BRIDGING GAPS IN MULTI-SCALE MATERIALS MODELING WITH MACHINE AND TRANSFER LEARNING

by

Zachary D. McClure

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Materials Engineering

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Alejandro Strachan, Chair

School of Materials Engineering

Dr. Eric Kvam

School of Materials Engineering

Dr. Michael S. Titus

School of Materials Engineering

Dr. David P. Adams

Sandia National Laboratory

Approved by:

Dr. David F. Bahr

To each person who has taken the time to help mold me into who I am: my family for the foundations they provided, the friends for their moments shared, and the professional leaders for their guidance.

ACKNOWLEDGMENTS

The work done throughout this thesis would not have been possible without the help and support of my advisor Prof. Alejandro Strachan. The relationship between a student and advisor, to me, is one akin to an academic apprenticeship. Five years of education, mentorship, arguments, guidance, growth, and stress accumulate together into a friendship crafted with fire. I am truly lucky to have been able to share this time with you as my lead.

My transition into graduate school was only feasible through the incredible patience and help of Sam Reeve, Saaketh Desai, Michael Sakano, and David Guzman. The foundations for all my skills and work was built directly on your support. Lucky for me, the growth did not stop at my senior mentors. Each day I was fortunate to learn more than I had ever hoped from the team I was surrounded by. Particular thanks go to Brenden Hamilton, Juan Carlos Verduzco, Shivam Tripathi, Lorena Alzate-Vargas, Tongtong Shen, and Robert Appleton. There is a common saying that you are the average of the five people you most surround yourself with. We as a group are fortunate that we slowly build up our aggregate average together.

I have also been fortunate to learn and grow from individuals to whom I was their mentor. Some of my richest times of interpersonal growth, or solidification of concepts was done during time as a teacher rather than a student. To each of the junior members of our group and undergraduates that I had a chance to directly guide in research, or help during coursework as a teaching assistant: thank you for giving me the chance and to earn your trust as a guide.

These drivers of interpersonal growth and academic challenge would not have been possible without the funding sources to support our studies. Through my Ph.D. I was supported by Sandia National Laboratories (SNL) on two separate Laboratory Directed Research Development (LDRD)s, and a separate National Science Foundation (NSF) initiative for Designing Materials to Revolutionize and Engineer our Future (DMREF). While it can be stressful in the moment to shift projects and funding sources mid-Ph.D., the chances they gave to me to grow into a new role were much appreciated.

I would like to thank my parents for their undying love and support for all my life, but especially through these years as an academic. Knowing that no matter how far down the research rabbit hole I went, no matter how hard things became, you two would always be there was one of the best feelings I could have ever asked for. I hope that the time I spent away is paid back in full with a copy of this thesis on your shelf.

To Kenza: my Goober, my partner. Your compassion and warmth has always been a much needed source of support during these challenging years. You truly are one of the most kind, driven, and patient people I have had the pleasure of sharing my time with. I know that everything I am now would not have been possible without you pushing me to be my best during some of our hardest moments.

Lastly, but certainly not least - I feel that it's appropriate to acknowledge the support that I received from some very special animals in my life. To Murphy - the beagle who had so much love to give: I wish you could have made it just one more year so I could show you this dedication, but I suppose I cannot be greedy after having been given almost fourteen. You were such a good boy, and I wish you all the best. To Lucy - I wish you would stop scratching my carpet, but your affection and support for me during graduate school is a debt I will never be able to repay. We do not deserve the animals in our lives, but the years we get with them are some of the most delightful.

PREFACE

Prior to entering my graduate program I was always told that the Ph.D. was a marathon, not a sprint. It takes time, patience, passion, and an undying thirst for answering new questions. Everyone talks about how hard it is, how difficult the content is, and that it is one of the most challenging things you can put yourself through academically. What I feel is lost in these general messages is what actually makes a Ph.D. difficult. It is not the material that is difficult, although the concepts are state of the art. It is not the time that is difficult, even with some years dragging longer than others. It is not the challenge of overcoming writer's block, even though this preface took me much longer than I would have liked. For me, the most challenging part of the Ph.D. was maintaining my mental resolve that everything was going to be okay. Early adulthood is already a strange and confusing time for personal growth, and compounding that with being surrounded by some of the most brilliant and high-achieving people in the world you easily find yourself wondering if you really belong. One day you find yourself in a room and you feel like an imposter amongst the crowd. That sense of not feeling like you belong is the first painful step of many to shaking off a previous version of yourself, embracing the growing pains, and becoming someone you did not think possible. Growing and changing is difficult on the best of days, and during the Ph.D. you find yourself in a war of emotional attrition between the expectations of academia, and your personal ability to mold and grow to an ever changing landscape. What truly makes a Ph.D. difficult is knowing that you lack the current skills for a problem, admitting the initial defeat, but rising to the occasion again and again to keep learning. As someone who has struggled with their mental health through graduate school there were dark periods where I found an infinite supply of things I did not know, but I could not face the uphill battle of improving myself. With the support of my advisor, my family, my friends, and Purdue Counseling and Psychiatric Services I was given tools and support nets to not be afraid to fail. Graduate school involves failure, research involves failure, but what makes us strong is the ability to re-characterize failure as a lesson rather than a punishment. The application and development of mental tools for success is what makes a Ph.D. challenging, and it is what crafts some of the most unique individuals.

Please enjoy the work included - and I hope these words give you what inspiration you need, whether it be emotional or intellectual. As a parting, I would like to leave you with a conversation piece between two hobbits on their way through the undertaking of a lifetime. While it would be crass to compare the efforts of writing a thesis, or the process of innovative materials design to Mr. Frodo's burden, I think that the words offered by Sam have a certain resonance with a wide audience. May they offer you some comfort during your difficulties, and let them be remembered just as we are about to stumble.

FRODO: I can't do this, Sam.

SAM: I know. It's all wrong. By rights we shouldn't even be here. But we are. It's like in the great stories, Mr. Frodo. The ones that really mattered. Full of darkness and danger they were. And sometimes you didn't want to know the end. Because how could the end be happy?

How could the world go back to the way it was when so much bad had happened? But in the end, it's only a passing thing, this shadow. Even darkness must pass. A new day will come. And when the sun shines it will shine out the clearer.

Those were the stories that stayed with you. That meant something. Even if you were too small to understand why. But I think, Mr. Frodo, I do understand. I know now. Folk in those stories had lots of chances of turning back only they didn't. They kept going. Because they were holding on to something.

FRODO: What are we holding on to, Sam?

SAM: That there's some good in this world, Mr. Frodo. And it's worth fighting for[1].

[1] ↑From The Lord of the Rings: The Two Towers. The hobbits' conversation at Osgiliath [1].

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

MGI	Materials Genome Initiative
F.A.I.R.	Findable, Accessible, Interoperable, Reproducible
ML	machine learning
API	application programming interface
DL	deep learning
AL	active learning
NN	neural network
RF	random forest
SHAP	Shapely Additive Explanation
GPR	Gaussian process regression
DFT	density functional theory
GGA	generalized gradient approximation
LDA	local density approximation
XC	exchange-correlation
KE	kinetic energy
PBE	Perdew-Burke-Ernezerhof
SCF	self-consistent field
VASP	Vienna *ab initio* simulation package
SIESTA	Spanish initiative for electronic simulations with thousands of atoms
MD	molecular dynamics
MC	Monte Carlo
MCTS	Monte Carlo tree search
NVT	constant atoms, volume, and temperature
NPT	constant atoms, pressure, and temperature
NVE	constant atoms, volume, and energy
NPH	constant atoms, pressure, and enthalpy
μVT	constant chemical potential, volume, and temperature
LAMMPS	Large-scale atomic/molecular massive parallel simulator

EAM	embedded atom method
MEAM	modified embedded atom method
OVITO	Open Visualization Tool
PTM	polyhedral template matching
TTM	two-temperature model
PCM	phase change memory
RRAM	random access memory
DoF	degree of freedom
FCC	face-centered cubic
BCC	body-centered cubic
HCP	hexagonal close-packed
SC	simple cubic
TDTR	time-domain thermoreflectance
NEGF	non-equilibrium Greens functions
eDMM	electron diffuse mismatch model
XRR	X-ray reflectivity
XRD	X-ray diffraction
XPS	X-ray photoelectron spectroscopy
TEM	transmission electron microscopy
EDS	energy-dispersive spectroscopy
DSC	differential scanning calorimetry
RDF	radial distribution function
ADF	angular distribution function
HEA	high entropy alloy
CCA	complex concentrated alloy
RCCA	refractory complex concentrated alloy
MPCA	multiple principle component alloy
MP	Materials Project
OQMD	Open Quantum Materials Database

TC	Thermo-Calc
VFE	vacancy formation energy
CTE	coefficient of thermal expansion
IPF	ionic packing fraction
QoI	quantity of interest
MAE	mean absolute error
RMSE	root mean squared error
UQ	uncertainty quantification
MM	maximum mean
UCB	upper confidence bound
MLI	maximum likelihood of improvement
MEI	maximum expected improvement
MU	maximum uncertainty
CE	cluster expansion
BS	bispectrum
SOAP	smooth overlap of atomic positions
GAP	Gaussian approximation potentials
SNAP	spectral neighbor analysis potentials
NNP	neural network potential
MTP	moment tensor potential
SVM	supoort vector machines
ACSF	angular centered symmetry functions
wACSF	weighted angular centered symmetry functions

ABSTRACT

In 2011, the Materials Genome Initiative (MGI) was founded as an effort to unite and drive materials design at an unprecedented pace. By linking computational tools with experimental data, and aligning their data structures to match and interact, scientists across the world have been able to change the way they do science at a fundamental level. The 3 Mission Statements of the Materials Genome Initiative include: 1) Developing a Materials Innovation Infrastructure 2) Achieving National Goals with Advanced Materials 3) Equipping the Next-Generation Materials Workforce. Since the inception of the MGI the Materials Engineering community has developed numerous cyberinfrastructure repositories for experimental, and varied levels of computational data. This practice aligns with a separate initiative for Findable, Accessible, Interoperable, and Reproducible (F.A.I.R.) principles for data handling and science. By integrating the cyberinfrastructure efforts with continued collaboration from experimental and computational scientists we push the field to evolve improved workflows for research.

This thesis is a collection of applied solutions for materials design with atomistic modeling, and machine learning (ML). In Part 1, we will discuss bridges for the gaps between atomistic simulation and experiment, and what it means for material solutions. A showcase of combining experimental information with *ab initio* electronic transport calculations will be discussed, as well as the principles of density functional theory (DFT) and molecular dynamics (MD) simulations. In Part 2, our focus will shift to applications of machine learning and the use of composition and chemical featurizers for materials design. Here we leverage cyberinfrastructure efforts with APIs and ML with transfer and active learning for efficient high-dimensional space exploration. In Part 3 local atomic environments and configurations, associative fingerprinting solutions, and workflows for designing deep learning (DL) interatomic potentials for MD are discussed. Finally, a brief section will conclude with efforts made to align with F.A.I.R. principles for Materials Engineering research, and educational development for Mission Statement 3 of the MGI.

1. INTRODUCTION

1.1 Materials of the Ages

The advent of new materials through discovery, design, or accident has earmarked human history and civilization for over three million years. From our earliest ancestors harnessing the loose stone around them to improve their daily lives and secure safety [2], to recent leaps in nanotechnology and bio-atomic level engineering [3]–[5], we have crafted ages around our history to represent what materials innovated people's lives the most at the time. In each of these ages, human civilizations have been defined by the ability to master their materials and explore new worlds.

At the close of the stone age, early civilizations improved their original casting of pure metals and discovered bronze. The fabrication of new and improved tools, jewelry, weaponry, and architecture gave root to the foundations of larger grouped societies. With materials science came the promise of protection, wealth, trade, commerce, and above all sustainability for a people. With the advent of iron smelting a new materials race began as its durability and hardness compared to bronze continued to win over both civilian and soldier [6]. However, as empires rose and fell, a new paradigm emerged with materials. Pushes in the antiquity and porcelain age saw materials development shift from majority commerce and weaponry, to a greater emphasis on the arts and civilian life [7]. Catching up to theory, processing techniques eventually followed that boasted higher processing temperatures, better control of systems, and at the mid-point of the 20th century we shifted our focus from steel to silicon.

