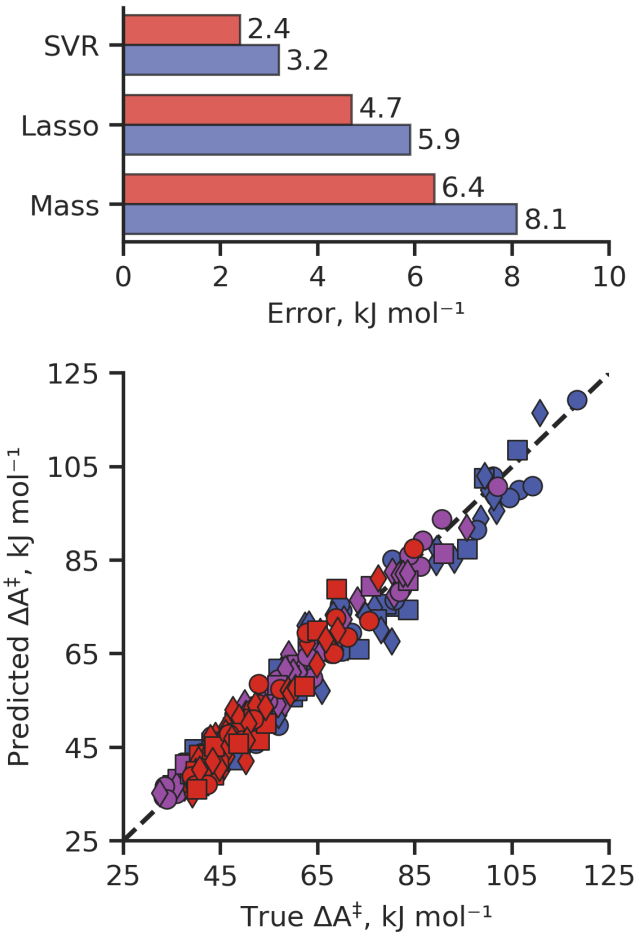

Figure 3.17: Top: Comparison between machine learning models (Lasso and SVR) and reduced mass (Mass) method for computing FE barrier of aggregation for all temperatures. Red shows the MAE and blue the RMSE. All values are in kJ mol^{-1}. Bottom: Comparison between calculated and predicted (SVR) FE barrier of aggregation at all temperatures. The dashed line provides reference of correct predictions. Colors represent temperatures:750 K is blue, 1000 K is purple, 1250 K is red. Marker shape represents dimer component type: circles are condensed hydrocarbons, squares are oxygenated PACs, and diamonds are condensed hydrocarbon and oxygenated PAC heterodimers

3.7 Adapting Free Energies to Model Parameters

Throughout this chapter, I have discussed computational methods for obtaining free energies which describe the physical interactions of PACs. Multi-scale simulations of these physical interactions, however, require free energies to be transformed into practical parameters such as the rate constant and equilibrium constant which can be applied in differential equations and equations of state. Therefore, in this final section I show how useful model parameters can be derived from the free energy. I compare with existing calculations and measurements to show that my rate and equilibrium constants agree with known correlations and improve upon existing calculation methods.

3.7.1 Equilibrium Constant

First, I consider how the equilibrium constant (K_p) varies with temperature. I compute the free energy of dimerization with equation 3.2 at a pressure of 1 bar correcting the free energy surface using the distance calculation discussed in the methodology. While the equilibrium constant is difficult to experimentally measure at such high temperatures due to a variety of competing reactions, I compare against a number of existing theoretical calculations of this value. First, I compare the K_p of pyrene (A4 in fig. 3.3) against correlations from Sabbah *et al.* and Totton *et al.*[11,34] in fig. 3.18. Some deviation exists between all three correlations especially at very low values of the equilibrium constant, however, overall my computed K_p values agree well with existing correlations. It is worth emphasizing here, that due to the difficulty in measuring this phenomena experimentally one cannot determine which of these correlations is closest to the ground truth. Rather, the fact that this approach arrives at similar values to two very different theoretical calculations suggests its utility.

Secondly, I compare circumcoronene (F in fig. 3.3), a much larger PAC, against the correlation from Totton *et al.* in fig. 3.19. Here, excellent agreement is observed

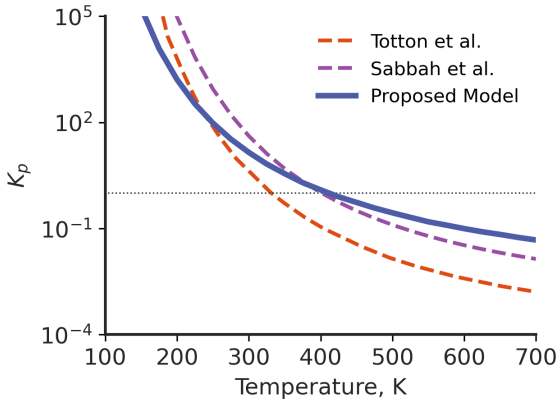

Figure 3.18: Equilibrium constant K_p vs. temperature for pyrene homo-dimerization. The proposed equilibrium constant is given with a solid blue line while two existing correlations[11,34] are provided for comparison with a red and purple dashed line. The dark dotted line is provided at a value of 1. The dimer state is favored above the line while the monomer state is favored below it.

between the two calculations. Interestingly, my calculations are only computed based on simulations between 750 K and 1250 K, however, good agreement is seen up to 2500 K which suggests that some extrapolation is possible at higher temperatures. This is likely due to the fact that at higher temperatures the entropic free energy contributions from equation 3.7, which are linearly related to temperature, are much greater relative to other contributions.

3.7.2 Rate Constant

Next, I show how the barrier can accurately reproduce the kinetic rate constant for pyrene dimerization. To this end, I compare my calculated rate constants for pyrene aggregation with experimentally measured rate constants from Sabbah *et al.*[34] and theoretical calculations from collision theory[29]. Due to experimental limitations, measurements were only taken at low temperatures and pressures. The rate calculations based on MD are simulated with a high number of collisions from the thermostat and as such do not exactly correspond to the low pressures at which the experimental measurements were made. Thus, I consider the rate calculated from equation 3.1

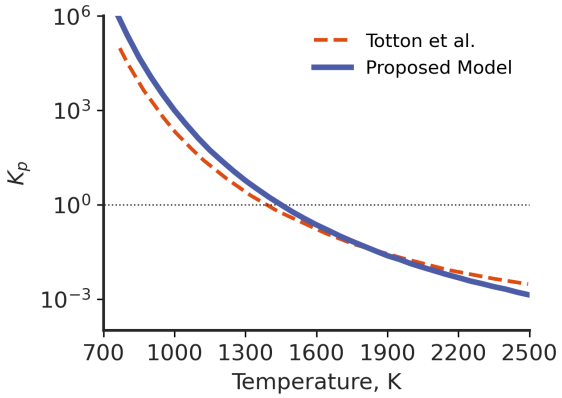

Figure 3.19: Equilibrium constant K_p vs. temperature for circumcornene homo-dimerization. The proposed equilibrium constant is given with a solid blue line while an existing correlation[11] is provided for comparison with a red dashed line. The dark dotted line is provided at a value of 1. The dimer state is favored above the line while the monomer state is favored below it.

and collision theory as the high pressure limit and correct for low pressure effects with Troe parameterization[64] using the low pressure limits from Biennier *et al.* for pyrene[65].

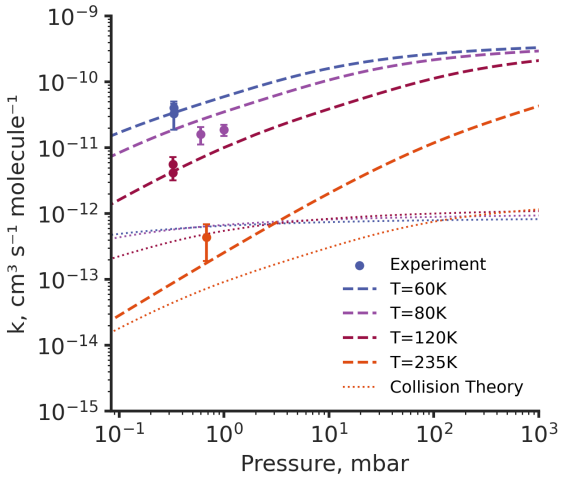

Figure 3.20: Calculated rate constants for pyrene homo-dimerization (dashed line) compared against experiment[34] (circles) and collision theory (dotted line).

The results are shown in figure 3.20. One can see that at all temperatures, the rates calculated with my transition state theory approach closely conform to the

experimental measurements. Notably when comparing to collision theory, my calculations show much closer agreement with experiment at all temperatures. This highlights some of the inadequacies of existing collision theory calculation methods and suggests that the transition state theory approach using the free energies offers a means to improve accuracy of even the most well-studied PAC physical interactions.

3.8 References

[1] Jacob C. Saldinger, Paolo Elvati, and Angela Violi. A machine learning framework to predict the aggregation of polycyclic aromatic compounds. *Proc. Combust. Inst.*, 39(1), 2023.

[2] Jacob C. Saldinger, Paolo Elvati, Karam Abdullah, and Angela Violi. Predicting aggregation rates of polycyclic aromatics through machine learning. *In Progress*, 2023.

[3] Andrea D'Anna. Combustion-formed nanoparticles. *Proceedings of the Combustion Institute*, 32(1):593–613, 2009.

[4] Hai Wang. Formation of nascent soot and other condensed-phase materials in flames. *Proceedings of the Combustion Institute*, 33(1):41–67, 2011.

[5] Mo R. Kholghy, Georgios A. Kelesidis, and Sotiris E. Pratsinis. Reactive polycyclic aromatic hydrocarbon dimerization drives soot nucleation. *Phys. Chem. Chem. Phys.*, 20(16):10926–10938, 2018.

[6] Paolo Elvati, V. Tyler Dillstrom, and Angela Violi. Oxygen driven soot formation. *Proceedings of the Combustion Institute*, 36(1):825–832, 2017.

[7] Andrea D'anna, Angela Violi, Antonio D'alessio, and Adel F Sarofim. A reaction pathway for nanoparticle formation in rich premixed flames. *Combustion and Flame*, 127(1-2):1995–2003, 2001.

[8] Angela Violi. Modeling of soot particle inception in aromatic and aliphatic premixed flames. *Combust. Flame*, 139(4):279–287, 2004.

[9] J. Houston Miller. The kinetics of polynuclear aromatic hydrocarbon agglomeration in flames. *Symp. Combust.*, 23(1):91–98, 1991.

[10] Jennifer D Herdman and J Houston Miller. Intermolecular Potential Calculations for Polynuclear Aromatic Hydrocarbon Clusters. *J. Phys. Chem. A*, 112(28):6249–6256, 2008.

[11] Tim S. Totton, Alston J. Misquitta, and Markus Kraft. A quantitative study of the clustering of polycyclic aromatic hydrocarbons at high temperatures. *Phys. Chem. Chem. Phys.*, 14(12):4081–4094, 2012.

[12] Andrei Kazakov, Hai Wang, and Michael Frenklach. Detailed modeling of soot formation in laminar premixed ethylene flames at a pressure of 10 bar. *Combustion and Flame*, 100(1-2):111–120, 1995.

[13] K. Olaf Johansson, Tyler Dillstrom, Paolo Elvati, Matthew F. Campbell, Paul E. Schrader, Denisia M. Popolan-Vaida, Nicole K. Richards-Henderson, Kevin R. Wilson, Angela Violi, and Hope A. Michelsen. Radical–radical reactions, pyrene nucleation, and incipient soot formation in combustion. *Proc. Combust. Inst.*, 36(1):799–806, 2017.

[14] J. Houston Miller, Kermit C. Smyth, and W. Gary Mallard. Calculations of the dimerization of aromatic hydrocarbons: Implications for soot formation. *Symp. Combust.*, 20(1):1139–1147, 1985.

[15] Qian Mao, Adri C.T. van Duin, and Kai H. Luo. Formation of incipient soot particles from polycyclic aromatic hydrocarbons: A ReaxFF molecular dynamics study. *Carbon*, 121:380–388, 2017.

[16] Charles A. Schuetz and Michael Frenklach. Nucleation of soot: Molecular dynamics simulations of pyrene dimerization. *Proc. Combust. Inst.*, 29(2):2307–2314, 2002.

[17] Stephen E. Stein and Askar Fahr. High-temperature stabilities of hydrocarbons. *J. Phys. Chem.*, 89(17):3714–3725, 1985.

[18] Jason Y. W. Lai, Paolo Elvati, and Angela Violi. Stochastic atomistic simulation of polycyclic aromatic hydrocarbon growth in combustion. *Phys. Chem. Chem. Phys.*, 16(17):7969, 2014.

[19] K. Olof Johansson, Jason Y.W. Lai, S.A. Skeen, D.M. Popolan-Vaida, K.R. Wilson, N. Hansen, A. Violi, and H.A. Michelsen. Soot precursor formation and limitations of the stabilomer grid. *Proc. Combust. Inst.*, 35(2):1819–1826, 2015.

[20] Jeremy Cain, Alexander Laskin, Mohammad Reza Kholghy, Murray J Thomson, and Hai Wang. Molecular characterization of organic content of soot along the centerline of a coflow diffusion flame. *Phys. Chem. Chem. Phys.*, 16(47):25862–25875, 2014.