As showcased by Fig. 1.1, the defining ages of human civilization have traditionally spanned hundreds, or even thousands of years of development. However, in just the last 100 years the strides we have made as a scientific community dwarf the development of the earlier ages. One could argue that we are on an exponential journey of materials ages and design - with the ultimate goal of materials development being accelerated discovery of new materials and processes. As our materials become more complicated, and we unlock some of the understanding attributed with the smallest levels of atomic resolution, the multi-disciplinary teams that have to work on these problems become high in demand. There is no promise that a project will yield the next age-defining material, and securing commitment from

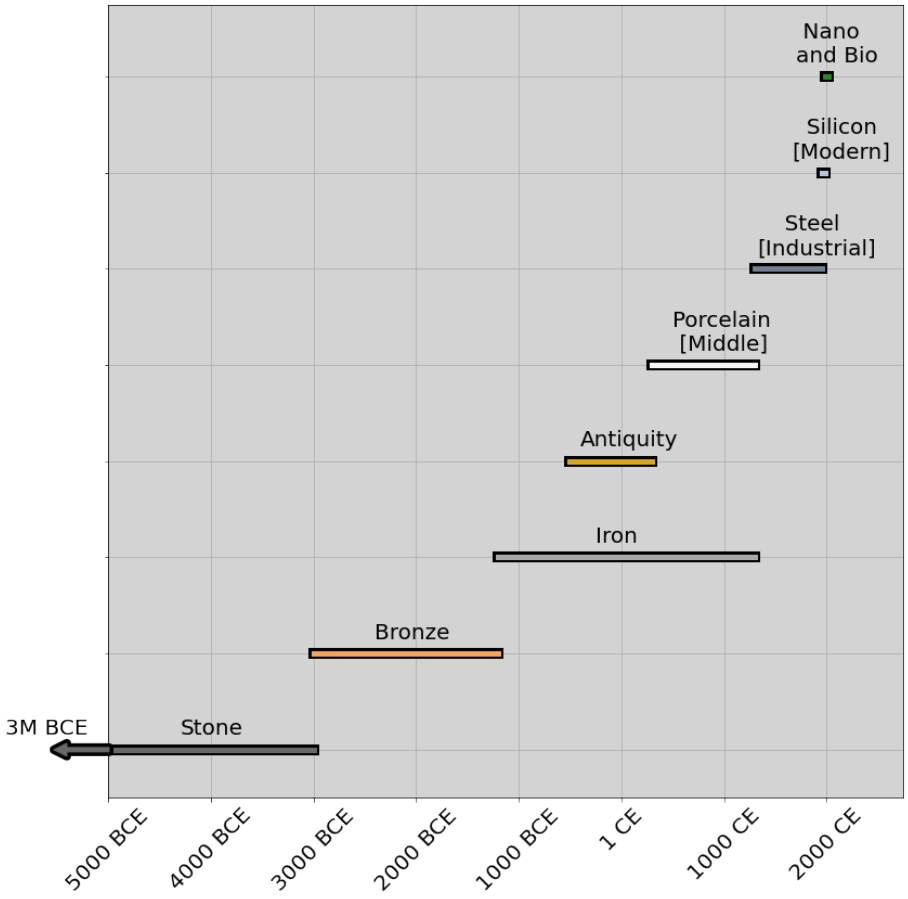

Figure 1.1. An approximate timeline schematic of material's ages up to the present day. Each bar represents a rough era of materials innovation or discovery. While not primary materials, the antiquity and porcelain ages are included to highlight a re-valuing of materials for artistry and national wealth.

varied institutions is challenging. To overcome this dual-phase problem of interpersonal and scientific complexity the way we do science as a community has begun to shift.

1.2 Building a knowledge database

Within a domain we can conceptualize the acquisition of knowledge as individuals slowly pushing outward on a known space. An advanced degree or committed research effort in particular is an individual's chance to discover something new their peers have yet to accomplish. This breach in the current understanding classically builds on knowledge extracted from within the domain, or directly from colleagues, but as materials and civilization have advanced our dissemination of information has changed. Barriers between groups have begun to slowly erode, and the transfer of knowledge across domains has received new breath. Fig. 1.2 illustrates a cartoon schematic for personal and domain growth of knowledge. From a motivational presentation from Mike Might [8], a schematic was created to show what an individual goes through to push a single point on the boundary of humanity's knowledge. Beginning with elementary and initial advanced coursework a person grows into a specialized role within a field. Eventually their contributions push this boundary, and a new radii of knowledge is created.

What this image does not capture, is that as we breach our own boundaries, we have the opportunity to make our work accessible to others across the field in another discipline. With strides in digital infrastructure, communication, high-throughput experiments and computation over the last 100 years, fields previously though separate have begun to collaborate and cross-pollinate each other through both necessity and personal interest. Fig. 1.2 also highlights that as domains become more interconnected, the chance for primary and secondary acquisition of knowledge grows with each branching interdisciplinary effort. Simple exchange programs between colleagues or professors grows culture within organizations, and allowing our work to be accessible to wider ranges of audiences helps to lower the cost barrier associated with searching for new ideas. Materials engineering is inherently a multi-disciplinary field, and the accessibility to data and information within the field is essential for accelerated design of materials.

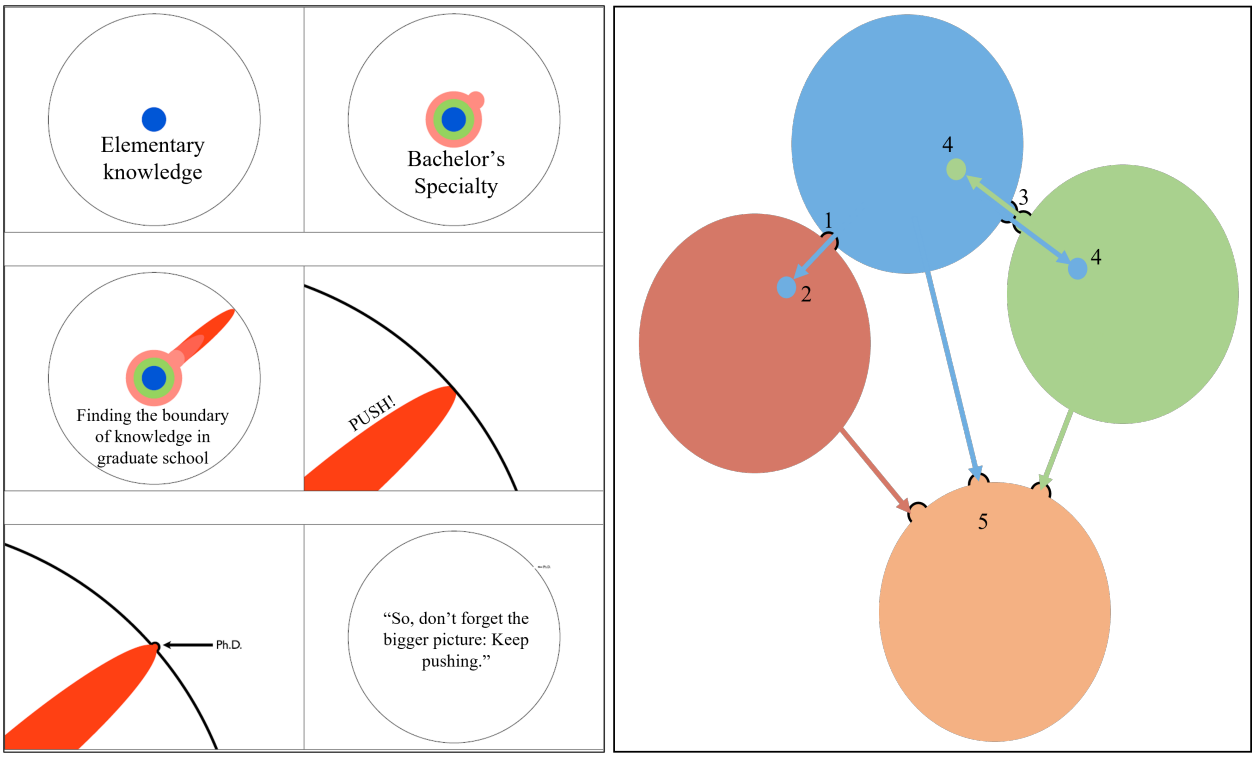

Figure 1.2. [Left] A pictorial representation of the acquisition of an individual's knowledge through education. Highlighted here is the individual pushing the boundary of the known space. Reproduced from Matt Might [8]. [Right] A schematic of separate domains of knowledge growing due to cross-pollination. 1) **Domain discovery**: An advanced academic finding something from a separate field that helps their work. 2) **Domain shifting**: An individual who joins a new field of study and brings fresh perspective. 3) **Domain collaboration**: Leaders in the field collaborating and exchanging ideas. 4) **Interdisciplinary Domains**: When fields begin to overlap to an extent that individuals find themselves drifting towards new skills. 5) **Domain Sampling**: Using modern tools to learn about other domains without direct contact.

1.3 The Materials Genome Initiative

Beginning in 2011, the Materials Genome Initiative (MGI) was founded as a guiding roadmap to accelerate materials design and innovation. Its mission: to codify and unify researchers from across domains to develop new products with unprecedented speed [9]. Its three guiding principles include: 1) Developing a materials innovation infrastructure, 2) Achieving national goals with advanced materials, and 3) Equipping the next-generation materials workforce. With this initiative came not only renewed discussions on domain Best Practices, but it also provided critical infrastructure funding for repositories, centralized databases, and workplace development. Highlighted in Fig. 1.3, the MGI founding concept recognized that all domains that relate to materials have common intersections between specialized practices. Computation would not exist without the validation of experiments, and neither could communicate if it were not for digital infrastructure.

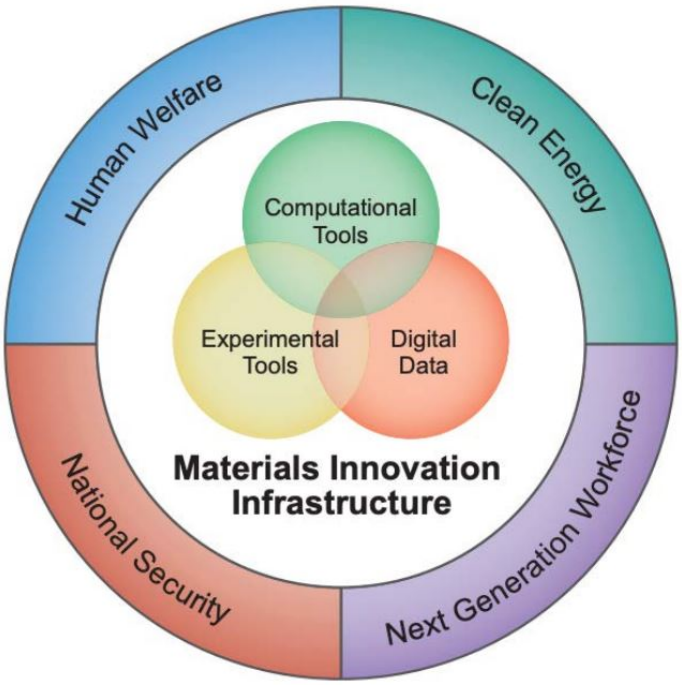

Figure 1.3. The founding conceptual structure of the Materials Innovation Infrastructure. Reproduced from MGI Strategic Plan 2011 [9].

Changing the flow of work for any environment is difficult, let alone research efforts across industry, national laboratories, and academia. However, to keep pace with the growing complexity of our lives the products we build must be developed with the same complex intricacy. This is not limited to accelerating the timeline from theory to industrial production, but also understanding each step of processing, the life cycle analysis of products, potential future contaminants, and replacement of already known toxic materials. As showcased by the explosion of the polymer industry in the 1950s, large steps in material production can have long standing effects if not carefully procured.

As we have begun to highlight, the process for materials development, analysis, engineering, recycling, and destruction quickly becomes a highly dimensional and complex space to explore. Not only do we have to consider composition of our materials, or the intrinsic recipe, we have to understand processing, material sourcing, defect analysis, long and short term fatigue, and more. To make matters more complicated, all of these properties can change slightly or catastrophically with a shift in composition, temperature, or atmospheric condition. The options presented can quickly lead to paralysis by choice. To overcome these challenges the scientific community has developed tools and workflows to expedite our initial screening process, and as our technology allows for more complex executions of materials design our field has flourished.

Ten years later, the MGI remains, and the landscape for materials design has been softly guided towards continued acceleration for design. An update schematic of their mission is shown in Fig. 1.4, with a higher emphasis on the cross-exchange of information through multi-disciplinary guidance. This drive in materials innovation is part of what has allowed our material "ages" to continue accelerating at their current pace. This is largely in part to the multi-disciplinary nature of the field, and with new integrations for Findable, Accessible, Interoperable, and Reproducible (F.A.I.R) data [11] principles. These efforts create a more transparent data availability system, and also create a synchronous data culture. Particularly with the presence of the MGI, a number of institutions around the world have put their work forward as figureheads of database development. Primary examples that were used in this work include: catering to niche communities for education platforming and simulation with nanoHUB [12], [13], electronic structure and crystallographic information with queryable

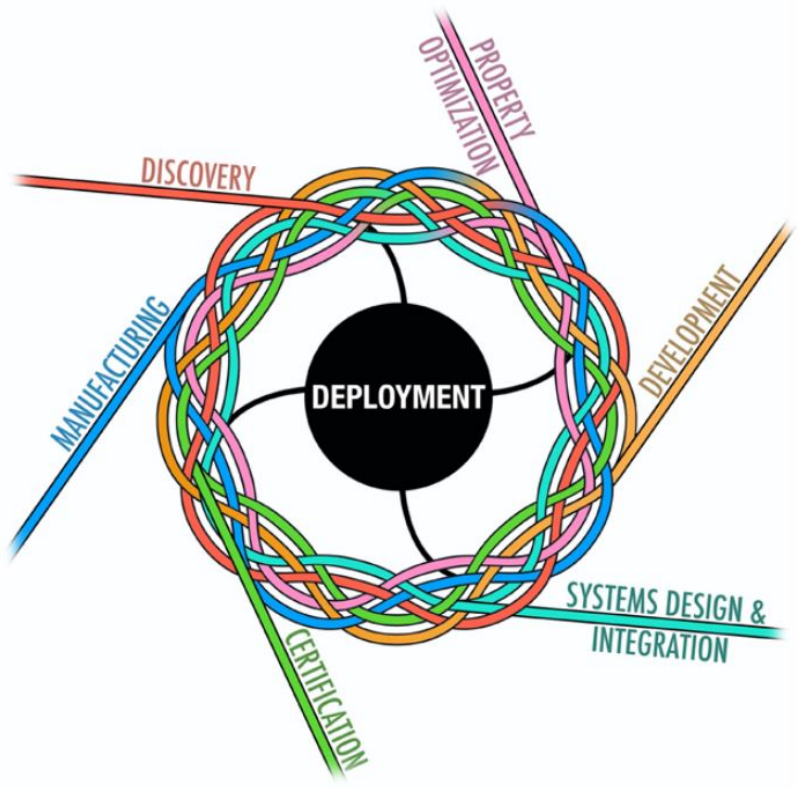

Figure 1.4. "The MGI paradigm promotes integration and iteration of knowledge across the entire materials development continuum. Unification of the MII will provide a framework for seamless, convective flow of information and a weaving of knowledge among all stakeholders contributing to the materials R&D enterprise, accelerating the deployment of new materials." Reproduced from MGI Strategic Plan 2021 [10]

APIs on Materials Project [14], [15], hybrids of computational and experimental results [16], and accessibility to modern simulation tools [17]. These tools and more have continued to reduce barriers, and allow for creative selection of methods to solve unique problems in the field.

1.4 Tool assessment and selection

Given a material problem, it is our goal to use well defined and accessible tools to solve for a target, or help explain gaps in the current model or experimental understanding. However, depending on the corresponding length and time scale of interest the fidelity of our tools will vary. In this section we will briefly explore the length and time scales of materials engineering from atomistic to macroscale, and briefly touch on how information from each level is propagated to a new tier. Fig. 1.5, shows a schematic representation of the cost associated with moving towards higher time and length scales. At the far left of the figure is our highest first-principles determined physics energy states. The smallest atomic particle, the electron, is modeled with degenerative effects, excitation energetics, and solved explicitly to determine chemical bonding between groups of atoms. As the atoms and electrons scale, the cost of performing each calculation scales. For each step we take towards larger and longer scale, to accommodate current state of the art computing methods certain assumptions must be calibrated, and the overall first-principles determined physics approximations be trimmed. For reasons we will discuss in detail in Ch. 2, calculating every electron for a macroscale system is computationally prohibitive, and would take longer than the age of the universe to attempt.

1.4.1 Femto- to nano-

At the highest level of first-principles determine physics we model atomic systems (10^{-12}-10^{-8} m) with explicit electron interactions represented through wavefunctions. Here in this domain, we sacrifice the ability to model fracture in a steel beam, but we can determine the stiffness of the alloy with near experimental precision, or determine electronic density of states. We can characterize fundamental interactions of electron bonding and anti-bonding,

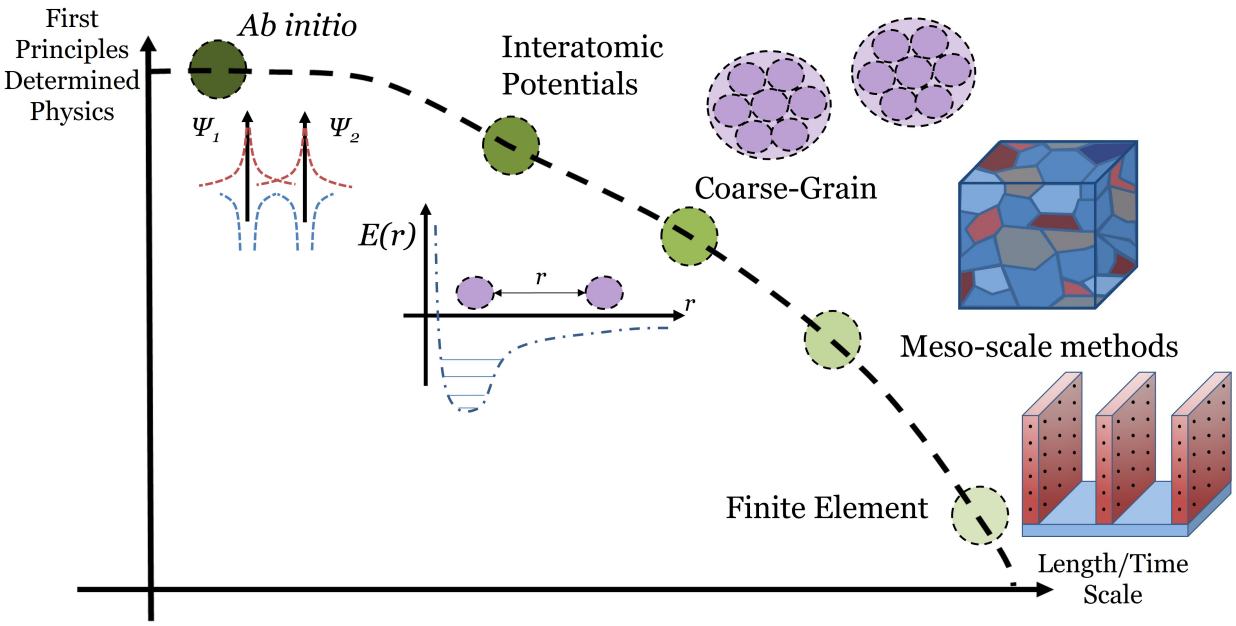

Figure 1.5. Pareto front of materials simulations tools with respect to length and time scale.

atomic relaxations due to bonding and electronic configurations, and preliminary interactions of long range order effects from dynamics. Electron interactions and coupling are modeled using first-principles physics based on wavefunction theory, and eventually density functional theory (DFT). While all are approximations of the time dependent Schrödinger equation, each iteration boasts different levels of computational efficiency and accuracy. However, at each point along the Pareto front we can use information from varied scales both experimentally and computationally to inform our model or decision-making process.

Unfortunately, while accurate and powerful in their own right, full electronic descriptions of each atomic system require immense computational power to evaluate. As we move to larger length scales to characterize higher dimensional defects, long-range order, and dynamics at longer scales so too must our simulation sizes scale. Here we replace all explicit electronic effects in the system, and model our ions as hard spheres with a potential energy function between neighbors that governs bonding and repulsion. Dynamics at this range can be modeled using molecular dynamic simulations where the forces are driven by this interatomic potential function. Removing the need for individual electronic interactions allows us to perform higher order molecular dynamics (MD) simulations, but with understood lower fidelity than ab initio MD. Furthermore, interatomic landscapes can be informed from experimental data as well.

1.4.2 Micro- to macro-

If interactions at the particle level are still important, but due to the size requirements of a full-body simulation explicit modeling of particles cannot be considered due to computational cost. Similar to the grouping of electronic interactions into a hard sphere, groups of atoms are bundled together into a coarse particle modeled as a larger sphere. Rather than assessing each atomic interaction we instead look at groups of atoms and their dynamics.

Scaling up to the microstructure (10^{-7}-10^{-3} m) we leverage free energy derivatives to determine phase equilibria or physics such as granular or dendritic growth. Free energy expressions of phases are approximated to obtain numerical expressions that can be differentiated for phase equilibrium. Leaning on the success of first principles and molecular dynamic

modeling, many of the inputs used in mesoscale modeling come directly from thermodynamic state variables in a system such as vacancy fraction and formation energy, cohesive energy, and relative phase stability.

With varied computational assumptions we can model material phenomena over a range of length/time scales. For large devices (10^{-2} - 10^{5} m) we solve differential equations at nodal points through finite element models. These problems are generally solved through partial differential equations that govern heat, mass, and fluid transport for a system. Nodal points on a device are solved, and coupled to neighbors to create dynamical simulations of processes. These simulations can range from extremely high fidelity, with high resolution of nodal solutions, explicitly derived properties of the material properties such as thermal conductivity from first-principles physics, or the solution can be as simple as a 1-D diffusive heat equation.

1.5 Connecting computational tools

The following work included in this dissertation is a collection of materials problems and solutions within the reference of computational research. In each of the chapters we will revisit Fig. 1.5 as a guiding reference to our process. To help initiate the unaffiliated, Ch. 2 will offer explanations of *ab initio* modeling with the foundations of density functional theory included. A discussion of the mechanisms of molecular dynamics will be provided with descriptions of timestep integration, interatomic potential functional, and statistical mechanics of systems for thermodynamic variable extraction. Finally, we will discuss a non-materials engineering tool that has helped drive the field rapidly with its widespread adoption in recent years: machine learning (ML). While an entire thesis could be written on the merits and disadvantages for each subset of ML, we will focus on the use of neural networks and random forests for regression based tasks.

Following the discussion of methods we will apply our toolbox of computational assets to explore a variety of materials systems and processes. To evaluate thermal transport mechanisms in reactive metallic materials in Ch. 3, an exchange of information between experimental, interatomic potential driven molecular dynamics, and electronic structure

theory is shown for transport kinetics of ions and electrons in Ni and Al/Pt systems. This chapter will be used as an introduction and calibration to the remainder of the dissertation for leveraging different aspects of our tools to propagate information. Ch. 4 will discuss the use of ML tools for exploring high dimensional chemical composition spaces, with low volume datasets or initial information. Often referred to as the 'Inverse Problem' in ML, we highlight different tools and careful selection of physics-based features to help guide ML models through this extrapolation exercise. Our ultimate goal: to define and create models that can seek out high-performing material candidates from previously unmanufactured or theorized predictions. While successful, these models are often static, can only predict values for a single chemical composition, and ignore the local chemistry of atomic environments. Ch. 5 rectifies this gap with deep learning methods for atomic level property prediction, and energy states based on invariant chemical geometries. The use of deep learning as a tool for materials engineering shows great promise given the high dimensionality problem of local chemistry, and given voluminous initial datasets we have shown excellent capability of our models to explore complex systems. In spirit with the 3rd guiding principle of the MGI, we will highlight education outreach and research accessibility efforts to train the next generation workforce in Ch. 6. Finally, we will conclude in Ch. 7 with our outlook on the field as it is today, what we hope it looks like in the future, and to begin waiting for MGI 2031.

2. ATOMISTIC SIMULATION AND MACHINE LEARNING METHODS

2.1 Introduction

The primary scale and focus of this thesis was amongst atomic level information and properties. The work was guided by underlying first-principles physics informed decisions with complementary tools to facilitate expedited materials development. At the smallest scale we will consider here is the electronic structure information of individual elements, their interactions when paired with neighbors of varied element types, and how this affects localized structure and property relationships. What we will begin to lay out is a foundation for atomic mapping of properties, and when it becomes necessary to shift pieces of information from one set of simulations to another. Many of the material properties that we learn from the electronic structure can be modeled implicitly without the need for costly wavefunction solutions. These interatomic potentials create the boundary of interaction between the atoms rather than defining them by their explicit electron bonding environment. Eventually, we will build a basis for leveraging machine learning techniques to bridge some of the fuzzier gaps between electronic structure theory, and the dynamic interaction of large scale particle systems with well defined training and feature sets. The methods included here are not a comprehensive summary of methods within the field, but a deeper look into the tools used in this thesis.