[21] Mario Commodo, Katharina Kaiser, Gianluigi De Falco, Patrizia Minutolo, Fabian Schulz, Andrea D'Anna, and Leo Gross. On the early stages of soot formation: Molecular structure elucidation by high-resolution atomic force microscopy. *Combust. Flame*, 205:154–164, 2019.

[22] Jacob C. Saldinger, Qi Wang, Paolo Elvati, and Angela Violi. Characterizing the diversity of aromatics in a coflow diffusion Jet A-1 surrogate flame. *Fuel*, 268:117198, 2020.

[23] Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Stochastic and network analysis of polycyclic aromatic growth in a coflow diffusion flame. *Phys. Chem. Chem. Phys.*, 23:4326–4333, 2021.

[24] Paolo Elvati and Angela Violi. Thermodynamics of poly-aromatic hydrocarbon clustering and the effects of substituted aliphatic chains. *Proceedings of the Combustion Institute*, 34(1):1837–1843, 2013.

[25] Seung Hyun Chung and Angela Violi. Peri-condensed aromatics with aliphatic chains as key intermediates for the nucleation of aromatic hydrocarbons. *Proc. Combust. Inst.*, 33(1):693–700, 2011.

[26] Paolo Elvati and Angela Violi. Homo-dimerization of oxygenated polycyclic aromatic hydrocarbons under flame conditions. *Fuel*, 222:307–311, 2018.

[27] Jacob W. Martin, Kimberly Bowal, Angiras Menon, Radomir I. Slavchov, Jethro Akroyd, Sebastian Mosbach, and Markus Kraft. Polar curved polycyclic aromatic hydrocarbons in soot formation. *Proc. Combust. Inst.*, 37(1):1117–1123, 2019.

[28] Nick A. Eaves, Seth B. Dworkin, and Murray J. Thomson. Assessing relative contributions of PAHs to soot mass by reversible heterogeneous nucleation and condensation. *Proc. Combust. Inst.*, 36(1):935–945, 2017.

[29] Abhijeet Raj, Markus Sander, Vinod Janardhanan, and Markus Kraft. A study on the coagulation of polycyclic aromatic hydrocarbon clusters to determine their collision efficiency. *Combust. Flame*, 157(3):523–534, 2010.

[30] Jeffrey S. Lowe, Jason Y.W. Lai, Paolo Elvati, and Angela Violi. Towards a predictive model for polycyclic aromatic hydrocarbon dimerization propensity. *Proc. Combust. Inst.*, 35(2):1827–1832, 2015.

[31] Qi Wang, Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Molecular structures in flames: A comparison between snaps2 and recent afm results. *Proc. Combust. Inst.*, 38(1):1133–1141, 2021.

[32] Paolo Elvati, Kirk Turrentine, and Angela Violi. The role of molecular properties on the dimerization of aromatic compounds. *Proceedings of the Combustion Institute*, 37(1):1099–1105, 2019.

[33] Mohammadreza R. Kholghy, Meghdad Saffaripour, Christopher Yip, and Murray J. Thomson. The evolution of soot morphology in a laminar coflow diffusion flame of a surrogate for Jet A-1. *Combust. Flame*, 160(10):2119–2130, 2013.

[34] Hassan Sabbah, Ludovic Biennier, Stephen J. Klippenstein, Ian R. Sims, and Bertrand R. Rowe. Exploring the Role of PAHs in the Formation of Soot: Pyrene Dimerization. *J. Phys. Chem. Lett.*, 1(19):2962–2967, 2010.

[35] Qian Mao, Yihua Ren, K. H. Luo, and Adri C. T. van Duin. Dynamics and kinetics of reversible homo-molecular dimerization of polycyclic aromatic hydrocarbons. *The Journal of Chemical Physics*, 147:244305, 2017.

[36] Nick A. Eaves, Seth B. Dworkin, and Murray J. Thomson. The importance of reversibility in modeling soot nucleation and condensation processes. *Proceedings of the Combustion Institute*, 35(2):1787–1794, 2015.

[37] Damien Aubagnac-Karkar, Abderrahman El Bakali, and Pascale Desgroux. Soot particles inception and PAH condensation modelling applied in a soot model utilizing a sectional method. *Combustion and Flame*, 189:190–206, 2018.

[38] Arash Khabazipur and Nickolas Eaves. Development of a Fully Reversible PAH Clustering Model. *Proceedings of the Combustion Institute (in press)*, 2022.

[39] Armin Veshkini, Nick A. Eaves, Seth B. Dworkin, and Murray J. Thomson. Application of PAH-condensation reversibility in modeling soot growth in laminar premixed and nonpremixed flames. *Combustion and Flame*, 167:335–352, 2016.

[40] Xiaofeng Tan. Towards a comprehensive electronic database of polycyclic aromatic hydrocarbons and its application in constraining the identities of possible carriers of the diffuse interstellar bands. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 71(5):2005–2011, 2009.

[41] K. Olof Johansson, T. Dillstrom, M. Monti, F. El Gabaly, M.F. Campbell, P.E. Schrader, D.M. Popolan-Vaida, N.K. Richards-Henderson, K.R. Wilson, A. Violi, and H.A. Michelsen. Formation and emission of large furans and oxygenated hydrocarbons from flames. *Proc. Natl. Acad. Sci.*, 113(30):8374–8379, 2016.

[42] Luke Di Liddo, Jacob C. Saldinger, Mehdi Jadidi, Paolo Elvati, Angela Violi, and Seth B. Dworkin. Exploring soot inception rate with stochastic modelling and machine learning. *Combustion and Flame (in press)*, page 112375, 2022.

[43] Debdutta Chakraborty, Hans Lischka, and William L. Hase. Dynamics of pyrene-dimer association and ensuing pyrene-dimer dissociation. *The Journal of Physical Chemistry A*, 124(43):8907–8917, 2020.

[44] Qi Wang, Jacob C. Saldinger, Paolo Elvati, and Angela Violi. Insights on the effect of ethanol on the formation of aromatics. *Fuel*, 264:116774, 2019.

[45] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.*, 58(1):267–288, 1996.

[46] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[47] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[48] Kristof. M. Bal, Satoru Fukuhara, Shibuta Yasushi, and Erik C. Neyts. Free energy barriers from biased molecular dynamics simulations. *J. Chem. Phys.*, 153:114118, 2020.

[49] Roberto Todeschini and Paola Gramatica. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR and QSAR in Environmental Research*, 7(1-4):89–115, 1997.

[50] David T. Stanton and Peter C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, 1990.

[51] Scott A. Wildman and Gordon M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

[52] Paul Labute. A widely applicable set of descriptors. *J. Mol. Graph*, 18(4-5):464–477, 2000.

[53] Xiliang Yan, Alexander Sedykh, Wenyi Wang, Xiaoli Zhao, Bing Yan, and Hao Zhu. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale*, 11(17):8352–8362, 2019.

[54] Scott A. Wildman and Gordon M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

[55] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Methods and Principles in Medicinal Chemistry. Wiley, 1 edition, 2000.

[56] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[57] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.*, 114(10):105503, 2015.

[58] Mathias Rapacioli, Florent Calvo, Fernand Spiegelman, and Christine Joblin. Stacked Clusters of Polycyclic Aromatic Hydrocarbon Molecules. *J. Phys. Chem. A*, 109(11):2487–2497, 2005.

[59] Paolo Elvati, Elizabeth Baumeister, and Angela Violi. Graphene quantum dots: effect of size, composition and curvature on their assembly. *RSC Advances*, 29, 2017.

[60] Seung Hyun Chung and Angela Violi. Nucleation of fullerenes as a model for examining the formation of soot. *The Journal of Chemical Physics*, 132(17):174502–174502–9, 2010.

[61] Edward K.Y. Yapp, Clive G. Wells, Jethro Akroyd, Sebastian Mosbach, Rong Xu, and Markus Kraft. Modelling PAH curvature in laminar premixed flames using a detailed population balance model. *Combust. Flame*, 176:172–180, 2017.

[62] Paola Gramatica. Whim descriptors of shape. *QSAR & Combinatorial Science*, 25(4):327–332, 2006.

[63] Qi Wang, Paolo Elvati, Doohyun Kim, K. Olaf Johansson, Paul E. Schrader, Hope A. Michelsen, and Anegla Violi. Spatial dependence of the growth of polycyclic aromatic compounds in an ethylene counterflow flame. *Carbon*, 149:328–335, 2019.

[64] Jurgen Troe. Predictive possibilities of unimolecular rate theory. *The Journal of Physical Chemistry*, 83(1):114–126, 1979.

[65] Biennier, Ludovic, Sabbah, Hassan, Chandrasekaran, Vijayanand, Klippenstein, Stephen J., Sims, Ian R., and Rowe, Bertrand R. Insights into the role of polycyclic aromatic hydrocarbon condensation in haze formation in jupiter's atmosphere. *Astronomy and Astrophysics*, 532:A40, 2011.

CHAPTER IV

Interactions of Biological Nanoparticles

4.1 Summary

In this section, I focus on the task of general nanoscale interaction prediction. In previous chapters, I have looked at characterizing the nanoscale interactions of PACs; relying heavily on atomistic simulations to learn the physics of the interactions. However, for many types of nanoparticle interactions this is not practical because reliable experimental or simulation data is not able to be obtained without significant resources. Given that many nanoparticle interactions share common chemistries, this section focuses on how data can be leveraged from multiple sources towards general nanoscale interaction predictions. Biological nanoscale interactions provide a good application for this general prediction tool as they are characterized by data-rich regions such as protein-protein interactions and data-sparse regions such as protein-nanoparticle and nanoparticle-nanoparticle interactions. I introduce a method called NeCLAS[1], a machine learning approach which operates agnostic to any specific kind of molecular structure or motif. After being trained only on protein-protein interactions, NeCLAS predicts a wide range of biological nanoscale interactions providing a means to better understand the interactions and functions of different biological nanostructures.

4.2 Introduction

Many technological, biological, and natural phenomena are governed by interactions that occur at molecular and nanoscopic scales[2–4]. For example, protein-protein interactions (PPIs) are crucial for cellular functions and biological processes in all organisms, from mediating selectivity along signaling pathways and elucidating infection mechanisms, to influencing the development of treatments and therapies[5]. Similarly, protein-nanoparticle interactions (PNIs) dictate the bio-reactivity of nanoparticles and their applications in nanodiagnostics, nanotherapy, and nanomedicine[6]. However, tailoring these interactions requires comprehensive knowledge of the interplay between nanomaterials and biological systems. In recent years, data-driven machine learning (ML) methods have provided insight into the mechanisms of nanoscale interactions, overcoming the cost and complexity of experiments[7] and simulations[8], without requiring *a priori* knowledge of physics- and template-based methods[9].

Partner-independent ML methods predict interaction sites for a target structure, regardless of the complementary nanostructure, and can successfully predict protein-ligand[10] and protein-protein[11,12] interactions. These methods identify features that correlate with the tendency of the target protein to interact with arbitrary nanostructures, but do not consider the properties of the second molecule (partner) directly. This approach is data-efficient, but pairwise information is often highly relevant and results in improved predictions[13,14]. To address this limitation, partner-specific methods predict whether a subunit (*e.g.* protein residue) of one structure interacts with a specific subunit of another complex[13,15,16]. Crucially, by using curated datasets[13,17] that include diverse structures and account for homology, partner-specific methods have successfully predicted the local pairwise residue interactions that control global protein-protein aggregation.

Despite this progress, most approaches are specifically designed for proteins and are not immediately generalizable to a wider range of nanoparticles. As these methods

use properties of the individual amino acids or rely on protein-specific characteristics, they cannot be straightforwardly extended to molecules that lack these motifs, even when they share other physical and chemical features[2,18]. Similarly, current ML methods for predicting PNIs use application-specific properties and small training datasets[19,20], which limits the cross-domain validity of the resulting ML models.

To relax this specificity, here I introduce NeCLAS, **Ne**ural **C**oarse-graining with **L**ocation **A**gnostic **S**ets, a flexible and generalized machine learning approach for predicting partner-specific nanoscale interactions. NeCLAS has two main features. The first is a generalized, atomistically-derived coarse-graining method to generate a rototranslational equivariant representation of nanoparticles and macromolecules. The second is a permutation invariant deep neural network that predicts pairwise interactions between the coarse-grained sites of two different molecules. This chapter showcases NeCLAS with three increasingly complex prediction challenges: (1) binding site for PNIs; (2) dynamic characteristics of PNIs, and (3) nanoparticle-nanoparticle interactions and their tendency to self-assemble. NeCLAS outperforms state-of-the-art PNI prediction methods when predicting interactions between proteins and organic nanoparticles. Furthermore, NeCLAS's PPI prediction is competitive with the best protein-specific methods, and shows potential in predicting nanoparticle-nanoparticle interactions. Overall, NeCLAS demonstrates interaction predictions across multiple domains with a reduced computational footprint. This conceptual framework finds applications in various fields, from biologists who search for interactions between proteins, to materials scientists who can design and engineer nanoparticles for targeted applications, to a broad range of additional nanotechnologies.

4.3 Methodology

4.3.1 Overview: NeCLAS

A common way to develop an ML model is to first create a learnable representation of real-world data, and then use this representation to train the model (Fig. 4.1a). In NeCLAS, the first step is accomplished by converting atomistic information to lower-dimensional coarse-grained (CG) structures before computing properties for each CG site, accounting for both local characteristics and chemical neighborhood. The second step, involves training a permutation invariant deep neural network which outputs a pairwise interaction prediction given a pair of CG sites. NeCLAS uses this network to predict interactions for all combinations of sites between two nanostructures.