2.2 Electronic Structure Theory

At the turn of the 20th century, the physics community was entering a golden age of academic pursuit. From Planck, to Einstein, to Bohr, the world's brightest physicists were attempting to unravel the mysteries of atomic level interactions, and the effect of electronic effects in systems. In 1900, Max Planck discovered something that would set off a theoretical arms race for the next century [18]: the explanation for the breakdown of classical mechanics when considering the radiation effects of photoelectric sources. His theories set forth in motion the idea that energy could be quantized as individual packets, rather than

being set continuous functions. Building on observations of discretized energy states in the Franck-Hertz experiment [19], the theories that would be modified in the early 1900s would allow for the development of electronic structure calculations as we know it.

We will begin with the initial approximation for a group of atoms and the description we define for their localized properties. Given a wavefunction, ψ, the time-evolution of the function's property with respect to an atom's energy state is solved using the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi, \tag{2.1}$$

with $\hbar$ the Planck's constant, t the time, and H the Hamiltonian operator, or the contributive sums of potential and kinetic energy. Theoretically, this equation is the fundamental law that governs all electronic and ionic dynamics. However, the analytical solution for this equation has eluded researchers even for simple molecular systems. At best, modern state of the art quantum computing infrastructure has successfully solved a partially analytical solution for the time-dependent Schrödinger equation, but scaling limitations in hardware and software continue to create burdensome challenges [20].

A critical approximation to Eq. 2.1 was proposed by Born-Oppenheimer [21], where they postulated that the relative speed of electrons to ions is so large that ions can be perceived by the electrons as stationary objects. This removes a critical time dependence of the wavefunction, and assumes that what an electron particle 'sees' is a frozen-atom. The removal of the time-dependence allows us to re-write Eq. 2.1, in the following form:

$$\mathcal{H}\psi = E\psi. \tag{2.2}$$

Here, E is a representation of the total energy of the system derived from the Hamiltonian. A wavefunction solution exists where the Hamiltonian and energy state are equal. The total Hamiltonian of a system here is described the combination of all particles with the contributions of kinetic energy from ions and electrons, as well as electron-electron, electron-ion, and ion-ion interactions:

$$\mathcal{H}_{tot} = \sum_{\mathrm{i}} \frac{P_{\mathrm{i}}^2}{2m_{\mathrm{i}}} + \sum_{\mathrm{j}} \frac{P_{\mathrm{j}}^2}{2M_{\mathrm{j}}} + \sum_{\mathrm{i}<\mathrm{i}'} \frac{\mathrm{e}^2}{|r_{\mathrm{i}} - r_{\mathrm{i}'}|} + \sum_{\mathrm{i,j}} \frac{-Z_{\mathrm{j}}\mathrm{e}^2}{|R_{\mathrm{j}} - r_{\mathrm{i}}|} + \sum_{\mathrm{j}<\mathrm{j}'} \frac{Z_{\mathrm{j}}Z_{\mathrm{j}'}\mathrm{e}^2}{|R_{\mathrm{j}} - R_{\mathrm{j}'}|}. \tag{2.3}$$

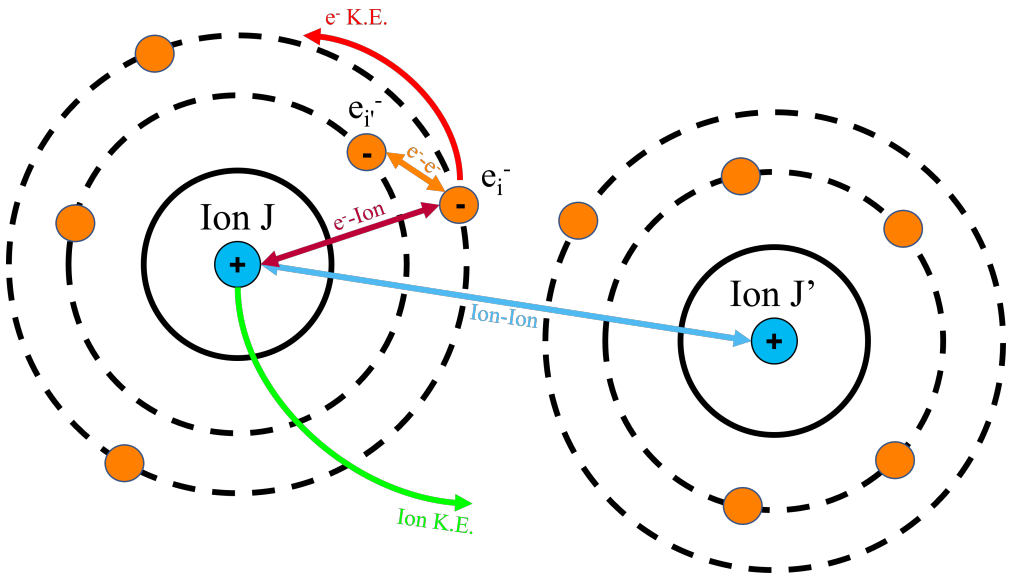

Figure 2.1. The Hamiltonian for *ab initio* electronic structure calculations

Uppercase variables such as R and M will denote ionic position and mass respectively, while lowercase r and m be representative of the electrons. Z_{j} is the ionic charge of the selected ion, and e^- the electronic charge fundamental unit. To help explain each term of the Hamiltonian, the colors of each expression in Eq. 2.3 have been coordinated to a respective pairwise or kinetic term schematic in Fig. 2.1. Assessing the interactions between two ions, J and J$'$, their respective electrons, $\mathrm{e}_{\mathrm{i}}^-$ and $\mathrm{e}_{\mathrm{i}'}^-$, we break down the contributions from each kinetic and potential energy effect in the system. Unfortunately, the expression shown in Eq. 2.3 is currently unsolvable with the explicit measurement and simulation of every particle within the atomic subsystem, given that for each electron we want to simulate our process scales by 3N-dimensionality, where N is the number of electrons in the system. While modern quantum supercomputers have begun to push the boundary of what we previously thought solvable, the full analytical solution of the *ab initio* Hamiltonion is unknown to this day.

A fundamental stride made in the modern calculation of electronic structure is the postulate that for a system at 0 K, the properties of the electronic system are at their ground

state. Excitation properties such as optical, electrical, or thermal effects are only prevalent for non-zero temperatures, and we can group the electron states as a gaseous density rather than individual particles with inherent uncertainties. This reduces our high dimensional many-body-electron problem to a single particle operating in Cartesian space (x,y,z).

2.2.1 Density Functional Theory

Beginning with the initial work by Hohenberg and Kohn [22], they showed that the for an interacting electron gas the ground state energy, E_{gs}, is a functional of the electron charge density $\rho(\mathbf{r})$, a perturbation from an external static potential $v(\mathbf{r})$, and a universal functional for the density $F[\rho]$.

$$E_{gs}[\rho] = \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} + F[\rho] \tag{2.4}$$

To find a minimum solution for Eq. 2.4, the charge density must satisfy the folowing:

$$N = \int \rho(\mathbf{r})d\mathbf{r} \tag{2.5}$$

with N as the total number of electrons.

This reduction in problem dimensionality was the basis of the Kohn-Sham set of equations for solvable functionals of $rho(\mathbf{r})$, $E_{gs}[\rho]$, and $F[\rho]$. Unpacking the universal functional of the density, $F[\rho]$, we evaluate the electron-electron interactions in the gas, and the kinetic energy. While solvable in principle, in practice these terms become quite complex to solve. Therefore, early steps to solve these expressions assumed that the primary exchange and correlation mechanisms would be ignored, with electron interactions taking the form of the Hartree equation:

$$E_{Hartree} = \frac{e^2}{2}\int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}drd\mathbf{r}' \ . \tag{2.6}$$

For a non-interacting system with charge density $\rho(\mathbf{r})$, the one-particle orbital, $\phi(\mathbf{r})$, kinetic energy is shown by:

$$K_s[\rho] = -\frac{\hbar^2}{2m}\sum_{\mathrm{i}}^{occ}\phi_{\mathrm{i}}^*(\mathbf{r})\nabla^2\phi_{\mathrm{i}}(\mathbf{r}) \tag{2.7}$$

and finally, the charge density is shown below:

$$\rho(\mathbf{r}) = \sum_{\mathrm{i}}\phi_{\mathrm{i}}^*(\mathbf{r})\phi(\mathbf{r}). \tag{2.8}$$

Expanding the approximate solutions for Eq. 2.4 we derive the Kohn-Sham formalism for an interacting electron gas under the influence of an external potential. The first term in blue is the external potential induced on the electron density, the second term in red the non-interacting kinetic energy of electrons in the system, third in green is the interacting potential energy of electrons from the reduced Hartree equation, and finally the exchange-correlation functional $E_{XC}[\rho]$ which contains the remaining electron-electron interactions not included within the Hartree energy, and any exchanged kinetic energy absent from K_s.

$$E_{gs}[\rho] = \int v(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} - \frac{\hbar^2}{2m}\sum_{\mathrm{i}}^{occ}\phi_{\mathrm{i}}^*(\mathbf{r})\nabla^2\phi_{\mathrm{i}}(\mathbf{r}) + \frac{\mathrm{e}^2}{2}\int\frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}drd\mathbf{r}' + E_{XC}[\rho] \tag{2.9}$$

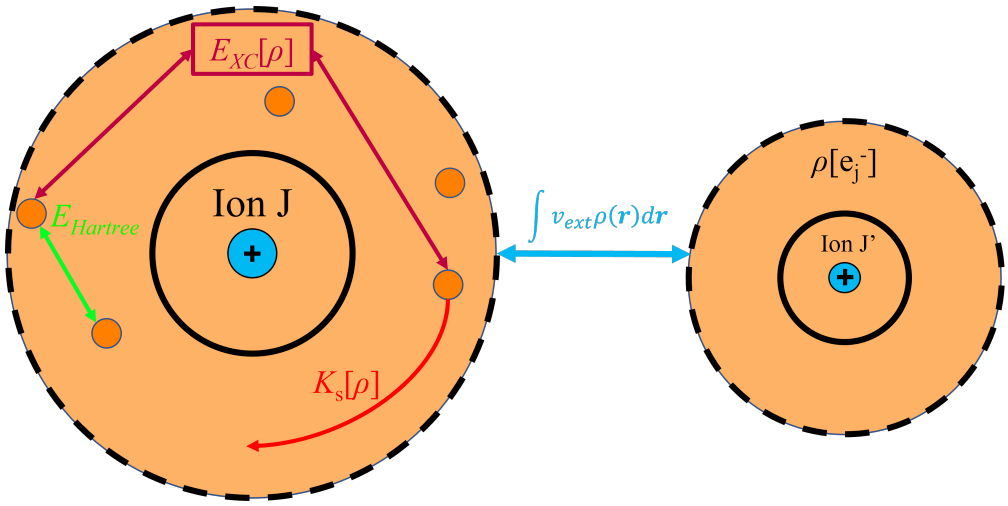

Figure 2.2. Representation of terms for the Kohn-Sham formalism of DFT. Included are the interactions of atoms from the Hartree energy, non-interacting kinetic energies of electrons, the external potential field from an ion, and the exchange-correlation functional. Values are color coded to terms in Eq. 2.9.

Using the set of functionals developed for Eq. 2.9, a set of Euler-Lagrange equations are implemented for the solution of the eigenvalues and eigenfunctions that describe a one-particle electronic system. The self-consistent field (SCF) equation is shown below in Eq. 2.10, and given E_{XC} can be solved explicitly and exactly.

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + \mathrm{e}^2\int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{XC}(\mathbf{r})\right]\phi_\mathrm{i}(\mathbf{r}) = \epsilon_\mathrm{i}\phi_\mathrm{i}$$
$$\text{where } V_{XC} = \frac{\partial E_{XC}}{\partial\rho(\mathbf{r})} \tag{2.10}$$

Unfortunately, like the full solution to the Schrödinger equation, the exact exchange-correlation functional is still unknown. However, the approximations that have been made for this value have shown great success at modeling macroscopic properties and states [23]. The two primary methods for modeling the exchange-correlation are within the local density approximation (LDA) [24], where the electron density is a function of the local density, $\rho(\mathbf{r})$, and the generalized gradient approximation (GGA) with the functional a form of both the local density, $\rho(\mathbf{r})$, and gradient, $\nabla\rho(\mathbf{r})$. For electronic states, transport, and *ab initio* dynamics simulations the Vienna ab initio Simulation Package (VASP) [25] and Spanish Initiative for Electronic Simulations with Thousands of Atoms (SIESTA) [26] software suites were used.