The CG representation can easily be tailored to capture structural symmetries, especially when interpretability is a primary concern. For example, *p*-SCLX$_6$ (Fig. 4.1b) is a para-sulfonate calixarene, composed of six repeating units with a positively-charged outer region and negatively-charged inner region[21]. By using twelve CG sites, the procedure consistently allocates two sets of sites that match the symmetry of the molecule, which has a hydrophobic core and anionic rim that facilitates protein recognition via entrapment of arginine or lysine side chains. These CG sites capture the underlying molecular properties, as shown by the two distinct clusters of the CG sites in the principal component analysis (Fig. 4.1c) of their local chemical features. While useful for interpretability, the specific choice of number of sites has a minimal impact on the accuracy of NeCLAS, as long as the extreme choices (*e.g.* one site per nanoparticle) are avoided.

To evaluate my model, I specifically tailored the data and workflow to avoid common causes of artificially inflated estimates of the model performance. First, to ensure model generalization, NeCLAS utilizes a neural network that is inherently invariant to the ordering of input sites. This structure provides a more stable prediction than per-

mutation variant methods do (appendix B). As per standard ML practice, NeCLAS kept a strict separation between train, validation (used to halt network training), and test sets (used to evaluate performance). I chose datasets to avoid redundancy, and assessed NeCLAS's performance with increasingly stricter criteria to test the possibility of information leakage. Lastly, since proteins and nanoparticles may change conformations as they interact, I trained NeCLAS only on unbound structures since the goal is to predict interactions for species with unknown bound conformations[22]. For the datasets, these structural changes were quantified as the root mean squared deviations (RMSD) of the atomic positions of bound and unbound species. Figure 4.1d-f show that conformational changes can be significant during binding.

Following these principles, for PPIs, I chose the Docking Benchmark Dataset (DBD) version 5, a curated set of 230 experimental structures of non-redundant protein complexes (for a total of approximately 15 million residue-residue interactions), in both bound and unbound form[17]. However, for PNIs no such dataset exists, so I used a subset of the data provided by Costanzò *et al.*[23] containing organic nanoparticles. From this data, I generated bound and unbound structures. Since this dataset is small and structural redundancy cannot be avoided, I used it only for testing, preventing information leakage from similar substructures.

4.3.2 Coarse-Grained representation.

To obtain the CG sites (Fig. 4.1b), a predetermined number of sites is randomly initialized and iteratively optimized to match the atomic distribution of a given molecular property[24]. While in this work I select atomic mass as the target property (because it properly affords minimal weight to the weak interactions between hydrogen atoms), other properties such as charge, electronegativity, or surface area may be superior representations in other nanosystems such as metal nanoparticles. The number of CG sites can similarly be manually specified for other nanoparticles based on their

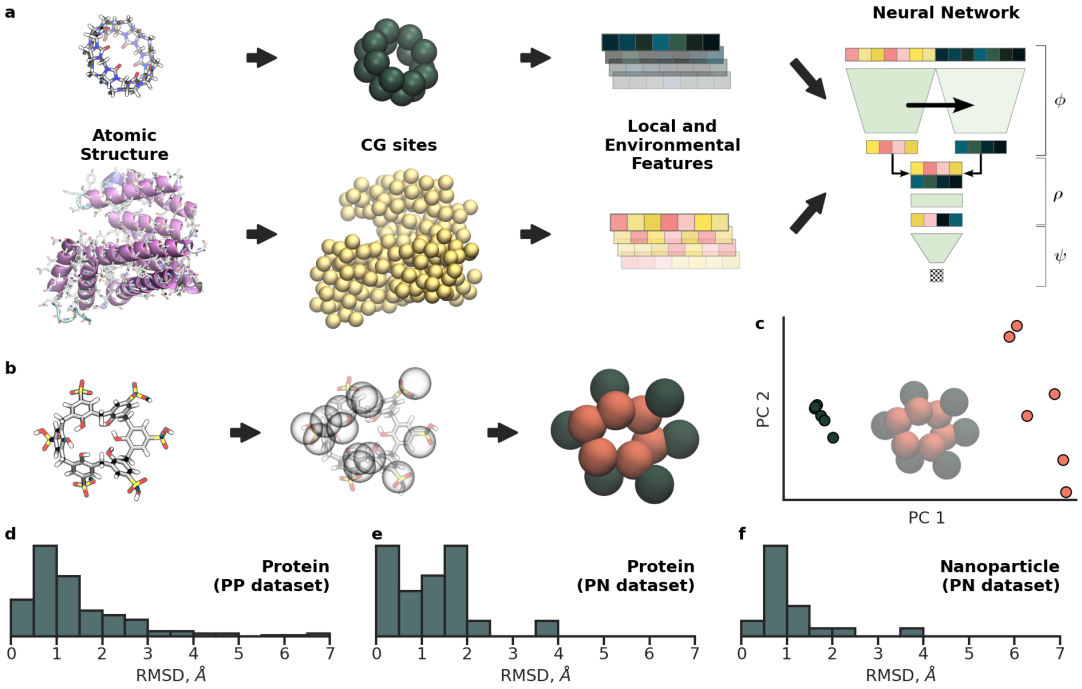

Figure 4.1: **Methods and data**. **a**, NeCLAS schematic. Reduced dimensionality representation (CG sites) and properties are derived from atomic structures (*e.g.* nanoparticles, top row, and proteins, bottom row); a set of the combined local (80) and environmental (400) features is then generated for each pairwise interactions. A neural network is used to predict the interaction between a given pair of CG sites (input as feature vectors). NeCLAS uses this network to predict pairwise interactions between all combinations of CG sites between two nanostructures. **b**, Schematic of the coarse-graining procedure applied on *p*-sulfocalix[6]arene (*p*-SCLX$_6$). CG subunits centers are randomly placed within the coordinate space of a starting molecule, then iteratively shifted to match a target property spatial distribution. Pink and green color indicate the two types of CG-sites (see panel **c**). **c**, First two principal components of the feature set obtained for *p*-SCLX$_6$. The data point colors correspond to the site of the CG nanoparticle shown in panel **b** and at the center of the plot. **d**, Distribution of RMSD between unbound and bound proteins (n=230) in the PPI dataset (DBD version 5). One structure (PDB: 1IRA) was omitted for clarity (RMSD = 8.36). **e-f**, Distribution of RMSD between unbound and bound proteins and nanoparticles (n=21) for the PNI dataset.

structural characteristics using between 5 and 11 heavy atoms per site, while for proteins one site is assigned for every 7.5 heavy atoms which is approximately equal to the average size of amino acids in the dataset.

The coarse-graining method is based on the neural gas algorithm from Martinez *et al.*[24]. Initial positions of N sites centers are assigned across the molecule with the *kmeans++* initialization algorithm[25]. The site's center positions are then iteratively updated to reach a final configuration.

For each iteration, an atom is stochastically selected with a probability according to a target property (mass). Each site is then numerically ranked, k, according to its proximity to the selected atom. Positions of each site are updated through equation 4.1.

$$R_i^{new} = R_i^{old} + \epsilon \exp\left[-k/\lambda\right](v - R_i^{old}) \tag{4.1}$$

where R_i is the site position and v is the coordinate of the selected atom. Parameters ϵ and λ are adjusted according to equation 4.2.

$$p = p_o(p_s/p_o)^{\frac{s}{S}} \tag{4.2}$$

where p is the parameter (either ϵ or λ), p_o and p_s are hyperparameters, s is the current step, and S is the total number of steps. After completion of all iterations, atoms are assigned to the closest site (Euclidean distance). The number of iterations, hyperparameters (table 4.1), and the target property of mass are unchanged from previous implementations[26].

4.3.3 Physicochemical Features

We used 80 features to describe the local chemistry (local descriptors) of each CG site and 400 features to describe the chemical neighborhood (environmental descriptors). The local chemistry was captured by using mass, charge, relative accessible surface area, depth[27], protrusion[28], CPSA descriptors[29], CPSA hydrogen bonding descriptors[30], pocket propensity[31], and mass-weighted WHIM[32]. Depth and protrusion

Table 4.1: Hyperparameters for the coarse-graining procedure. The variable in the first column is assigned the value from the second column when performing coarse-graining. N is assigned the number of CG sites, which is different for each structure. For nanoparticles, the number of sites are described in table 4.4. For proteins, I have one site for every 7.5 heavy atoms.

Variable	Value
ϵ_o	0.3
ϵ_s	0.05
λ_o	0.2 N
λ_s	0.01
N	Preset # of sites
S	200 N

values were computed for each atom in the residue and for depth, protrusion, charge, and mass a feature set was generated by taking the minimum, maximum, mean, sum, and standard deviation of the values. Charges were computed by processing proteins with the PDB2PQR (version 2.1.0) package[33] using the AMBER force field[34], and computed for nanoparticles with the Gasteiger method[35]. A complete list of features are given in table 4.2.

Table 4.2: Descriptors used in model. Included are the name of the descriptor set, the number of descriptors in each set, and details about the implementation of these features. Details include a reference and/or a description of the weighting scheme used to express distributional properties (*e.g.* charge, which is defined for each atom in a bead).

Descriptor	Number	Details
CPSA	29	[29]
CPSA Hydrogen Bonding	16	[30]
WHIM	14	Mass-weighted[32]
Depth	5	Sum, min, max, std deviation, mean[27]
Protrusion	5	Sum, min, max, std deviation, mean[28]
Charge	5	Sum, min, max, std deviation, mean
Mass	3	Sum, std deviation, mean
Pocket Propensity	2	[31]
Relative Accessible Surface Area	1	-

4.3.4 Environmental Features

To capture the effect of the surrounding atoms, which are known to play an important role[13,15,36], I extended the 80 features discussed in the previous section to compute an additional 400 properties weighted by spatial functions that provide a description of the molecular environment that is (globally) equivariant under translations and rotations[37,38].

Environmental descriptors are based on a series of radial functions that have previously been used to describe local atomic environments[38]. For a residue property P, environmental descriptors D are computed according to equation 4.3,

$$D_i(r_c, \eta, \mu) = \sum_{\substack{j=1 \\ j \neq i}}^{N} P_j \cdot e^{-\eta(r_{ij}-\mu)^2} f_c(r_{ij}) \tag{4.3}$$

Here, r_c, η, and μ are hyperparameters and describe the distance and weighting across which the environment is considered. The pairwise distances between the center of masses of residue i and j are given by r_{ij} and a cutoff function f_c is computed in equation 4.4.

$$f_c(r_{ij}) = \begin{cases} \frac{1}{2}[\cos(r_{ij}\pi/r_c) + 1] & \text{if } r_{ij} \leq r_c \\ 0 & \text{if } r_{ij} > r_c \end{cases} \tag{4.4}$$

In addition to equation 4.3, which represents an extrinsic summation, intrinsic environmental properties are also computed by normalizing the summation by the total property weights W.

$$W_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} e^{-\eta(r_{ij}-\mu)^2} f_c(r_{ij}) \tag{4.5}$$

For each property, a total of five different environmental descriptors were computed by varying hyperparameters, as detailed in table 4.3 and graphically shown in

Table 4.3: Parameters for NeCLAS's environmental descriptors. "Type" refers to the aggregation method for distributional properties

r_c	η	μ	Type
25	0.005	0	sum
18	0.05	0	mean
18	1	7.5	sum
18	1	10	sum
18	1	12.5	sum

figure 4.2.

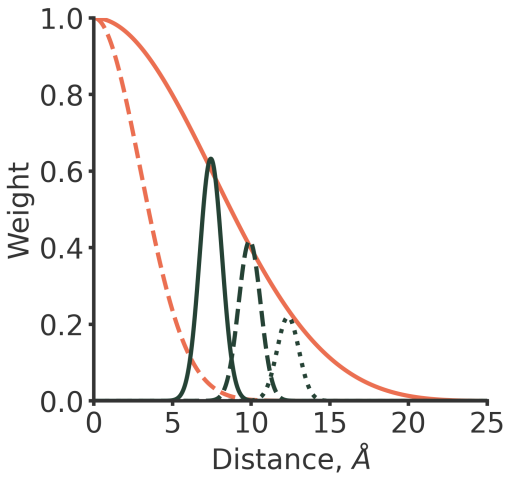

Figure 4.2: Weights used for computing environmental descriptors. Each curve represents a radial function with different hyperparameters. In descending order, pink solid, pink dashed, grey solid, grey dashed, grey dotted lines refer to weighting functions in table 4.3

No standard method exists for hyperparameter selection, but rather it is informed by knowledge of the system and obtaining a number of different coverage[38]. I chose 1.8 nm as the primary cutoff radius since it has been similarly employed with some success as a cutoff for Voronoi based environmental descriptors in similar protein-protein pairwise predictions[15] and nanoparticle machine learning studies[39]. Distances considered are from the center of masses, however, environmental interactions might occur between any atoms within two residues. Therefore, for a more complete description, I also consider at least one longer cutoff with an additional 0.7 nm which corresponds

approximately to the difference in position between the center of mass and the outer heavy atoms in my datasets, for the most common CG size. Beyond that, I choose multiple values of μ to capture a number of different positional distances between the origin and the cutoff radius. To avoid biasing the data towards bound structures, environmental descriptors are only calculated using a single structure, not information of the bound structure.

These descriptors are also used to smooth the NeCLAS predictions. A number of other protein interaction prediction methods[13,15] perform a basic smoothing of prediction results by considering nearby predictions. This has been empirically shown to improve results, as being surrounded by residues with a high probability of interaction is itself a good indicator of interaction[15]. To this end, I smooth my predictions by the weighted average prediction with weights determined by the environmental descriptors discussed above. I used equation 4.6.

$$P_i = (P'_i + D_i)/(W_i + 1) \tag{4.6}$$

where P_i is the smoothed prediction, D_i is given in equation 4.3, and W_i is given in figure 4.2. The unsmoothed prediction P'_i is given a maximum weight of 1. r_c, η, and μ are 25, 0.005, and 0 respectively.

4.3.5 Protein-Protein Dataset

For pairwise PPI predictions, I used the DBD (version 5)[17]. Unbound proteins were used for feature generation, and bound proteins were used to compute the ground-truth pairwise interactions. For each complex, I consider all combinations of residues between the two proteins. Due to the severe class imbalance (positive sample rate of 0.136%), I downsampled the training data so that there is one positive example for every three negative examples, providing a training dataset of approxi-

mately 83,000 pairwise interactions between CG sites (different validation splits cause the exact value to change). I did not alter the imbalance from the testing set to avoid biasing the data.

4.3.6 Protein-Nanoparticle Dataset

For the PNIs, I used the recent collection of crystallographic data by Costanzò *et al.*[23]. This collection contains approximately 40 unique structures; however, I removed files containing duplicate interactions or incomplete information, leaving 21 unique PNI complex pairs (table 4.4). Unbound proteins structures were taken from the RCSB database[40], while unbound nanoparticle structures were generated by relaxing the bounded configuration with the MMFF94 force field[41] in the absence of the protein. Two structures (5ET3 and 5N10) contain no equivalent solved structure, and therefore the bound structure was used.

4.3.7 Featurization and Labeling

Given a pair of nanostructures A and B, I computed all the pairwise combinations of one CG site from A and one from B. Each of these pairwise interactions was considered to be a possible interaction. For each pair of CG sites, the local and environmental residue features of both sites were concatenated to create a single input feature vector to the model, resulting in a training or testing sample. When assigning ground truth labels to both the training and testing data, two sites $S1, S2$ are considered to be interacting if and only if a heavy atom from site $S1$ falls within 0.6 nm of any heavy atom of site $S2$. When considering interface labels for PNIs, a residue is considered to be interacting if it participates in at least one pairwise interaction.

Table 4.4: Protein-nanoparticle pairs used in PNI testing set. The PDB ID represents the bound complex. The unbound protein ID is the PDB ID of the unbound protein used. The NP ID is either the chain (if a single letter) or RCSB ligand ID of the nanoparticle. The number of sites used in coarse-graining the nanoparticle is also provided.

PDB ID	Unbound Prot.	NP ID	Num. Sites
3BCD	3BCF	D	6
3CYU	1AVN	1CR	12
3EDK	3EDD	C	16
3TYI	2MHM	T3Y	8
3CZH	3C6G	C	7
4PRQ	5WRB	T3Y	8
5ET3	5ET3	60C	6
5N10	5N10	C8L	16
5LFT	2MHM	6VB	8
5LYC	2MHM	7AZ	12
5KPF	2MHM	6VJ	8
5OEH	3IQU	9SZ	6
5MKA	5MKB	B	8
6HAH	2MHV	FWQ	12
6HAJ	2MHV	EVB	16
6HA4	2MHV	T3Y	8
6EGY	2MHM	B4T	9
6RGI	5T8W	FWQ	12
6GL5	6F7Y	T3Y	8
6GD6	2MHM	EVB	16
6SUY	2MHM	LVT	8

4.3.8 Machine Lerning Details

The machine learning part of NeCLAS consists of a permutation invariant neural network inspired by the *Deep Sets* architecture[42]. The NeCLAS neural network was implemented by co-author Matt Raymond as an alternating structure of permutation variant networks and maxpool to derive a generalized, non-linear, permutation-invariant network. The exact architecture can be seen in figure 4.3, and was implemented using Tensorflow 2.9[43]. All neural networks (NNs) utilize variants of stochastic gradient descent, which have a probabilistic component that can provide slightly different results depending on the random seed used. The TensorFlow default of

32-bit floats for the PPI model, and 64-bit floats for the NeCLAS PNI model was used. To ensure accurate and unbiased performance estimates for protein-protein interaction, leave-one-protein-out cross validation was performed over 230 protein complexes, where I iterate over each complex P_i and remove all interactions involving P_i from the training dataset. The testing dataset is then constructed from all the interactions that include P_i. To provide robust estimates, all preprocessing steps are calibrated using only the training dataset, and twelve validation proteins were selected with replacement for early stopping. To construct the validation dataset, I partition the training dataset based on 3 complexity levels (as defined by DBD database) and 3 "family" bins (enzyme, antibody, and other interactions), and excise 2 proteins from each difficulty and each family. This randomization reduces human bias when selecting validation sets, while ensuring that the validation set is diverse enough to facilitate the early stopping of NeCLAS's permutation invariant NNs. Since the training of machine learning methods is influenced by its training set and initial state, for NeCLAS and *NoPair* the results show a distribution obtained from 250 initial conditions, namely 25 different training and validation sets, each with 10 different initial sets of model weights. However, for results to be reproducible, the same random seed must be used every time the entire cross-validation cycle is run. All protein models were trained using Adam[44] with a learning rate of 10^{-3} and a batch size of 256. For pairwise PNI predictions, I trained on the entire (downsampled) PPI dataset, and used the PNI dataset for testing. The same method was used for selecting PPI and PNI validation sets. To show that my features are truly general, my nanoparticle interaction model uses only proteins for training and validation datasets. Therefore, to prevent overfitting, the batch size is significantly increased to 2^{15} and reduced the model size as seen in figure 4.3. These modifications decrease training time, but also causes the model to occasionally get stuck in local minima. Similar to existing methods[45] (and unlike the protein-protein model), the

same fit is rerun multiple times during training. Typically, these models are all used during inference, with their predictions being aggregated to form a single prediction. However, after training, NeCLAS keeps only the model with the highest (protein) validation AUC. Since the NeCLAS model is small, this provides high AUC scores for protein-nanoparticle interaction predictions, while incurring minimal overhead during training and no overhead during inference.

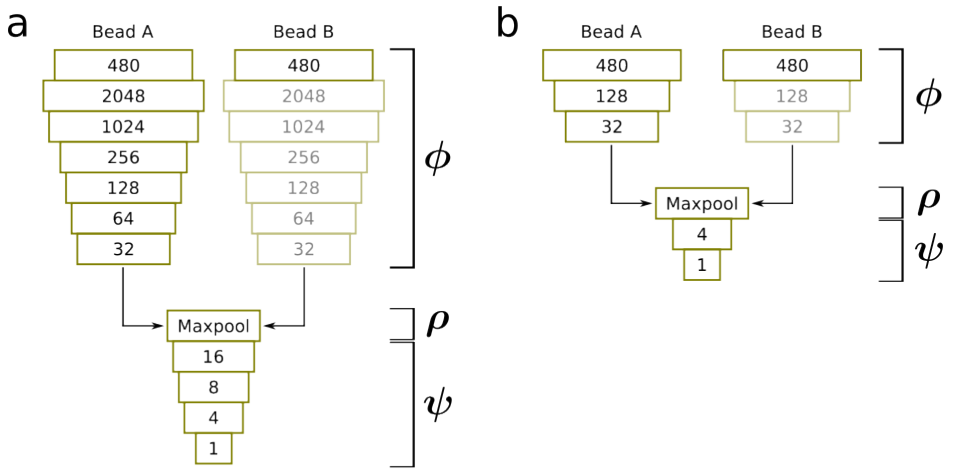

Figure 4.3: NeCLAS architecture. Architectural diagram of permutation invariant NN. Each numbered layer represents a dense layer with that many weights, except for the first layer, which represents the input size. Transparent layers indicate shared weights. a, protein network with 2 input sites and 1 output task, b, miniaturized version of model a for nanoparticle prediction.

Interface interaction predictions were obtained from pairwise predictions using a scoring function[13]. All predictions were smoothed[13,15]. When conversion between CG predictions and protein residues was needed, I assigned each heavy atom a prediction score equal to that of its corresponding CG site and computed the residue prediction as the mean of all its constituent atoms (excluding hydrogen).

4.3.9 Molecular Dynamics

For one of the tests, NeCLAS was validated using molecular dynamics of graphene quantum dots. To this end, NeCLAS predictions were used in order to parameterize a coarse-grained molecular dynamics force field. The specific molecular dynamics parameterization was performed by co-author Paolo Evlati using my predictions, however to contextualize the results, I discuss the procedure.

In the CG simulations, 8 particles were randomly placed in a box, and after an energy minimization, the system was run for 5 ns in a canonical ensemble at 300 K. Snapshots were taken from the last 500 ps of each run. Molecular dynamics simulations were performed with either Large-scale Atomic/Molecular Massively Parallel Simulator (CG simulations, software version 29 Sep 2021 - Update 2)[46] or Nanoscale Molecular Dynamics Program (all-atom simulations, software version 2.13)[47]. For intramolecular interactions in the CG simulations, harmonic potentials were used for bonds, angles, dihedral, and improper, using all atom equilibrium distance/angles as equilibrium values and constants of 150/75 kcal/mol/Å^2 for bonds, 100/50 kcal/mol for angles (in both cases the first value is for rigid aromatic atoms, the second for everything else), and 70/35/17.5 kcal/mol for dihedral and improper, based on the amount of atoms that were part of a rigid aromatic subgroup. For intermolecular interactions, equation 4.7 was used.

$$E(r) = 4\epsilon\sqrt{p}\left\{\left[\frac{(1-p)^2}{2} + \left(\frac{r}{\sigma}\right)^6\right]^{-2} - \left[\frac{(1-p)^2}{2} + \left(\frac{r}{\sigma}\right)^6\right]^{-1}\right\} \tag{4.7}$$

where p ($\in [0,1]$) is the prediction value from NeCLAS as a function of radius r (nm). σ (nm) and ϵ (kcal/mol) were kept the same for all the CG sites of a given GQD. σ was chosen to be 4 nm based on the distances observed in a previous work[48]. ϵ was estimated by matching the minimum for the potential in Eq. 4.7 for benzene-benzene interactions (NeCLAS prediction ≈ 0.2) with the energy value of the potential

for the closest minimum for the interactions between two CG benzene molecules, $\epsilon_{benzene} = 0.5$ kcal/mol (from[49]). The match produces a value of $\epsilon \approx 1.11$ kcal/mol. These two values were used for all the intermolecular potential, while p was obtained from NeCLAS's predictions.

For all atom simulations, the fully solvated GQDs are run in a canonical ensemble, starting from the final conformation produced in a previous work[48]. Complete details about the protocol and force field can be found there.

4.4 Results

4.4.1 Performance

To assess NeCLAS's ability to predict PNIs, I compared it with seven other methods: the recently published generalized method (Unified[45]), and six (non-partner specific) binding residue prediction methods, namely SPPIDER[36] (designed for PPIs), P2Rank[11] and COACH[50] (protein-ligand binding), DeepSite and DeepSurf (deep-learning-based binding pocket identification)[51,52], and Fpocket[53] (geometric pocket identification). Among these methods, only NeCLAS and Unified are explicitly designed for prediction of pairwise PNI. The other methods were selected based on their ability to produce localized PNI predictions, even if they were not originally developed for this purpose.

To quantify the performance of the different methods, I generated the receiver operating characteristic curve, which considers the prediction probability of all interactions between CG sites and computes the fraction of true-positive to false-positive interaction predictions at various discrimination thresholds. The area under this curve (AUC) is a binary prediction metric, and is commonly used in similar problems[12,13,15,16]. Typically, AUC weights all pairwise samples equally (AUC_{all}). Since nanoscale complexes can vary in size, it is necessary to reweight pairwise interactions

at the complex level to guarantee that each complex makes an equal contribution to the resulting metric. This produces a more realistic metric for model performance on an untested species, as larger complexes are not disproportionately weighted[16]. Thus, I also computed the AUC for each test sample in an individual complex as AUC_{comp}, and report statistics with respect to this metric as well (*e.g.* mean AUC_{comp}). The procedure carefully distinguishes $\text{AUC}_{\text{comp}}^{\text{pair}}$ from $\text{AUC}_{\text{comp}}^{\text{inter}}$. $\text{AUC}_{\text{comp}}^{\text{pair}}$ scores pairwise interaction predictions, while $\text{AUC}_{\text{comp}}^{\text{inter}}$ scores interface membership predictions. As NeCLAS and Unified do not predict interaction interfaces directly, I converted pairwise predictions to interface predictions using a scoring function[13], which considers interface membership using all possible interactions with different weights. Finally, to test the predictive importance of pairwise information, I included a non-partner-specific version of NeCLAS (*NoPair*), which comprises identical chemical features of the protein only, but omits nanoparticle features.