2.3 Molecular Dynamics

Using the electronic structure information derived in Sec. 2.2.1, we can use molecular dynamics (MD) methods to capture the time and space evolution of particles using Newton's classical equations for mechanics. According the Newton's second law, atomic acceleration, a, will occur proportionately to the force applied, F, and the atomic mass, m.

$$F = m \cdot a \tag{2.11}$$

Through the gradient of the potential energy function, the force on an atom can be evaluated through an interacting surface:

$$F = -\nabla U. \tag{2.12}$$

Building on the relationship between Eq. 2.11 and Eq. 2.12 we can propagate the forces experienced on an atom to find the new preferred position and energy given their configuration. For atomic positions, $\mathbf{r}$, and velocities, $\mathbf{v}$, defined for each atom the forces can be numerically integrated and solved for the next step in time [27]. Here, $\dot{r}_\mathrm{i}$ is the derivative of the position with respect to time, or $\frac{\mathrm{d}x}{\mathrm{d}t}$, and $\dot{v}_\mathrm{i}$ the derivative of velocity, or the acceleration. Using the relationship between the position and velocity updates, and given an interatomic energy potential, $V(r_\mathrm{i})$[1], a full time evolution of a molecular system can be performed. Fig. 2.3 shows an example workflow for the initialization and update steps for an MD simulation. Depending on the assumptions made within the $V(r)$ expression, whether they be from a full electronic structure calculation using DFT, or a parameterized functional form, particle dynamic systems of hundreds to millions of atoms can be parallelized.

$$\dot{r}_\mathrm{i} = v_\mathrm{i} \tag{2.13a}$$

$$\dot{v}_\mathrm{i} = \frac{F_\mathrm{i}}{m_\mathrm{i}} \tag{2.13b}$$

$$F_\mathrm{i} = -\frac{\partial}{\partial r_\mathrm{i}} V(r_\mathrm{i}) \tag{2.13c}$$

2.3.1 A Brief Overview of Statistical Mechanics

Now that we have the tools to simulate many-body interactions of atoms and build larger sized systems, we can begin grouping the individual information obtained on a per-atom basis for interpretable macroscopic information. By sampling stochastic averages of group mechanics we unlock bridge between localized atomic vibrations and macroscopic properties such as temperature, pressure, and potential energy [28].

[1]↑It is common (and preferred) in MD literature to write the interatomic potential functional as $\phi(r_\mathrm{i})$. However, to avoid confusion for the reader coming out of a DFT section where ϕ is the single particle orbital, the use of an equally acceptable form $V(r_\mathrm{i})$ is used. So, we have $V(r_\mathrm{i})$ the potential, V for volume, v as velocity, and sometimes wavenumber with ν. Sorry.

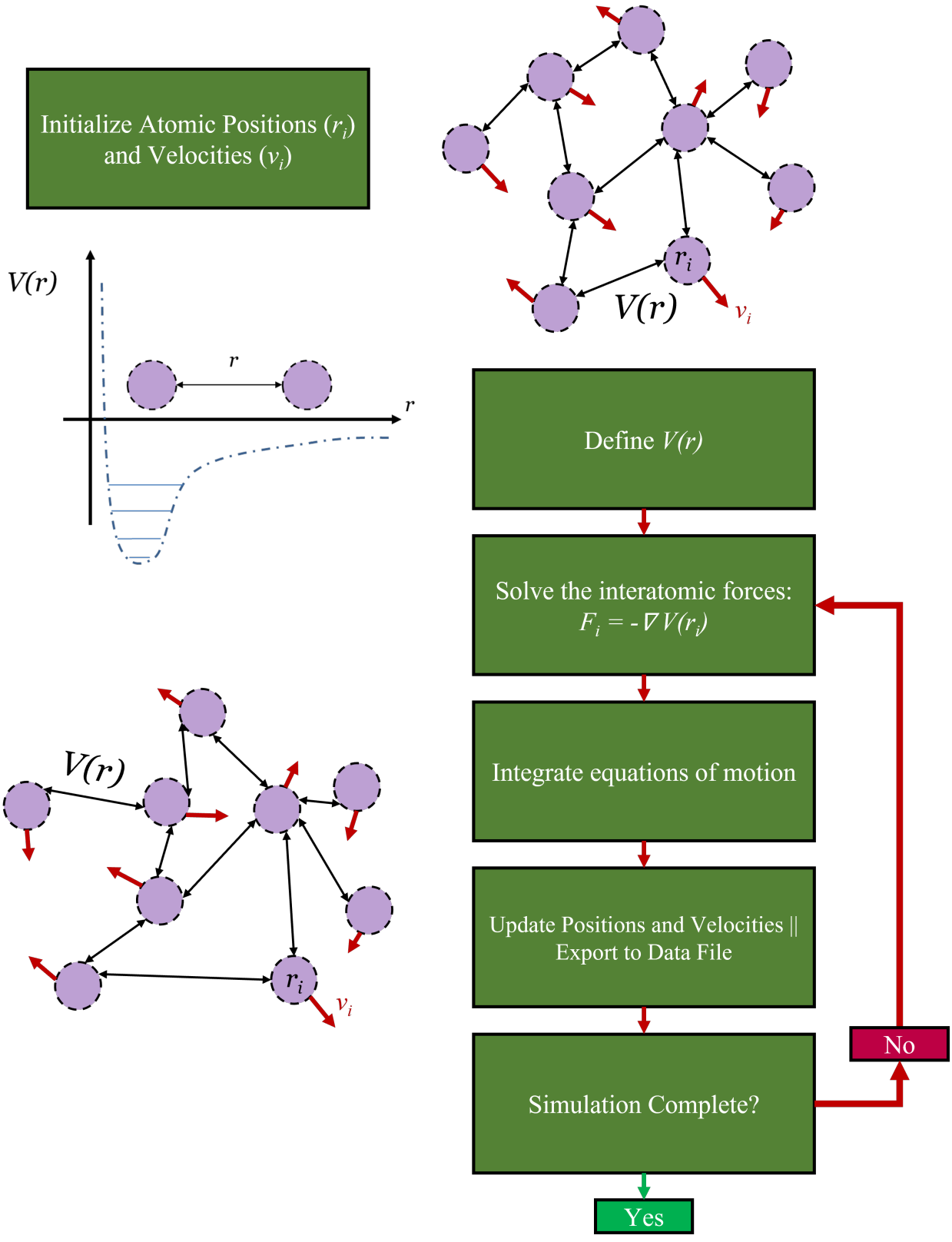

Figure 2.3. Molecular dynamics protocol.

Similarly to our formalism for density functional theory, we will begin with the Hamiltonian for motion of a system of particles. We define this group of positions and momenta as the microstate of the system.

$$\mathcal{H}(\mathbf{r}, \mathbf{p}) = K(\mathbf{p}) + V(\mathbf{r}) \tag{2.14}$$

Here, $\mathbf{r}$ and $\mathbf{p}$ represent the position and momentum of N particles in a system, and are dependencies for the kinetic, $K(\mathbf{p})$, and potential energy, $V(\mathbf{r})$. The kinetic energy is broken down into Cartesian components of momentum in Eq. 2.15, while the potential energy is some assumed function determining pair-wise dynamics.

$$K(\mathbf{p}) = \sum_{\mathrm{i}=1}^{N} \frac{1}{2m_{\mathrm{i}}}(p_{\mathrm{i}x}^2 + p_{\mathrm{i}y}^2 + p_{\mathrm{i}z}^2). \tag{2.15}$$

Once an assumption is created for V(r), whether it be electronic structure, interatomic potential, or ising model, the time-evolution Hamiltonian is generated to acquire microstates. The phase-space product of the momentum and position result in a 6N dimensional space for sampling. Ensembles are associated with the represented phase-space for groups of subsystems, that when summed, generate macroscopic properties. In this section we will establish the fundamental principles of statistical mechanics for atomic motion, and how these ensembles are used to extract number of atoms (N), volume (V), energy (E), temperature (T), and pressure (P). For a subsystem, a point can be mapped in phase-space, and with a group of points show macroscopic properties. Generally, an observable value A is calculated from the ensemble average as:

$$\begin{gathered} A_{observable} = \langle A \rangle_{ensemble} \\ \text{where } \langle ... \rangle \text{ represents an average.} \end{gathered} \tag{2.16}$$

As an exercise, we will consider the Canonical Ensemble Average of an atomic system, where by definition the number of atoms, the volume, and temperature will be held constant. We can envision our system as being equilibrated by a surrounding heat bath that exchanges energy to keep our system temperature constant as shown in Fig. 2.4. Since we have an

exchange of energy, the number of states at different energy levels will vary. We define the partition function of the canonical ensemble as:

$$Z_{NVT} = \sum_{\substack{micro- \\ states}} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})} \tag{2.17}$$

where $\beta = 1/k_B T$, with k_B the Boltzmann constant. The Boltzmann factor, $\mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})}$, is used in the Maxwell-Boltzmann distribution to define the probability distribution function of our ensemble:

$$P(\mathbf{r},\mathbf{p}) = \frac{\mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})}}{\sum_{micro-states} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})}} \tag{2.18}$$

To simplify our expression we will decouple the kinetic and potential energy terms of our Hamiltonian, and solve the kinetic term analytically. The potential energy term $V(r)$ will remain in the partition function. For an observable property, the average for the NVT ensemble is shown in Eq. 2.19. The property of interest, A, will determine the necessary sum and expansion of the expression for a value. Included below in Table 2.1 are the probability distribution functions, partition functions, and ensemble expressed free energy.

$$\langle A \rangle_{NVT} = \sum_{\mathbf{r}} A(\mathbf{r}) P(\mathbf{r}) = \frac{\sum_{\mathbf{r}} A(\mathbf{r}) \mathrm{e}^{-\beta V(\mathbf{r})}}{\sum_{\mathbf{r}} \mathrm{e}^{-\beta V(\mathbf{r})}} \tag{2.19}$$

Table 2.1. Statistical mechanics ensembles and their derived functions for thermodynamic quantity analysis. **NVE**: Constant number of atoms, volume, and energy. **NVT**: Constant number of atoms, volume, temperature. **NPT**: Constant number of atoms, volume, temperature. μ**VT**: Constant chemical potential, volume, temperature

	Probability Distribution Function	Partition Function	Free Energy
Microcanonical (NVE)	$P(\mathbf{r},\mathbf{p}) = \frac{1}{\Omega(N,V,E)}$	$\Omega(NVE) = \sum_{m.s.} \delta(E - \mathcal{H}(\mathbf{r},\mathbf{p}))$	$S = k_B \, log \, \Omega(N,V,E)$
Canonical (NVT)	$P(\mathbf{r},\mathbf{p}) = \frac{\mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})}}{Z_{NVT}}$	$Z_{NVT} = \sum_{m.s.} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p})}$	$F_{NVT} = -k_B T \, log \, Z_{NVT}$
Isobaric/Isothermal (NPT)	$P(\mathbf{r},\mathbf{p},V) = \frac{\mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p}) - PV}}{Z_{NPT}}$	$Z_{NPT} = \sum_V \sum_{m.s.} \mathrm{e}^{-\beta \mathcal{H}(\mathbf{r},\mathbf{p}) - PV}$	$G_{NPT} = -k_B \, log \, Z_{NPT}$
Grand-Canonical (μVT)	$P(\mathbf{r},\mathbf{p},\mu) = \frac{\mathrm{e}^{-\beta(\mathcal{H}_i(\mathbf{r},\mathbf{p}) - \mu N)}}{\Xi_{\mu VT}}$	$\Xi_{\mu VT} = \sum_{N=0}^{\infty} \sum_{m.s.-\mathrm{i}} \mathrm{e}^{-\beta(\mathcal{H}(\mathbf{r},\mathbf{p}-\mu N)} = \sum_{N=0}^{\infty} \mathrm{e}^{\beta \mu N} Z_{NVT}$	$\langle N_{\mu VT} \rangle = k_B T \frac{\partial \, log(\Xi)}{\partial \mu}$

As an example of expanding on the NVT ensemble, we show a few expressions of the macroscopic properties listed above, but with their statistical mechanic formulation:

$$U = \frac{1}{Timestep}\sum_{\mathrm{j}}^{Timestep}\frac{1}{N}\sum_{\mathrm{i}}^{N}V(r_{\mathrm{i}}) \;||\; \text{Internal Energy} \tag{2.20a}$$

$$T = \frac{1}{Timestep}\sum_{\mathrm{j}}^{Timestep}\frac{1}{N}\sum_{\mathrm{i}}^{N}\frac{2}{3k_B}\frac{1}{2}m_{\mathrm{i}}v_{\mathrm{i}}^2 \;||\; \text{Temperature} \tag{2.20b}$$

$$P = \frac{1}{Timestep}\sum_{\mathrm{j}}^{Timestep}\left(\frac{Nk_BT}{V} + \frac{\frac{1}{N}\sum_{\mathrm{i}}^{N}r_{\mathrm{i}}F_{\mathrm{i}}}{3V}\right) \;||\; \text{Pressure} \tag{2.20c}$$

For the NVT properties above, the timestep, or iterations of the MD protocol shown in Fig. 2.3, is incorporated to show the assistance of scale and time averaging. As the number of particles go up, and the length of our simulation increases the refinement of our sampling will grow. However, there is a price to pay for this refinement, and it is with the computational cost of the interatomic potential expression. For each step in the MD protocol the effect of each atom's neighbor interactions must be calculated, and with high-fidelity expressions for $V(r)$ the computational barrier per system timestep can outweigh any theoretical accuracy. Continuing to reduce the necessary physics to produce well-controlled simulations, while expediting the rate of simulation is key for scalable work.

2.4 *ab initio* vs. the interatomic potential: explicit to implicit

Building on the discussion highlighted in Sec. 2.3.1, the need for increased particle count simulations and longer timescales is the ultimate bridge for materials engineering. Ultimately, the most common solution is to continue removing the explicit calculation of electrons, and replace interactions with growing coarseness. The simplest formalism for particle interaction is akin to hard sphere interactions like billiards, croquet, or golf. Using a similar functional shown in the MD protocol in Fig. 2.3, the simplest models for pair-wise potentials were developed that consisted of attractive and repulsive interactions. The Lennard-Jones or 12/6 potential [29] is one of the most common with functional exponential form below:

$$V_{LJ}(r) = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{2.21}$$

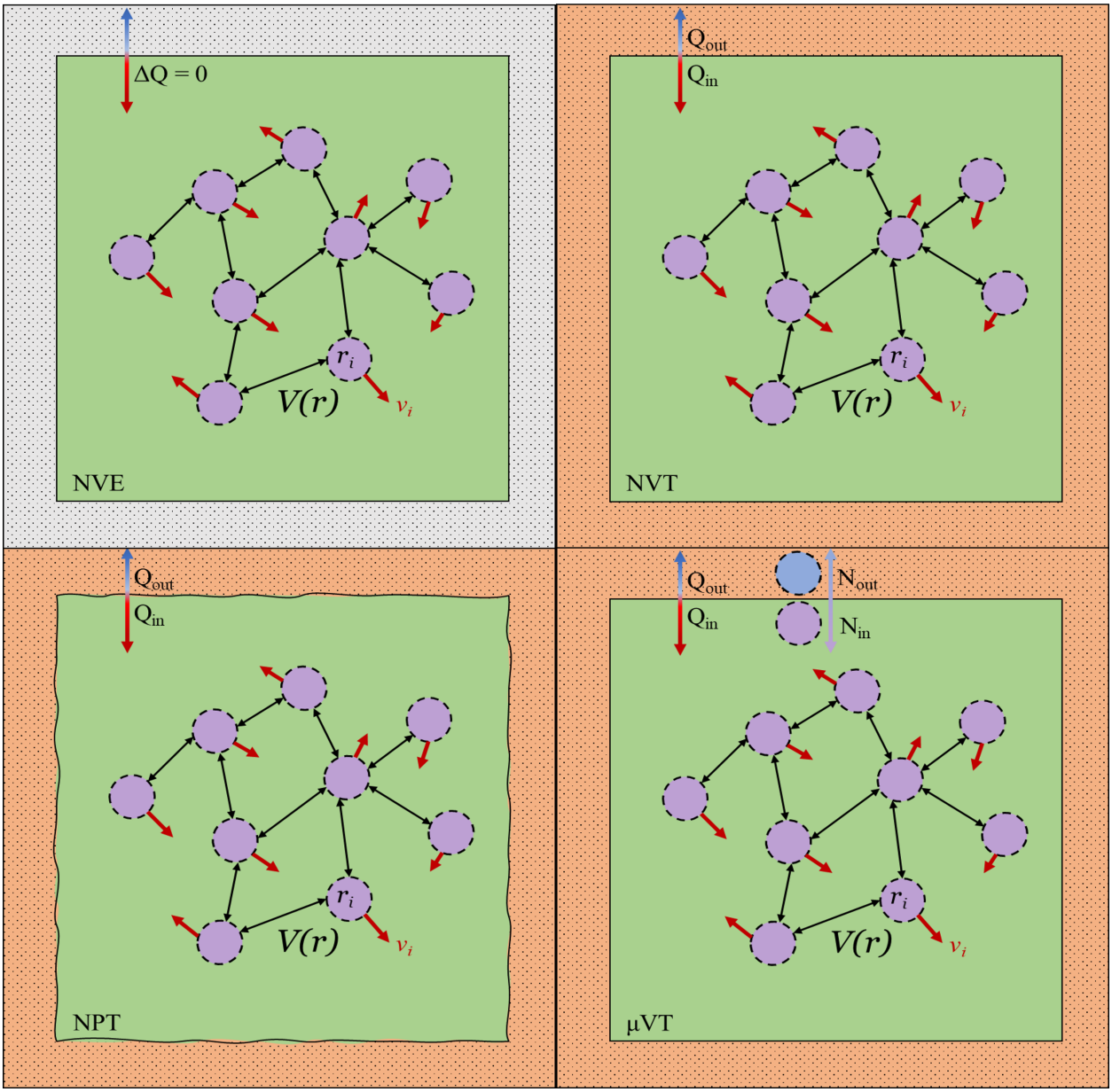

Figure 2.4. Thermodynamic ensembles NVE [upper left], NVT [upper right], NPT [lower left], and μVT [lower right]

In the baseline interatomic potential function, an attractive and repulsive force balances the atomic separation to determine bonding energy and atomic spacing. However, the initial assumption proposed by Lennard-Jones is only complete for ideal gases. Efforts to incorporate other exponential forms in the Morse potential [30], covalent bonding with Tersoff [31], multi-directional effects in semiconductors with Stillinger-Weber [32], and metallic electron density functionals in EAM [33], [34] and MEAM [35], and potentials for studying organics and biomolucules [36], van der Waals interactions [37], and reactive pathways [38] have been put forth over the years. Each of them have their advantages and disadvantages for systems, and all must be chosen understanding how the fundamental assumptions were created.

For a majority of the MD simulations performed in this work the Large-scale Atomic/-Molecular Massive Parallel Simulator (LAMMPS) [39], [40] was used with EAM interatomic potentials. MD processes including dynamics and thermodynamic output are parallelized over computational server resources. All simulations used a Nosé-Hoover thermostat and/or barostat where appropriate [41]–[43]. Details on appropriate timestep and coupling constants are included in each chapter's methods.

2.5 Machine Learning for Materials Science

As our domain expertise grows, and the availability of well parallelized computational tools emerge the chance to leverage machine learning tools for materials science grows. In the this section we will detail some of the emerging paradigms of materials engineering, and how machine learning (ML) is allowing us to more efficiently explore the vast compositional and potential energy landscapes.

At its core, ML is a set of mathematical tools to approximate functions, and correlate inputs with outputs. It is a way to use complex datasets and create a cost-effective model that can generalize the label necessary. We will save the specific examples of its historical use in materials engineering for the targeted chapters, and instead focus on framing our sense of how machine learning provides a bridge between multi-scale modeling methods.

When considering inputs for an ML model they must meet the following criteria: be a comprehensive fingerprint of the target value, less expensive to acquire than the target

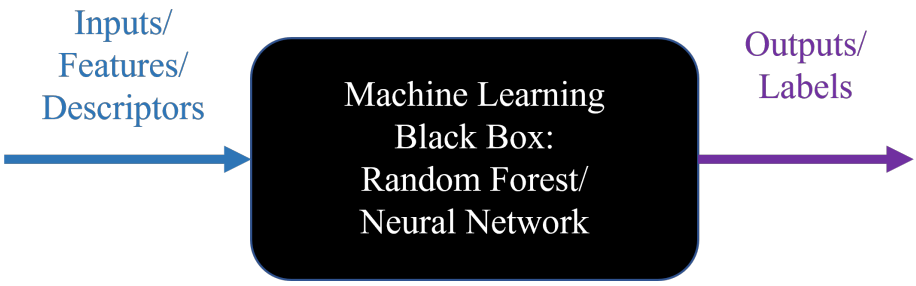

Figure 2.5. Schematic of machine learning design. A statistical representation of our data is leveraged with well defined inputs and outputs for training, and using a set of generated inputs can predict the output of an unknown value.

value itself, and show high spatial variability compared to neighboring features. The labels should be acquired with the highest fidelity available for a problem, and have as little noise as possible. Ideally, the total number of dense data sets (data containing a full amount of features paired with a label, without any blank entries) should be a large and expansive representation of the sample space you wish to explore. While powerful, the greatest power of ML methods comes from their ability to *interpolate* between data entries rather than *extrapolating*. Industrial datasets used for training ML models are generally curated from datasets with 10^6+ volume in entries, and have dedicated teams to cyberinsfrastructure platforms for data management, alignment, and acquisition functionality. As discussed in Ch. 1, the multi-disciplinary nature of materials engineering has made the standardization and accessibility of data initially difficult. However, large strides have been made in recent years for deployment of ML models on open databases, enhanced screening of new composition spaces, feature engineering, and more that we will discuss in the following chapters.

As a community we are often faced with a situation called the 'Inverse Problem'. Here we find ourselves limited by the volume of our data, while hoping to explore a much wider compositional space. Our initial set of data is small, and the desired space is rather large. Conversely with tasks such as image recognition, the data available on a pixel by pixel basis is enormous. Especially when considering the highest fidelity experimental validation, the cost and facility based prohibitions lead to sparse datasets for new materials. A secondary problem that we have is that we often do not publish or provide data that is unsuccessful. These datapoints are categorized as 'failures' because they did not meet the pre-established criteria and the values often go unreported in published literature. These 'failures' are actually positive points for unsuccessful samples that can help an ML model make more realistic predictions in extrapolatory space.

For the application of ML to materials engineering problems we will consider two types of featurization sets that we will use for a range of material systems to demonstrate their utility in the field. **First**, the basis for feature sets that describe a chemical specie, and correlate those descriptors to material properties based on crystallographic information and chemical formula. These tools will be built on small datasets acquired from repositories and open source projects in alignment with the MGI. Showcase studies on interpretable physics based

descriptors are included, with transfer and active learning applications for model development. Due to the limited initial datasets (10^{1-3}) we will use random forest methods for regression due to their high performance on limited data, and their interpretability with physics based descriptors. **Second**, we will narrow our scope, and assess the use of atomic-level descriptors to capture environment dependent and spatially invariant geometries for individual atom properties. Once we begin sampling from a particle and phase-space rather than produced composition based properties the volume of our data expands dramatically, and allows for more complex tools to be applied. Here we turn to neural networks for predicting the local atomic property of a system given a geometric fingerprint.

2.5.1 Random Forests

Random forests (RF) are a subset of machine learning, with capabilities for both regression and classification of data [44]. They are comprised of a set of decision trees which select subsets of our data and feature sets to make judgement calls of a label based on an observable feature. In the case of Fig. 2.6, we use three primary feature sets to make a decision if the specie in question is indeed 'My Cat'. However, as you will see from the crudely selected features there is room for error in each set where a model could make an erroneous prediction. If I possessed a hairless cat the model would immediately miss the first decision tree value. Continuing down the decision network, while uncommon, some cats do have rounded ear profiles and again our model would return the wrong value. Finally, we look to a split in numerical data like the weight category. Depending on when the model was applied, there are certain period of time where the model would not capture 'My Cat' since she was nearly 11 lbs. Since the choices made by an individual tree can often 'overfit' the data or return unstable results we generally take an aggregate score from an ensemble of trees, much like the collective ensemble of atoms in statistical mechanics. For regression tasks each decision tree produces a score, and the mean of the forest is taken as the model output. From here bootstrapping methods can be applied to assess uncertainties of the forest, and individual feature contribution.