Performances for PNI predictions are shown in Fig. 4.4a. NeCLAS outperforms all competing methods in PNI prediction, and *NoPair* performs slight worse, suggesting that there is a performance benefit in including the representation of the partner molecules. A two-sided Mann-Whitney U test between the next highest performing method, SPPIDER, and *NoPair* suggests the difference is statistically significant with a p-value of 0.017. To elucidate the contribution of the environmental descriptors, I also performed predictions without environmental descriptors and observe that the $\text{AUC}_{\text{comp}}^{\text{inter}}$ decreases to 0.563 demonstrating their importance. This indicates that the environmental descriptors significantly contribute to the more complete representation provided by NeCLAS's chemical features. It is important to note that the dataset presents a challenge for all the methods, and given the limited number of nanoparticle-protein pairs, it is not surprising that NeCLAS's long tail is solely due to two complexes. These results suggest that adding structural information to the validation set from similar nanoparticles may help generalize the stopping criteria of

the neural network and improve NeCLAS performance.

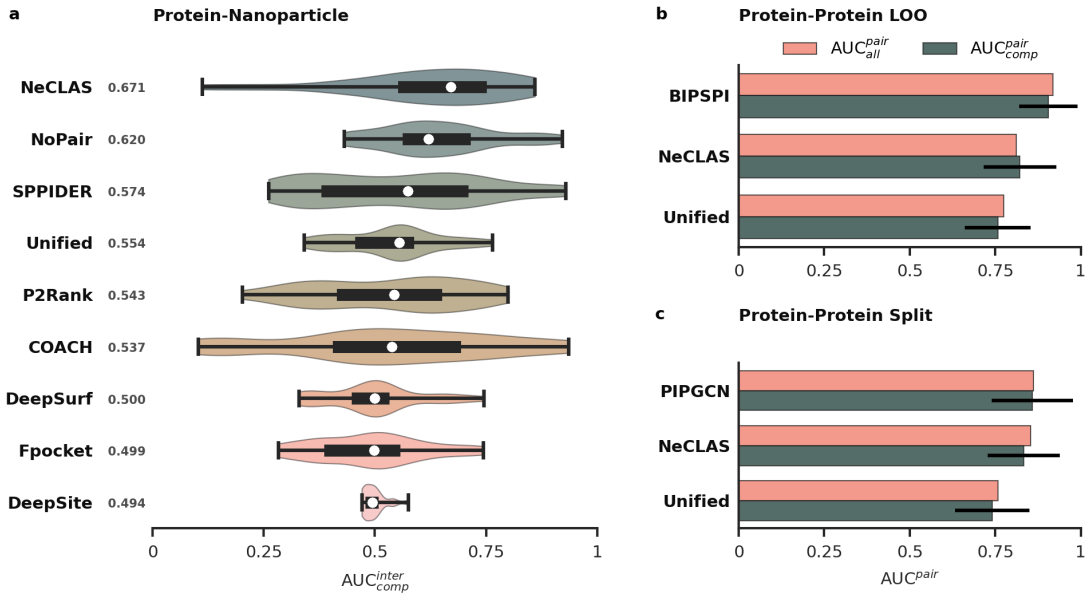

Figure 4.4: Predictive performances of NeCLAS compared to different methods. **a**, Distribution of the prediction performance for protein-nanoparticle interface interactions. Median is marked as a white circle and reported as a number near the method name. The thick black bar shows the 1st-3rd interquartile range. ($n = 21$ test complexes, 3,988 interface predictions total). NeCLAS and *NoPair* distributions are obtained by computing the median of each pair over 250 independent trials, providing $n = 5,250$ samples (250 trials $\times$ 21 nanoparticles). Distributions are colored to increase readability. **b**, Performances of different methods for leave-one-out cross validation with $n = 230$ test complexes (16,579,545 pairwise predictions) and **c**, the predefined DBD train-test split for protein-protein pairwise interactions ($n = 55$ test complexes). For **b** and **c**, green bars indicate the mean and black lines indicate the standard deviation of complex-wise predictions.

The issue of structural homology between training and testing datasets is a persistent problem in protein interaction predictions, leading to overly optimistic error estimates. Garcia *et al.*[13] addressed this issue by evaluating pairwise interactions only for protein pairs that do not share either a single Structural Classification of Proteins (SCOP) family or both SCOP families[54]. NeCLAS already meets the former criteria, since the chosen nanoparticles are not structurally homologous to any proteins. However, I took an even stricter approach by removing all trained and validation proteins

that share a single SCOP family with any of the proteins in the PNI test set. This test causes a negligible change in performance (median $\text{AUC}_{\text{comp}}^{\text{inter}} = 0.678$), showing no general effect.

Finally, despite its high degree of generality, NeCLAS achieves pairwise PPI prediction performance that is competitive to state-of-the-art protein-specific methods (Fig. 4.4b). To prove this result, I compared NeCLAS against a number of PPI methods which can provide partner specific pairwise interactions between residues. While other methods for predicting PPIs exist[9,12], most do not provide pairwise interaction predictions between sites, making a meaningful comparison impossible. I performed two tests using the protein dataset: leave-one-out cross-validation (LOOCV) at the protein complex level, and a predefined train-test split based on DBD versions. In LOOCV, I iteratively withhold a single protein complex for evaluation and use the remaining 229 complexes for training and validation. This approach avoids overly optimistic predictions that arise from mixing interactions in the training and testing set between nearby residues of the same protein complex. In the specified predefined split, I trained on DBD versions 1 to 4, and tested on DBD version 5. This split is used when comparing methods with significant computational overhead, like PIPGCN's[16], which renders LOOCV impractical. The results (Fig. 4.4b-c), show that NeCLAS PPI predictions fall just below BIPSPI[13], one of the leading methods in protein pairwise interaction predictions, and is comparable to PIPGCN. Similar to the PNI NeCLAS predictions, PPI prediction performance is largely independent of structural homology. When removing SCOP family homologs and applying leave-one-homology-out validation, the performance only decreases slightly ($\text{AUC}_{\text{all}}^{\text{pair}} = 0.770$ and median $\text{AUC}_{\text{comp}}^{\text{pair}} = 0.812$).

To further illustrate the potential of NeCLAS, below I describe in detail its application on three systems: nanoparticle tweezer and 14-3-3σ protein, carbon-based nanoparticle with amyloid fibrils, and organic quantum dots in water.

4.4.2 Molecular Tweezers

Supramolecular ligands (*e.g.* molecular tweezers) represent a promising way to modulate protein functions. They can be artificially synthesized with unique properties and recognition profiles towards amino acids and peptides, with the ability to bind to specific sites. Specifically, the interaction between a 14-3-3σ protein and the lysine-specific molecular tweezers shown in Fig. 4.5 have been characterized in detail both experimentally and computationally[55]. Therefore, it is an ideal test for pairwise interaction prediction models.

NeCLAS predictions (Fig. 4.5a,b) corroborate Bier *et al.*, indicating the critical role played by Lys214, as well as Leu218, Tyr213, and Thr217, which form a hydrophobic binding pocket, and Glu210 and Gln221, which provide hydrogen bond stabilization. Bier *et al.* derived some general principles characterizing the active binding site (Lys214) leveraging the fact that the protein has four other energetically possible, but non-binding, lysine residues, (Lys* in the following) and that several properties differentiate Lys214 from the other residues.

First, they determined that Lys214 and Lys* are more energetically likely to bind due to their protruding carbon side chains, which is captured by NeCLAS by the higher protrusion index value[28] and elongated structure of Lys* residues compared to other residues (Fig. 4.5c). Additionally, Bier *et al.* suggested that the difference between Lys214 and the other Lys* residues is caused by the nearby hydrophobic binding pocket and small number of close, positively charged functional groups, which destabilize the nanoparticle by forming external ion pairs between the nanoparticle phosphate groups and surrounding cations. These characteristics are captured by several environmental features of Lys214 that capture the effect of neighboring atoms. The hydrophobic pocket for Lys214, depicted by the total surface area of surrounding hydrophobic groups[29] and total surface area of surrounding hydrogen bonding groups[30], shows higher values than all other residues. Additionally, the envi-

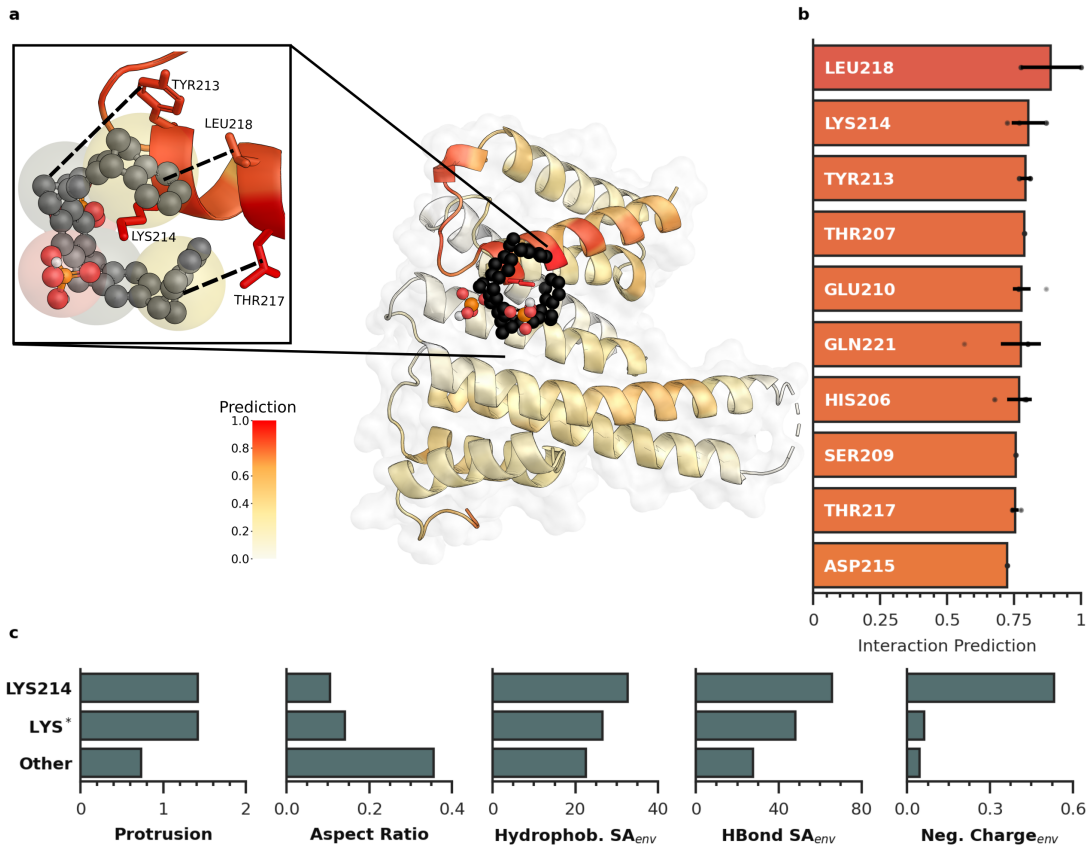

Figure 4.5: Interactions between molecular tweezers and the 14-3-3σ protein. **a**, Visual representation of the interaction location based on NeCLAS prediction; cutout highlights the lysine residue (Lys214) and the surrounding hydrophobic pocket. **b**, Top 10 interacting residues according to NeCLAS predictions. Bars indicate the mean and black lines indicate standard deviation of the prediction across atoms in a residue. The n for each residue is equal to the number of atoms in that residue. All unique predictions within a residue are overlaid by black circles. All atoms in a single CG site share the same prediction. **c**, Comparison of selected features for binding lysine Lys214), probable but non-binding lysines (Lys*, as defined by Bier *et al.*), and all the other residues. Protrusion and aspect ratio are features of the individual amino acid sites (deterministic), while the last three histograms refer to environmentally weighted features.

ronmentally weighted charge shows that Lys214 is surrounded by significantly more negatively charged atoms than other Lys* residues.

4.4.3 Bacterial Amyloid Fibrils

Nanoscale structural interactions exhibit complex, high-dimensional free energy surfaces, which are the product of dynamic molecular constraints and entropic factors. Molecular dynamics (MD) simulations can model such high dynamic processes evolving across relatively short time scales. One of such examples, are the interactions between phenol-soluble modulin (PSMα1) peptides and graphene quantum dots (GQDs)[56]. It has previously been shown that GQD nanoparticles dissolve biofilms via their interactions with PSMα1, a key constituent of the *Staphylococcus aureus* biofilm matrix, that assemble into amyloid fibers (Fig. 4.6a).

Characterizing these interactions via ML is challenging, and it is difficult to compare the predictions of ML and MD, especially since MD generates ensemble distributions of conformations. However, while most of the current datasets do not entirely capture the free energy landscape and the dynamics of a nanoscale system, ultimately, ML and MD are both (different) representations of the same physical system, and therefore it is ideally possible to find some correlation between the information generated with these methods. Specifically, I compared the interaction probabilities obtained from NeCLAS with the contact times (*i.e.* the time two CG sites spent within a 1 nm-distance) during MD simulations of the system composed of GQDs and PSMα1 as reported in Fig. 4.6b.

The figure shows that the model interaction confidence is generally correlated with contact times (Spearman coefficient, $r_s^{tot} = 0.907$). This trend is not simply due to strong interactions with the hydrophilic charged groups at the edges of the GQD, but rather a complex interplay of chemical properties. To support this idea, I chose a twelve site representation for the GQD, as it consistently partitions the molecule in two distinct classes of nearly identical internal and external sites. The former contains only matrix carbon atoms, while the latter includes edge carbons and outer functional groups. Despite the separation of two regions, the predictions for each subset still

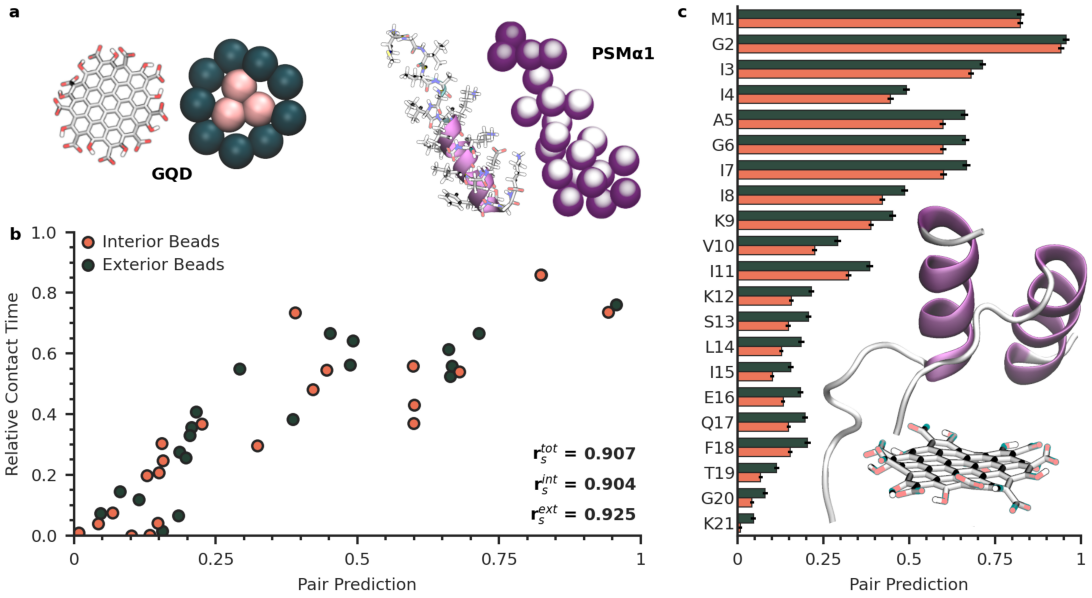

Figure 4.6: Interactions of PSMα1 1 and graphene quantum dot. **a**, All-atom and CG representation of the GQD (circumcoronene with alternating hydroxyl and carboxyl groups at the edges) and the PSMα1 1 peptide (5KHB). PSMα1 1 is purple to distinguish it from the GQD beads. **b**, Relation between MD contact time and predicted pairwise interactions between the PSMα1 residues and the GQD sites. Spearman correlation (r_s^{tot}) for all, interior only (r_s^{int}), and exterior only (r_s^{ext}) sites is reported. **c**, Interaction prediction between residues from PSMα1 1 and GQD interior (pink) and exterior (green) units, along with a snapshot of interactions observed during simulation. Bars represent the mean value, and the black lines indicate the standard error. $n = 3$ for interior beads (pink) and $n = 9$ for exterior beads (green).

shows a high Spearman correlation (external: $r_s^{ext} = 0.925$, internal: $r_s^{int} = 0.904$) with the contact time, despite the different properties of these subsets. Finally, I analyze the predictions for individual amino acids (Fig. 4.6c) to confirm the importance of the N-terminal residues (which have the highest interaction probability), likely due to the GQD's negative charge (dissociated carboxylic groups), in agreement with observations by Wang *et al.*[56]. Notably, these conclusions hold even when different definitions of contact time are used (see corresponding publication[1]).

4.4.4 Organic quantum dots

As a final example, I discuss the potential of pairwise interaction predictions to inform atomistic models (*e.g.* generate realistic conformation distributions or evaluate the aggregation of multiple nanoparticles). For this class of problems, many factors (*e.g.* thermal energy, solvent effects, entropic contributions), must be accounted for, requiring additional assumptions and data. Furthermore, the outputs of binary classification models cannot be directly interpreted as interaction strength; they are more readily conceptualized as model confidence. However, one would expect a well-informed model to assign low probabilities to weakly interacting pairs. Thus, this approach considers the interaction probability as being proportional to the interaction strength. Under this assumption, I use NeCLAS to tune the intensity of intermolecular forces of different GQDs in water to study their aggregation propensity.

Previously, using all-atom MD, I have reported the effect of the composition of edge groups present on GQDs and their tendency to aggregate in water[48]. Here, three types of GQDs are studied: one terminated with hydroxyl (g3OH), another with formyl (g3CHO), and one with an alternating (2:1 ratio) hydroxyl and cysteine groups (6C-g3OH). These nanoparticles (sized between 1.5 and 2 nm) were chosen as hydrophobic and hydrophilic forces are generally comparable, whereas for bigger structures, hydrophobic forces and water entropic exclusion increasingly dominate their behavior. Pairwise NeCLAS predictions (AUC^{pair}) were converted using a tunable repulsive-core potential with identical parameters for all GQDs sites, except for one that was taken from NeCLAS predictions, effectively converting them into a physically-meaningful potential (Fig. 4.7a,d).

Using these potentials, the dynamics of a few GQDs were simulated observing, for g3OH and g3CHO, the rapid formation of aggregates with similar structures to those observed in all-atom MD[48] (Fig. 4.7). Indeed, I observed close and parallel stacking

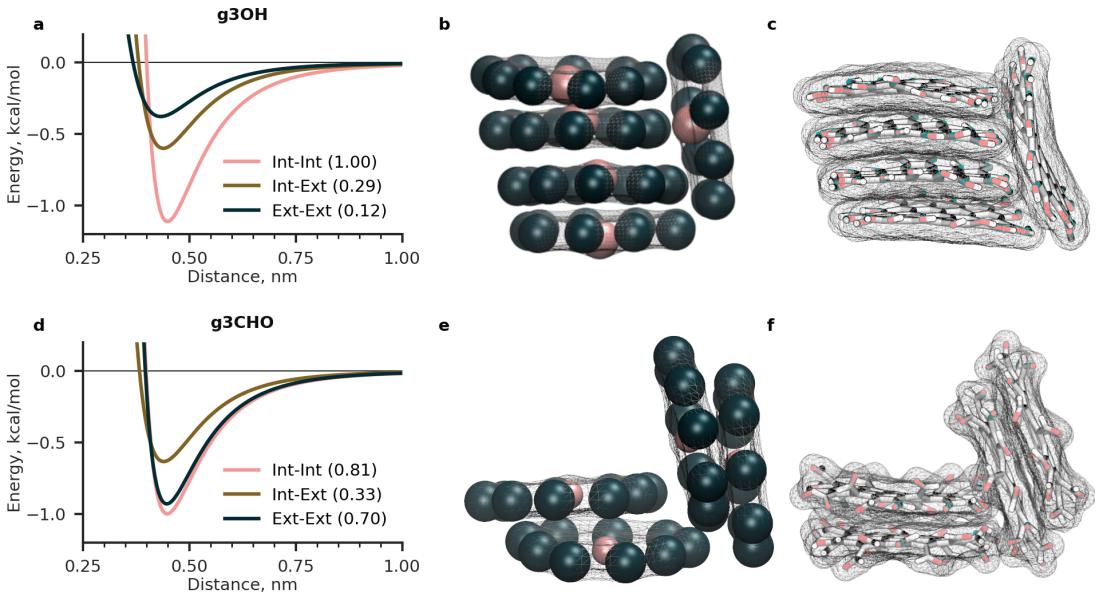

Figure 4.7: Predicted and simulated interactions of graphene quantum dots. **a**, pairwise interaction potential from NeCLAS mean predictions (in parentheses) for g3OH CG sites. **b**, Snapshot of g3OH from CG simulations modeled using NeCLAS predictions. **c**, Snapshot of all-atom g3OH simulations. **d**, Pairwise interaction potential from NeCLAS mean predictions (in parentheses) for g3CHO CG sites. **e**, Snapshot of g3CHO from CG simulations modeled using NeCLAS predictions. **f**, Snapshot of all-atom g3CHO simulations. Inner and outer beads are colored pink and green, respectively.

of the structures (see all-atom reference[48] for definition), and a similar lateral shift between consecutive stacking planes. 6C-g3OH, however, did not aggregate, also in agreement with all-atoms simulations and the experimental high solubility at pH 7[57].

The similarity between the crude CG and all-atom simulations, albeit qualitative, was obtained without fine-tuning of the CG potential, as these optimizations would obscure NeCLAS's contributions and a better agreement can be expected if additional optimizations are performed.

4.5 Discussion of NeCLAS

The results showed above illustrate how NeCLAS can predict pairwise PPIs and PNIs, and can be extended to predict nanoparticle-nanoparticle systems. Unlike most competing methods, NeCLAS forgoes protein-specific information, using descriptors that are common to all molecules. NeCLAS performance can be ascribed to two general design principles: structural simplification, operated through the CG representation, and the use of environmental features to capture the chemical neighborhood at different scales. The CG representation reduces physical (*e.g.* thermal vibration), observational (*e.g.* experimental and numerical error), and statistical (*e.g.* sampling size) uncertainties in the data, allowing efficient and robust model training. This approach is not needed for all problems, but it is critical for the data-limited applications typical of chemistry. In addition, coarse-graining reduces the computational requirements, which makes it possible to train on larger systems, and reconstruct the atomistic information when necessary[58]. As a low-dimensional representation results in a loss of information about local atomistic properties, which are not adequately captured through averaging, I utilize distribution statistics to express the local spatial distribution of each property. This approach provides a more nuanced characterization of the target distribution. However, the issue of long-range nanoscale interactions remain unresolved with this method. Such interactions typically decay rapidly in solvents with high relative permittivity, but can still significantly contribute to long-range organization. To capture these interactions, it is necessary to capture longer scale properties. This is a goal that NeCLAS accomplishes, as it employs environmental descriptors that incorporate information on the properties and positions of individual atoms. The choice of descriptors that are better suited to model pairwise interactions is still an open question. For proteins, I have shown that spatial features[59] alone are sufficient to obtain excellent PPI predictions. However, the much larger chemical and physical variety of nanoparticle properties requires

NeCLAS' more nuanced approach, as nanoparticles with similar structures, but different properties, are possible. Indeed, even if structural information is sufficient for some classes of molecules for which the limited chemical variety results in a correlation between atomic species and spatial organization, they become insufficient to distinguish wider classes of systems. One such examples, are fullerenes, which can gain different amounts of charge while in water without a relevant change in structure[60]. NeCLAS predicts a marked difference in the interactions between a fullerene organizing protein (Protein Data Bank (PDB) ID: 5ET3)[61] and a neutral C60 fullerene (as it is commonly modelled) or a negatively charged one (average experimental charge of -2 elementary charges).

As ML-based pipelines become increasingly prominent in modeling scientific data, it is essential to observe several good practices, as discussed above. Although some of these practices are intuitive, others, such as the effect of data symmetries on model performance and reliability, are more subtle but equally crucial. Many interaction-prediction methods use ensemble-based models (*e.g.* XGBoost)[13,15] or dense neural networks[45] to predict interaction interfaces or pairwise interactions. However, both of these methods are permutation variant, adding artificial ordering to a problem that is inherently unordered. In doing so, these methods violate the guiding principle that subjective ordering has no bearing on the behavior of a physical system. Further, these models produce unstable prediction results, as the hypothesis space of an over-parameterized model may contain many permutation variant functions that fit the training data (appendix B). NeCLAS avoids this problem by using a permutation invariant neural network inspired by the *Deep Sets*[42] architecture.

It is important to note that NeCLAS is built on a flexible and versatile predictive framework, which may result in slightly reduced performance when focusing on a narrow domain. Therefore, in cases where I only consider PPI with minimal conformation changes, geometric or template-based docking methods[9] may be more

suitable. Nonetheless, none of these methods possess NeCLAS's capability to accurately extend predictions to a wider range of nanoscale interactions and leverage datasets comprising diverse materials and conditions.

Most of the unrealized potential of NeCLAS stems from limitations in the available data. A greater variety of curated nanoparticle-nanoparticle interactions, and a diverse sample of sizes (*e.g.* larger ionic and colloidal nanoparticles), atom types, and solvents would significantly enhance the versatility of this tool. NeCLAS (and most other methods) tacitly assume that the species of interest are largely soluble in water. Under these conditions, many of the forces that govern protein complexes are also present in interactions between proteins and nanoparticles[23], as water solubility strongly limits the chemical properties of exposed nanoparticle surfaces. Different solvents (*e.g.* polymeric host-guest systems) not only have different ability to stabilize ionic groups or form hydrogen bonds, but also a different propensity than water to solvate species based on their size. The above limitations, can only be addressed by increasing the availability of curated data from spectroscopic data and simulations that go beyond PPIs. However, the inclusion of multiscale properties and strategies to deal with data uncertainty, as I do here with environmental descriptors and CG, remain a necessary requirement until radically larger datasets and computational power become available. For this reason, as more structural information and databases for nanoscale species emerge, I expect that this approach will prove to be a valuable technique for operating across different molecular domains and nanoparticle interaction problems.

4.6 References

[1] Jacob C. Saldinger, Matt Raymond, Paolo Elvati, and Angela Violi. Domain-agnostic predictions of nanoscale interactions in proteins and nanoparticles. *Nature Comp. Sci.*, 3:393–402, 2023.