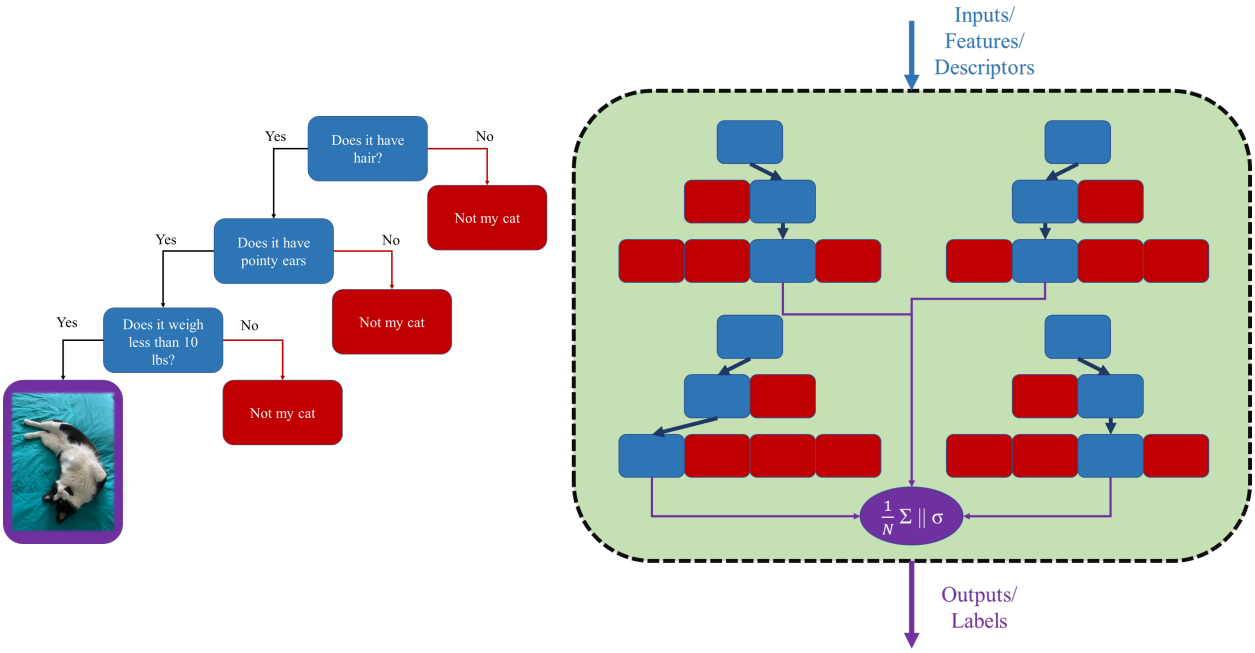

Figure 2.6. [Left] An individual decision tree to determine if the specimen in front of me is my cat. Given three features or questions: does it have hair, are its ears pointed, and the weight a prediction can be formed. [Right] A schematic of an ensemble of decision trees, or a random forest. The use of multiple decision trees can be used in regression with a group average, or classification with a majority vote.

A key advantage to RF methods is that they boast high performance and interpretability on relatively small datasets. A critical feature of the dataset is that each entry be dense to avoid gaps, but even with the selective process of a random forest these limitations can be overcome. Even more impressive is their ability to train on datasets that are on the order of tens or even hundreds of datapoints. This volume of data is something that as materials engineers we are more familiar with in our studies. Data procurement generally involves calculated or experimental properties of a handful of materials, with high-throughput experimentation yielding either noisy data, or is limited by factors of cost.

In this work, we use random forests to predict various properties of materials given an input composition. These compositions represent the macroscopic chemical formula, and do not address microscopic effects such as gradients in composition, grain boundaries, or other secondary effects. However, given careful selection of features and supplementary models in tandem we can create improved models that have many-body physics feature sets built into them. For example, when considering the bonding state of a material we want to understand the fundamental electronic effects involved. These states can be calculated with DFT principles, and elastic constants can be extracted from the output. As highlighted in Sec. 2.2.1, these calculations come with both high accuracy and cost. To overcome some of the challenges associated with cost, speed, or physical resources of DFT an ML model such as an RF can be trained on a subset of well-converged DFT results, and the associated stoichiometry be labeled as inputs. Other properties could then be leveraged from the bonding information model such as kinetic behavior or even melting temperature. Case studies and applications of RF models with uncertainty quantification, active learning protocol, and feature explanation are shown in Ch. 4.

2.5.2 Neural Networks

Built on the biological function of our own neural processes powered by electrical signals, the synthesis of digital neurons with digital inputs and outputs is synonymous [45]. At each nodal point in a neuron, a summation of inputs are grouped with an activation function, a weight, and a bias term [46]. These conversions turn our physical input data into a series of

signals that are 'fed-forward' through the single-neuron design shown in Fig. 2.7. During a training process the output is evaluated to a set of sequestered data called the validation set, and a back-propagation scheme is used to update the weights and parameters of the neuron based on an objective function. To phase it another way: given a penalty associated with the accuracy of an output for a configuration of weights and biases we modify the neuron parameters to find a function that satisfies our ground truth data. Originally these neurons were used as single divider functions operating on a linear activation function between 0/1, but the solution for non-linear functions remained a challenge. Even more complex would be the linking of many neurons together into grouped layers, or interconnections such as graph networks. The compounding complexity of the interacting parameters originally made analytical solutions impossible. The grouping and connection of individual neurons creates what would eventually be termed a neural network (NN).

The first breakthrough for NN design was the efficient solution of back-propagation of errors which led to the advent of most modern day NN architecture [47]. The enriched development of modern day gradient descent optimizers [48], [49], boosting functions to avoid local minima convergence [50], cyberinfrastructure scaling enhancements [51], and local model explainers [52] have allowed these once dusty computational tools to be used by a wider distribution of communities.

The materials engineering community in particular has begun to apply these NN tools to a wide range of problems. In addition to being used for chemical featurized property predictions in tandem with RF studies, deep NNs and complex ML methods have seen wide use in the characterization of local atomic environments. Revisiting our discussion on statistical mechanics in Sec. 2.3.1 we will explore how the application of geometric fingerprints such as the bispectrum coefficients and Gaussian symmetry functions allow us to probe deeper into the chemistry of materials and associate the bridge between local atomic structure and atomic energy. Using carefully selected statistical ensembles of data we are able express complex chemical environments with simple geometric descriptors.

The first case study shown in Ch. 5 uses a NN formalism to connect unrelaxed geometric descriptors of local atomic environments with fully relaxed atomic properties such as vacancy formation energy, cohesive energy, and local stress. The distribution of these properties

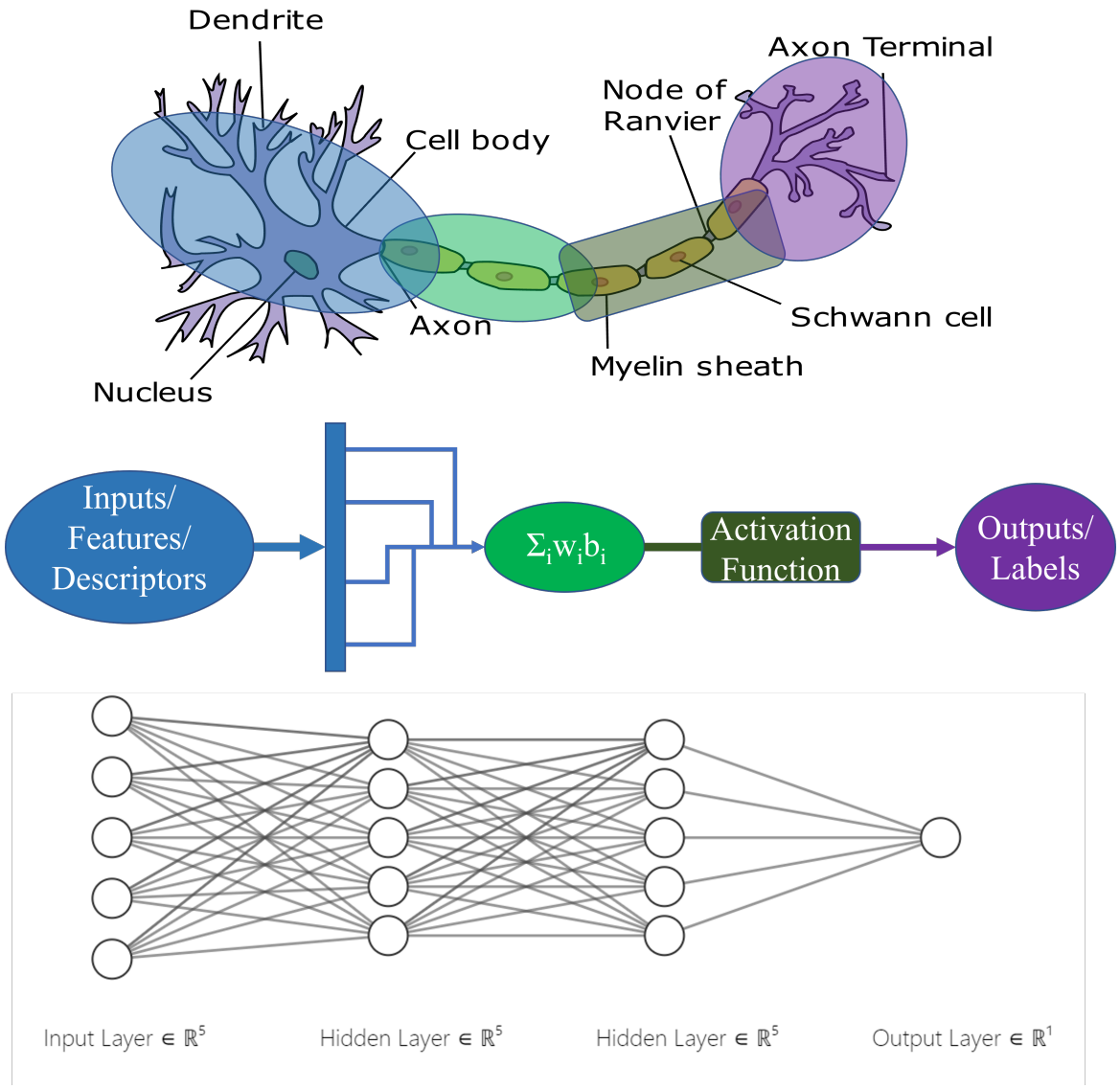

Figure 2.7. [Top] Representation of the human neuron and its structure. Reproduced from Wikimedia Commons [53]. The neuron was used as the baseline for individual perceptron construction in a neural network. [Middle] The single-neuron schematic using the input feature array, applying a sum of weights and biases, an activation function, and the correlated output. [Bottom] A complex, feed-forward, fully-connected neural network. Each line represents a connection between layers with an activation function, weight, and bias term. Generated from open NN visualizer [54]

are assessed and the implications of grouping properties, or atomistically sampling when scaling to macroscale properties is discussed. Finally, we will use the combination of all the above mentioned methods into the capstone of this thesis' work. Using an established atomic ensemble dataset generated from DFT trajectories we apply geometric descriptors of individual atoms to capture the interatomic potential dynamics with many-body interactions. The learned dynamics are used to propagate new trajectories and enrich the database with both 'poor' and 'good' data. Here the comparison between 'poor' and 'good' is used to associate high or low energy states discovered by the neural network potential (NNP). Ultimately, we show that using an algorithm with an exploratory network allows for well converged physics based simulations with heightened speed compared to the full DFT simulation, while sacrificing little accuracy. This final bridge between electronic structure theory, statistical mechanics, and scaled simulations with machine learning as the driver will conclude the academic portions of the thesis.

3. NANOSCALE MODELING OF REACTIVE METALLICS

ADAPTED FROM:

Zachary D. McClure, Samuel Temple Reeve, and Alejandro Strachan. Role of electronic thermal transport in amorphous metal recrystallization: A molecular dynamics study. The Journal of Chemical Physics, 149, August 2018. ©2018 AIP Publishing. [55]

Christopher B. Saltonstall, **Zachary D. McClure**, Michael J. Abere, David Guzman, Samuel Temple Reeve, Alejandro Strachan, Paul G. Kotula, David P. Adams, and Thomas E. Beechem. Complexion dictated thermal resistance with interface density in reactive metal multilayers. Physical Revew B, 101, June 2020. ©2020 American Physical Society [56]

3.1 Introduction

In this chapter we will explore the phase transitions and thermal kinetics of reactive metallics at the nanoscale. Our methods of study will include molecular dynamics driven by an interatomic potential in the first portion, followed by full density functional theory calculations of electron transport for a more complex system. In each of these works we leverage information from different scales to improve the fidelity of our results, or to expedite dynamics in structure generation. In regard to the study of Ni recrystallization we assess the impact of adding a 'switch' in our simulation that acts as an electron thermostat. It is well known that the thermal transport predictions in metals from classical dynamics are underestimated, and in this work we characterize this effect on size dependencies in Ni. Density functional theory information about the material electron-phonon coupling process was supplemented for higher fidelity transport calculations.