[2] Goutam Ghosh and Lata Panicker. Protein–nanoparticle interactions and a new insight. *Soft Matter*, 17(14):3855–3875, 2021.

[3] Kristen A. Russ, Paolo Elvati, Tina L. Parsonage, Alyssa Dews, James A. Jarvis, M. Ray, B. Schneider, P.J.S. Smith, P.T.F. Williamson, Angela Violi, and Martin A. Philbert. C60 fullerene localization and membrane interactions in raw 264.7 immortalized mouse macrophages. *Nanoscale*, 8:4134–4144, 2016.

[4] Changjiang Liu, Paolo Elvati, Sagardip Majumder, Yichun Wang, Allen P. Liu, and Angela Violi. Predicting the Time of Entry of Nanoparticles in Lipid Membranes. *ACS Nano*, 13(9):10221–10232, 2019.

[5] Tony Pawson and John D. Scott. Signaling through scaffold, anchoring, and adaptor proteins. *Science*, 278(5346):2075–2080, 1997.

[6] Michael Holzinger, Alan Le Goff, and Serge Cosnier. Nanomaterials for biosensing applications: A review. *Frontiers in Chemistry*, 2, 2014.

[7] Sang-Ho Cha, Jin Hong, Matt McGuffie, Bongjun Yeom, J. Scott VanEpps, and Nicholas A. Kotov. Shape-Dependent Biomimetic Inhibition of Enzyme by Nanoparticles and Their Antibacterial Activity. *ACS Nano*, 9(9):9097–9105, 2015.

[8] Stewart A. Adcock and J. Andrew McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chemical Reviews*, 106(5):1589–1615, 2006.

[9] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The HDOCK server for integrated protein–protein docking. *Nature Protocols*, 15(5):1829–1852, 2020.

[10] Sangsoo Lim, Yijingxiu Lu, Chang Yun Cho, Inyoung Sung, Jungwoo Kim, Youngkuk Kim, Sungjoon Park, and Sun Kim. A review on compound-protein interaction prediction methods: Data, format, representation and model. *Computational and Structural Biotechnology Journal*, 19:1541–1556, 2021.

[11] Radoslav Krivák and David Hoksza. P2Rank: Machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(39), 2018.

[12] Pablo Gainza, Freyr Sverrisson, Federico Monti, Emanuele Rodolà, Davide Boscaini, Michael M. Bronstein, and Bruno E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[13] Ruben Sanchez-Garcia, C. Sorzano, J M Carazo, and Joan Segura. Bipspi: a method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics*, 35(3):470–477, 2019.

[14] Bowen Dai and Chris Bailey-Kellogg. Protein interaction interface region prediction by geometric deep learning. *Bioinformatics*, 37(17):2580–2588, 2021.

[15] Fayyaz ul Amir Afsar Minhas, Brian J. Geiss, and Asa Ben-Hur. Pairpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins*, 82:1142–1155, 2014.

[16] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. *NIPS*, 30, 2017.

[17] Thom Vreven, Iain H. Moal, Anna Vangone, Brian G. Pierce, Panagiotis L. Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A. Bates, Juan Fernandez-Recio, Alexandre M.J.J. Bonvin, and Zhiping Weng. Updates to the integrated protein–protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2. *Journal of Molecular Biology*, 427(19):3031–3041, 2015.

[18] Marco P. Monopoli, Christoffer Åberg, Anna Salvati, and Kenneth A. Dawson. Biomolecular coronas provide the biological identity of nanosized materials. *Nature Nanotechnology*, 7(12):779–786, 2012.

[19] Matthew R. Findlay, Daniel N. Freitas, Maryam Mobed-Miremadi, and Korin E. Wheeler. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. *Environmental Science: Nano*, 5(1):64–71, 2018.

[20] Nicholas Ouassil, Rebecca L. Pinals, Jackson Travis Del Bonis-O'Donnell, Jeffrey W. Wang, and Markita P. Landry. Supervised learning model predicts protein adsorption to carbon nanotubes. *Science Advances*, 8(1):eabm0898, 2022.

[21] Jimi M. Alex, Martin L. Rennie, Sylvain Engilberge, Gábor Lehoczki, Hajdu Dorottya, Ádám Fizil, Gyula Batta, and Peter B. Crowley. Calixarene-mediated assembly of a small antifungal protein. *IUCrJ*, 6(2):238–247, 2019.

[22] Jordan J. Clark, Zachary J. Orban, and Heather A. Carlson. Predicting binding sites from unbound versus bound protein structures. *Scientific Reports*, 10(1):243–252, 2020.

[23] Luigi Di Costanzo and Silvano Geremia. Atomic details of carbon-based nanomolecules interacting with proteins. *Molecules*, 25(15):3555, 2020.

[24] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, 1993.

[25] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, volume 07, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[26] Anton Arkhipov, Peter L. Freddolino, and Klaus Schulten. Stability and Dynamics of Virus Capsids Described by Coarse-Grained Modeling. *Structure*, 14(12):1767–1777, 2006.

[27] Michel F. Sanner, Arthur J. Olson, and Jean-Claude Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38(3):305–320, 1996.

[28] Alessandro Pintar, Oliviero Carugo, and Sandor Pongor. CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, 18(7):980–984, 2002.

[29] David T. Stanton and Peter C. Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, 1990.

[30] David T. Stanton, Leanne M. Egolf, Peter C. Jurs, and Martin G. Hicks. Computer-assisted prediction of normal boiling points of pyrans and pyrroles. *Journal of Chemical Information and Computer Sciences*, 32(4):306–316, 1992.

[31] Takeshi Kawabata. Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins: Structure, Function, and Bioinformatics*, 78(5):1195–1211, 2010.

[32] Roberto Todeschini and Paola Gramatica. The Whim Theory: New 3D Molecular Descriptors for Qsar in Environmental Modelling. *SAR and QSAR in Environmental Research*, 7(1-4):89–115, 1997.

[33] Todd J. Dolinsky, Paul Czodrowski, Hui Li, Jens E. Nielsen, Jan H. Jensen, Gerhard Klebe, and Nathan A. Baker. PDB2PQR: Expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(Web Server):W522–W525, 2007.

[34] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[35] Johann Gasteiger and Mario Marsili. A new model for calculating atomic charges in molecules. *Tetrahedron Letters*, 19(34):3181–3184, 1978.

[36] Aleksey Porollo and Jarosław Meller. Prediction-based fingerprints of protein-protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 66(3):630–645, 2006.

[37] Jorg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011.

[38] Michael Gastegger, Ludwig Schwiedrzik, Marius Bittermann, Florian Berzsenyi, and Philipp) Marquetanda. wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials. *J. Chem. Phys.*, 148, 2018.

[39] Xiliang Yan, Alexander Sedykh, Wenyi Wang, Xiaoli Zhao, Bing Yan, and Hao Zhu. *In silico* profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale*, 11(17):8352–8362, 2019.

[40] Helen Berman, Kim Henrick, and Haruki Nakamura. Announcing the worldwide Protein Data Bank. *Nature Structural & Molecular Biology*, 10(12):980–980, 2003.

[41] Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, 1996.

[42] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[43] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[45] Minjeong Cha, Emine S.T. Emre, Xiongye Xiao, Ji-Young Kim, Ppaul Bogdan, J. Scott VanEpps, Angela Violi, and Nicholas A. Kotov. Unifying structural descriptors for biological and bioinspired nanoscale complexes. *Nature Computational Science*, 2:243–252, 2022.

[46] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1–19, 1995.

[47] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[48] Paolo Elvati, Elizabeth Baumeister, and Angela Violi. Graphene quantum dots: effect of size, composition and curvature on their assembly. *RSC Advances*, 29, 2017.

[49] Paolo Elvati. *Computer Simulations of Fuel Cells: Modeling of Surfactants and Polymeric Membranes for Fuel Cells Applications*. LAP LAMBERT Academic Publishing, Saarbrücken, Germany, 2012.

[50] Jianyi Yang, Ambrish Roy, and Yang Zhang. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)*, 29(20):2588–2595, 2013.

[51] J Jiménez, S Doerr, G Martínez-Rosell, A S Rose, and G De Fabritiis. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017.

[52] Stelios K Mylonas, Apostolos Axenopoulos, and Petros Daras. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics*, 37(12):1681–1690, 2021.

[53] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10(1):168, 2009.

[54] Antonina Andreeva, Eugene Kulesha, Julian Gough, and Alexey G Murzin. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, 48(D1):D376–D382, 2019.

[55] David Bier, Rolf Rose, Kenny Bravo-Rodriguez, Maria Bartel, Juan Manuel Ramirez-Anguita, Som Dutt, Constanze Wilch, Frank-Gerrit Klärner, Elsa Sanchez-Garcia, Thomas Schrader, and Christian Ottmann. Molecular tweezers modulate 14-3-3 protein–protein interactions. *Nature Chemistry*, pages 234–239, 2013.

[56] Yichun Wang, Usha Kadiyala, Qu Zhibei, Paolo Elvati, Christopher Altheim, Nicholas A. Kotov, Angela Violi, and J. Scott VanEpps. Anti-biofilm activity of graphene quantum dots via self-assembly with bacterial amyloid proteins. *J. Phys. Chem. A*, 13(4):4278–4289, 2019.

[57] Nozomu Suzuki, Yichun Wang, Paolo Elvati, Zhi-Bei Qu, Kyoungwon Kim, Shuang Jiang, Elizabeth Baumeister, Jaewook Lee, Bongjun Yeom, Joong Hwan Bahng, Jaebeom Lee, Angela Violi, and Nicholas A. Kotov. Chiral Graphene Quantum Dots. *ACS Nano*, 10(2):1744–1755, 2016.

[58] Sergei Izvekov and Gregory A. Voth. A multiscale coarse-graining method for biomolecular systems. *The Journal of Physical Chemistry B*, 109(7):2469–2473, 2005.

[59] Mayank Baranwal, Abram Magner, Jacob Saldinger, Emine S. Turali-Emre, Shivani Kozarekar, Paolo Elvati, J. Scott VanEpps, Nicholas A. Kotov, Angela Violi, and Alfred O. Hero. Struct2graph: A graph attention network for structure based predictions of protein-protein interactions. *bioRxiv*, 2020.

[60] Shigeru Deguchi, Rossitza G. Alargova, and Kaoru Tsujii. Stable Dispersions of Fullerenes, C60 and C70, in Water. Preparation and Characterization. *Langmuir*, 17(19):6013–6017, 2001.

[61] Kook-Han Kim, Dong-Kyun Ko, Yong-Tae Kim, Nam Hyeong Kim, Jaydeep Paul, Shao-Qing Zhang, Christopher B. Murray, Rudresh Acharya, William F. DeGrado, Yong Ho Kim, and Gevorg Grigoryan. Protein-directed self-assembly of a fullerene crystal. *Nature Communications*, 7(1):11429, 2016.

CHAPTER V

Conclusion

5.1 Concluding Remarks

In this work I show how computational methods can be applied to study a variety of nanoparticle interactions. The main products of my thesis are as follows:

- I developed a computational workflow that stochastically models the chemical interactions leading to the growth of nanoparticles and interprets these simulations with a set of numerical descriptors.
- I created a method to study the physical interactions of nanoparticles that utilizes molecular dynamics to generate free energy surfaces. These simulations are then coupled with a machine learning approach to efficiently encompass the physical interactions of millions of unique structures.
- I created a general nanoscale interaction prediction tool that uses a coarse-grained representation and neural network. This tool operates across multiple nanoparticle domains, accurately predicting a diverse array of nanoparticle interactions.

In the first application, I show how these methods can be applied to improve our understanding about how, in flame systems, combustion nanoparticle precursors grow

through chemical interactions with small molecules in the gas-phase. *kMC* simulations and molecular descriptors demonstrate how computational methods can closely match experimental observations, quantify how different PAC properties develop in the flame, and reveal the effects of different flame environments on the PAC chemical space. Meanwhile machine learning can be applied to relate these PAC properties to other more complex combustion nanoparticle formation processes that would otherwise be difficult to directly measure. I show through my *kMC* simulations, across five different flame systems, that as a result of these chemical interactions, there exist millions of unique PACs with properties including oxygenation, five-membered rings, aliphatic chains, and curvature. These diverse properties suggest that stabilomer hydrocarbon PACs do not represent the true chemical space in these flame systems. Models which do not account for these properties will suffer in accuracy because they do not properly account for the chemical growth of these PACs and the downstream interactions in which these structures participate.

Computational methods can also be applied to explain the physical interactions which are instrumental to transition these PACs into larger combustion nanoparticles. Enhanced molecular dynamics can offer information about the kinetics and thermodynamics of this physical aggregation as they produce a free energy landscape. Machine learning can be applied to these nanostructures to efficiently extend this atomistic simulation data to the millions of unique structures which are observed in real flame systems. The machine learning methods employed also improve our understanding of these physical interactions as part of the prediction task they automatically select properties which contribute to aggregation. The equilibrium constants and rates derived from these free energy values are expected to remove significant challenges in predicting combustion nanoparticle inception as they provide a means to understand the effects of different properties on PAC aggregation and quantitatively relate these properties to a large, complex dataset of PACs found in real flame systems.

Finally, I demonstrate how computational methods can enable a better understanding of the function of nanoparticles through their interactions. I introduce a general nanoscale interaction prediction tool called NeCLAS which uses coarse-graining and machine learning to accurately predict interaction locations between two nanostructures. This generalized scheme enables NeCLAS to be trained on interactions where data is available and characterize nanoparticle interactions where data would otherwise be difficult to produce. I demonstrate how in a biological context, NeCLAS can learn from protein-protein interaction datasets to predict the interaction sites and mechanisms of protein-protein, protein-nanoparticle, and nanoparticle-nanoparticle interactions. I find that this general representation has many desirable features in nanoscale prediction including being able to be derived only only from atomic coordinates, reproducing natural symmetries and chemistries of nanoparticles, being invariant to translation, rotation, and pairing order, being insensitive to minor atomic fluctuations, having the ability to operate on both bound and unbound structures, and considering chemistry at multiple length scales. This offers the potential to better understand the processes in which these nanoparticles participate, to design nanoparticles for specific biological applications, and to reduce the cost and increase the accuracy of atomistic simulations of nanostructures.

5.2 Future Directions

Going forward, there are multiple areas where this work can be extended in the future in order to provide novel insights into nanoparticle interactions. The first area is to combine and extend the work in chapters II and III to provide a comprehensive model for predicting the nanoscale interactions leading to the early growth of combustion nanoparticles. The second area involves improving the machine learning methods contained in chapters III and IV by identifying a more complete and optimal set of descriptors. The final area involves extending the computational methods

discussed in this work to new nanoparticle systems.

5.2.1 Comprehensive Model of Combustion Nanoparticle Inception

While this work has revealed insights into the chemical interactions leading to the growth of PACs and the physical interactions where these PACs transition into larger nanostructures, additional work is needed to create a comprehensive model of combustion nanoparticle inception. While my findings have addressed two very important mechanisms in this process, there are still mechanisms which must be properly captured and integrated into the existing findings. For example, there exists a series of chemical reactions where radical electrons on PACs form bonds with other PACs, stabilizing physically aggregated structures[1]. Atomistic simulations would need to be applied to better understand the kinetics of this process and how it is affected by intermolecular spacing and resonantly stable electrons[2]. Machine learning would also likely be needed to extend these findings to the large number of possible PACs which might be present in flame environments. Furthermore, the works from chapters II and III must be combined together along with these additional growth mechanisms to provide a unified model of how these interactions contribute to nanoparticle inception. These more accurate kinetic and thermodynamic parameters can be applied to improve inception simulations which have previously been carried out with *kMC* and deterministic kinetic simulations[3–5]. These improved models should be validated on experimentally measured parameters such as precursor concentrations, soot volume fraction, and particle size distribution.

5.2.2 Improved Chemical Descriptors

Properly selecting chemical descriptors is integral to all the studies in this work as they provide a means to extract insights from atomistic simulations and ensure that machine learning is based on the underlying physics and remains interpretable.

While the descriptors used in this work are wide-ranging and align with chemical intuition, they are by no means exhaustive and computational methods can be improved by considering new descriptors and systematically selecting the optimal descriptors for each task. One area where a number of descriptors could be added is from experimental measurements. Most of the descriptors used in this work were entirely computationally derived, however, there has been recent findings in other domains which suggest that experimental measurements can be used in tandem with computational descriptors to improve accuracy of machine learning models[6]. In addition to adding new descriptors, automatic selection of descriptors can also greatly enhance interpretibility. While the Lasso methods discussed in chapter III achieve some feature selection, it is limited to a linear combination of features and additional work can be done to extend feature selection to other interactions with stronger non-linear relationships. Furthermore, a method such as NeCLAS which is designed for general prediction could be improved by developing a method to select an optimal set of descriptors which can represent both unique and common chemistry across multiple different length scales and nanoparticle domains.

5.2.3 Extending Computational Frameworks to New Systems

The final area to build upon this work to is to extend the methods discussed in this thesis to new systems. While the gas-phase studies in this work focus on carbon nanoparticles in flame environments, there are a large number of other gas-phase nanoparticles such as silicon and alumina nanoparticles which grow through a combination of physical and chemical interactions[7,8]. Each system presents its own unique chemistries and challenges. However, the atomistic simulation methods in this work provide a means to simulate both physical and chemical interactions, the descriptors in this work can numerically represent the chemistry of a nanoparticle, and machine learning offers a method to efficiently predict a large number of nanoscale interactions

when there is a combinatorial complexity of possible interacting molecules. Furthermore, concerning NeCLAS, the work in this thesis focused on validating predictions against known interactions, but future work should extend NeCLAS to new systems by predicting unknown protein and nanoparticle interactions. This can then be used as a tool to guide future studies by identifying promising nanostructures for specific functions and providing initial configurations for molecular dynamics simulations. The datasets which NeCLAS is trained on can also be updated as more data becomes available in other nanoscale domains such as metal nanoparticles.

5.3 Final Remarks

The above findings provide significant insights into a variety of different nanoscale systems. In these applications, difficulties related to the complexity of nanoparticle systems, the lack of existing datasets, and unique chemistries of these nanoparticles are most effectively addressed by applying multiple computational techniques together. I hope that this thesis provides a framework to overcome these challenges by applying atomistic simulation, numerical descriptors, and machine learning to gain scientific insights and quantify nanoparticle interactions.

5.4 References

[1] Hai Wang and Michael Frenklach. A detailed kinetic modeling study of aromatics formation in laminar premixed acetylene and ethylene flames. *Combust. Flame*, 110(1):173–221, 1997.

[2] K. Olaf Johansson, Tyler Dillstrom, Paolo Elvati, Matthew F. Campbell, Paul E. Schrader, Denisia M. Popolan-Vaida, Nicole K. Richards-Henderson, Kevin R. Wilson, Angela Violi, and Hope A. Michelsen. Radical–radical reactions, pyrene nucleation, and incipient soot formation in combustion. *Proc. Combust. Inst.*, 36(1):799–806, 2017.

[3] Abhijeet Raj, Matthew Celnik, Raphael Shirley, Markus Sander, Robert Patterson, Richard West, and Markus Kraft. A statistical approach to develop a detailed soot growth model using PAH characteristics. *Combust. Flame*, 156(4):896–913, 2009.

[4] Paolo Elvati, V. Tyler Dillstrom, and Angela Violi. Oxygen driven soot formation. *Proceedings of the Combustion Institute*, 36(1):825–832, 2017.

[5] Meghdad Saffaripour, Mohammadreza R. Kholghy, Seth B. Dworkin, and Murray J. Thomson. A numerical and experimental study of soot formation in a laminar coflow diffusion flame of a Jet A-1 surrogate. *Proc. Combust. Inst.*, 34(1):1057–1065, 2013.

[6] Samantha Stuart, Jeffery Watchorn, and Frank X. Gu. Sizing up feature descriptors for macromolecular machine learning with polymeric biomaterials. *npj Comput Mater*, 9(102), 2023.

[7] Xuetao Shi, Paolo Elvati, and Angela Violi. On the growth of si nanoparticles in non-thermal plasma: physisorption to chemisorption conversion. *Journal of Physics D: Applied Physics*, 54(36):365203, 2021.

[8] Tue Johannessen, Sotiris E. Pratsinis, and Hans Livbjerg. Computational fluid-particle dynamics for the flame synthesis of alumina particles. *Chemical Engineering Science*, 55(1):177–191, 2000.

APPENDICES

APPENDIX A

Number of Potential Binary Interactions based on System Size

Given a system with N unique interacting species, the number of potential unique binary inter-molecular interactions which can occur is obtained by considering the homo-interactions with species of the same kind and hetero-interactions with species of a different kind.

The number of homo-interactions is N.

The number of hetero-interactions is given by:

$Cr(N, 2)$

Which can be re-written as:

$\frac{N!}{2!(N-2)!}$

Which simplifies to:

$\frac{N^2-N}{2}$

Thus the total number of potential homo- and hetero-interactions is:

$\frac{N^2-N}{2} + N$

Or in more concise terms:

$\frac{N^2+N}{2}$

APPENDIX B

Permutation Invariance

NeCLAS is a permutation invariant neural network inspired by the *Deep Sets* architecture (see main text). Given CG sites A, B which yield descriptors $\mathbf{a}, \mathbf{b}$, we apply a permutation variant function $\boldsymbol{\phi}$ element-wise to the block vector $[\mathbf{a}\ \mathbf{b}]$ to produce a permutation-equivariant function $\hat{\boldsymbol{\phi}}([\mathbf{a}\ \mathbf{b}]) := [\boldsymbol{\phi}(\mathbf{a})\ \boldsymbol{\phi}(\mathbf{b})]$. We then apply a permutation invariant aggregator function $\boldsymbol{\rho}$ and a permutation variant function $\boldsymbol{\psi}$, resulting in a non-trivial permutation invariant function:

$$\boldsymbol{\sigma}([\mathbf{a}\ \mathbf{b}]) := \boldsymbol{\psi}(\boldsymbol{\rho}([\boldsymbol{\phi}(\mathbf{a})\ \boldsymbol{\phi}(\mathbf{b})])) \tag{B.1}$$

In this case, $\boldsymbol{\rho}$ is `Maxpool` and $\boldsymbol{\phi}, \boldsymbol{\psi}$ are Multilayer Perceptrons. NeCLAS uses ReLU activation for all layers except for the final layer of $\boldsymbol{\psi}$, which has sigmoid activation to enable binary predictions.

The practical implications of permutation variance is explored by comparing the permutation-invariant network of NeCLAS against permutation variant XGBoost models. One XGBoost model is trained on the standard CG sites' ordering $[\mathbf{a}\ \mathbf{b}]$, then evaluated it on both orderings ($[\mathbf{a}\ \mathbf{b}]$ and $[\mathbf{b}\ \mathbf{a}]$). Figures B.1 and B.2 show that XGBoost exhibits unstable predictions and a marked performance decrease when sites are permuted. Such permutations can even change the AUC of an individual complex

by as much as 0.49. To confirm that NeCLAS is permutation invariant, the same evaluation is performed on the permutation-invariant neural network and it is found that the ordering has no effect on the model outputs.

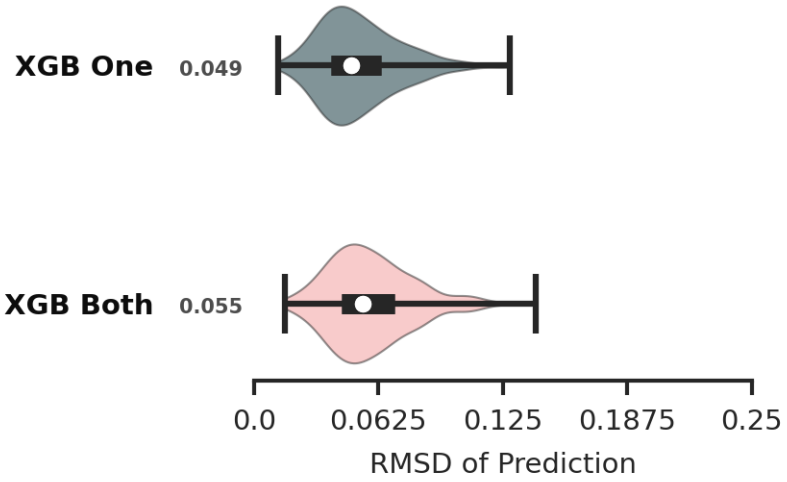

Figure B.1: Permutation Variance of XGBoost Prediction. The root mean squared deviation (RMSD) between predictions over the range $[0, 1]$ for feature vectors [**a b**] and [**b a**] for each protein complex ($n = 230$ protein-protein complexes). Medians are by the captions in grey text. We only show XGBoost trained on one direction and XGBoost trained on both directions, since the predictions of the permutation invariant network have a RMSD of 0. Colors are for visual distinction.

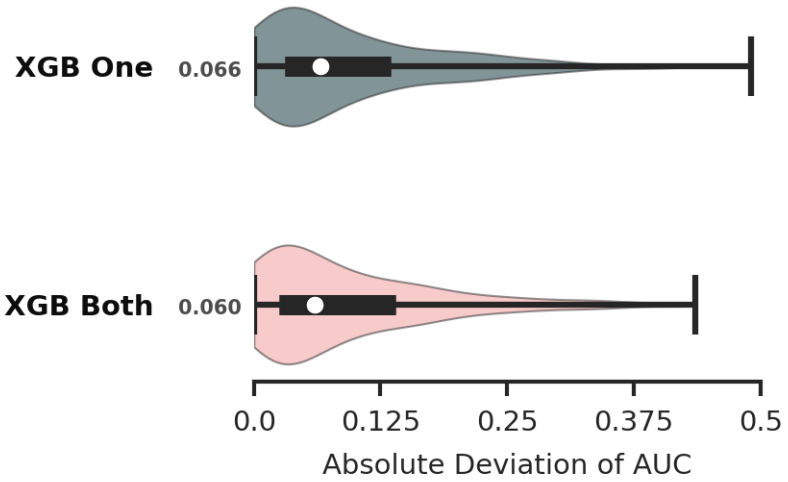

Figure B.2: Permutation Variance of XGBoost AUC. This figure shows the absolute deviation between the AUC of the predictions for [**a b**] and the AUC of the predictions for [**b a**] for each protein complex ($n = 230$ protein-protein complexes. Medians are by the captions in grey text. The permutation-invariant network is not included, as it achieved 0 deviation of AUC for each complex. Colors are for visual distinction.