Acknowledging the importance of full-fledged electron transport calculations in metallic processes we used a combination of interatomic potential driven molecular dynamics for structure generation, and DFT formalisms for electron transport in metallic multilayers. Our results were expedited due to the team's discipline and the field's alignment with MGI initiatives. Mechanisms of experimental data were explained with well defined electron transport calculations, while using molecular dynamics data to create stochastic averages of inelastic electron-phonon interactions.

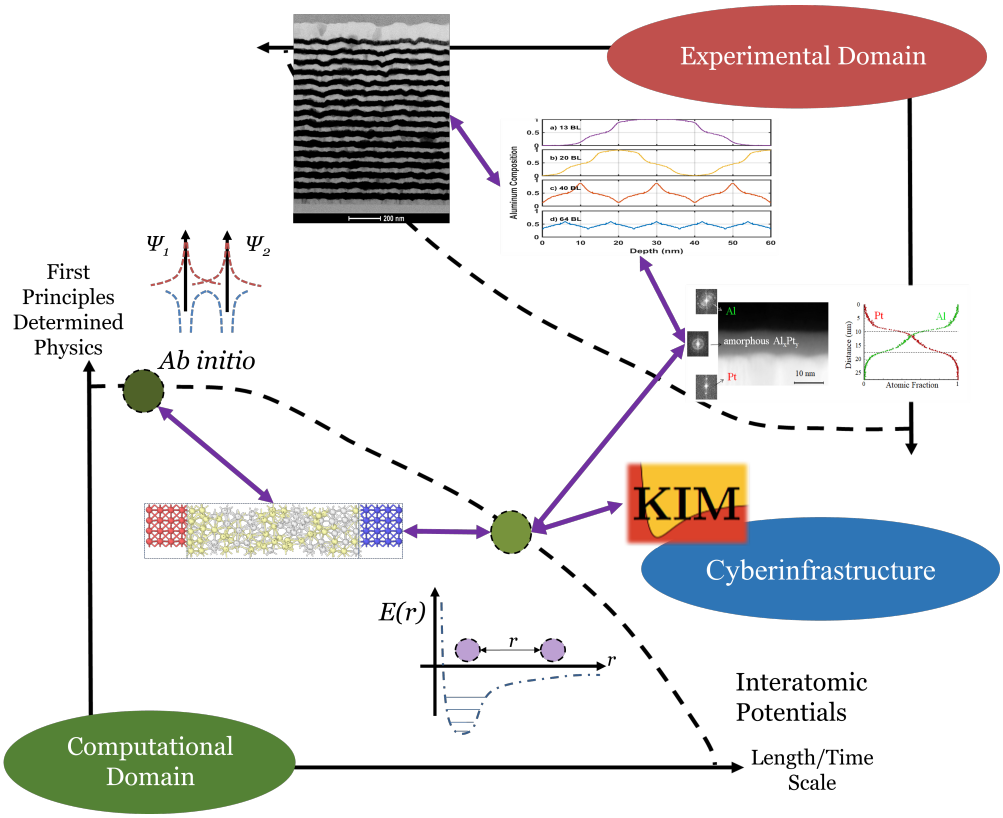

Figure 3.1. Flowchart for experimental, simulation, and cyberinfrastructure data acquisition.

3.2 Recrystallization of amorphous metals

Amorphous materials are attractive for a wide range of applications from sports equipment [57], structural support [58], [59], functional bio-replacements [60], phase-change memory (PCM) [61], magnetic devices [62], and reactive composites [63], [64]. In many of these applications, the mechanisms and kinetics associated with recrystallization dominate performance. PCM devices operate by reversible switching between an amorphous phase of high resistance and a crystalline phase of low resistance and have attracted significant interest for scalable non-volatile memory applications [61], [65]. In these devices the speed of recrystallization from the glassy phase in PCMs is a critical performance metric associated with switching speed. The degree of undercooling, thermal history, and material stability are known to affect recrystallization behavior and this knowledge has been used to improve the operation of PCM devices [66]. Amorphous components have also found application in reactive alloys, where a higher free energy and driving forces than their crystalline counterparts improve performance8 for potential use in lead-free explosive primers [67], [68], and soldering [69], [70]. In this case, the stability of the amorphous phase is important. Amorphous metals are thermodynamically metastable and can be synthesized through multiple routes [71], including fast cooling from the liquid [72], or through aqueous chemical reduction [73], [74]. Given appropriate thermal or mechanical stimuli these glasses can recrystallize via a solid-to-solid, exothermic phase transition [75]. As the material recrystallizes, the released energy creates a self-propagating reaction front. The velocity of this reaction front is dependent on intrinsic material properties including exothermicity of the reaction, thermal conductivity of the material, microstructure, and defects. Despite the significant interest and importance of glassy metals several questions remain unanswered. For example, large scale predictive atomistic simulations have been able to characterize recrystallization behavior in amorphous materials [64]; however, these methods ignore important processes in electron dominated thermal transport.

Self-propagating recrystallization have been observed experimentally in, for example, amorphous silicon [76], [77] and amorphous Ni-Fe foils [78]. Molecular dynamics has contributed to our understanding of recrystallization in semiconductors and insulators, where

phonon transport dominates the thermal properties [79], [80]. MD simulations of recrystallization in metals have been explored to a lesser degree. However, Manukyan et al. studied self-propagation velocities in agglomerates of Ni nanoparticles and compared to experiments [64]. The authors found that the simulations predicted significantly higher propagation velocities. While several effects contribute to this discrepancy, we hypothesize that the underestimation of thermal transport in metals through missing electronic effects is a critical factor. In this paper, we explore the effect of thermal conduction by electrons in recrystallization of amorphous Ni.

Carriers for thermal transport in metals are phonons (or the equivalent collective ionic modes in non-crystalline solids) and conduction electrons; while they have a smaller specific heat the latter dominate thermal transport [81]. Standard MD simulations describe thermal transport through ionic processes, but do not include electronic contributions. Multiple methods have been developed that can account for electron interaction and transport including the two-temperature model (TTM), which models electron thermal transport through a diffusive heat equation [82], [83], as well as dynamics with implicit degrees of freedom (DID), where electrons are modeled as degrees of freedom (DoF) for each atom [84]. Here, we add electronic effects with the TTM to simulate more realistic behavior for amorphous metal recrystallization.

3.2.1 Atomistic model of Ni and sample preparation

All MD simulations were performed using the LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) software package developed by Sandia National Laboratory [40]. Visualization was performed using OVITO (Open Visualization Tool) [85] with the polyhedral template matching (PTM) algorithm [86] to identify atoms belonging to crystalline and amorphous phases. Throughout, atoms with an FCC local environment are colored green and atoms with unidentified coordination are white. An embedded-atom method (EAM) potential for Ni was used in our simulations, parameterized by Mishin [87] (accessed through the NIST Interatomic Potentials Repository under Ni-Al 2009). This model was parameterized using density functional theory (DFT) formation energies and experimental

results including cohesive energy, lattice parameter, and elastic constants. To assess the accuracy of the interatomic potential to describe amorphous Ni configurations we computed the melting temperature using the coexistence simulation technique [88] and heat of fusion as the enthalpy difference between liquid and crystalline samples at the melting temperature. The melting temperature predicted by the potential is 1750 K and the predicted heat of fusion is 0.19 eV/atom, both in good agreement with the experimental values of 1728 K and 0.181 eV/atom, respectively [89].

All recrystallization simulations were performed with a crystalline seed in contact with an amorphous sample, extended in one direction. A similar simulation was previously performed to investigate crystallization of amorphous Ni nanoparticles [64]. A timestep of 1 fs was used throughout, with damping constants of 0.1 ps and 1.0 ps for the Nose-Hoover thermostat and barostat, respectively. All simulations were run at 1 atm pressure. Crystalline seeds were generated by replicating the FCC unit cell 5x10x10 times. The resulting crystalline seed initially measured 1.76 nm by 3.52 nm by 3.52 nm with periodic boundary conditions and was equilibrated using the isothermal-isobaric (NPT) ensemble at 300 K for 100 ps. Amorphous samples of different lengths were generated via melting, deformation, and quenching. The initial configurations were obtained by replicating the FCC unit cell 10x10 along the directions of the crystal/amorphous interface leading to cross-sectional dimensions of 3.52 nm by 3.52 nm with periodic boundary conditions. The unit cell was replicated between 100 and 300 times along the recrystallization direction resulting in initial system lengths between 35.2 and 105.6 nm. A temperature ramp from 300 to 3000 K in 100 ps was used to melt the sample under NPT conditions. To prevent strain in the final structure, the system was deformed in the melt at 3000 K to match the transverse directions of the crystalline seed lattice, conserving the total volume of the cell. The system was then quenched from 3000 to 300 K in 100 ps under NPT conditions. The quenched amorphous samples were equilibrated at 300 K with fixed cross section and NPT conditions along the recrystallization direction. Five independent samples for each system length were taken during equilibration in increments of 10 ps. Samples for recrystallization were generated by combining the crystalline seed and amorphous bulk. The two systems were added to the same simulation cell (already possessing identical transverse dimensions) with a gap of 5 Å, see Fig. 3.2. A vacuum layer

of 3.5 nm was added in the longitudinal direction to ensure only one crystal/amorphous interface. The total system was then relaxed with energy minimization and equilibrated for 5 ps under NPT at 300 K.

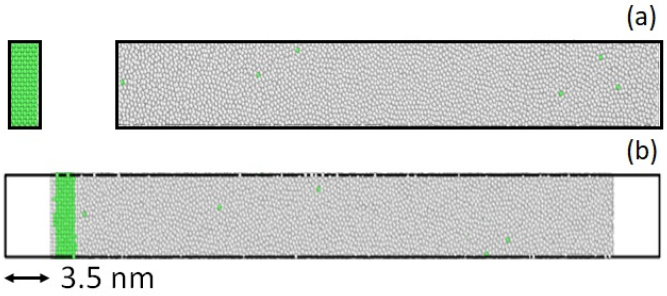

Figure 3.2. (a) Crystalline seed and amorphous bulk with periodic boundary conditions after separate equilibration. (b) Combined structure for recrystallization experiments with added vacuum layers.

3.2.2 Recrystallization simulations

The approach for investigating amorphous nickel recrystallization was first tested without electronic effects. Two different methods were applied to investigate the amorphous nickel recrystallization. The first was under isothermal and isobaric conditions (NPT ensemble) throughout the recrystallization. The amorphous and seed regions were heated to various set temperatures, and then held to recrystallize for at least 0.5 ns. The second method simulated an adiabatic reaction by exposing the seed to a fast thermal impulse while maintaining the amorphous region at 300 K. The set temperature of the seed was ramped to various values in 8 ps and held for 2 ps. The subsequent process can be described as an adiabatic MD simulation, using an isobaric-isenthalpic (NPH) ensemble.

3.2.3 Two-temperature model MD

The TTM used here solves the equations of motion for atoms using MD, adding an electronic temperature field defined over a grid overlapping with the atomistic system.

Electrons and atoms interact and exchange energy which affects the ionic equations of motion via an additional friction coefficient:

$$m\frac{\partial v_{\mathrm{i}}}{\partial t} = F_{\mathrm{i}}(t) - \gamma_{\mathrm{i}} v_{\mathrm{i}} + \widetilde{F}(t) \tag{3.1}$$

where v, m, and F are the velocity, mass, and standard MD force (gradient of potential energy) of atom i, respectively. The term including friction coefficient γ_{i} is the energy loss due to electron-ion interactions and the stochastic force term, $\widetilde{F}(t)$, is determined by the local electron temperature as a Langevin thermostat.

$$\widetilde{F}(t) = \gamma_L v(t) + \xi(t) \tag{3.2}$$

The stochastic force is determined by a Langevin friction coefficient of the system, γ_L, and a random force obtained from a Gaussian probability distribution, ξ, giving the effect of background noise. Energy loss between electrons and phonons occurs via two pathways depending on the velocity of an atom [90], [91]. For high atomic velocities (e.g. approximated as 5.4 km/s for Fe [83]) the valence electrons slow the atoms through a process known as electronic stopping. At lower velocities, of interest in this work, electron-ion interactions act to bring the two sub-systems to thermal equilibrium. Atomic motion is slowed down or sped up proportionally to the difference in the electronic and atomic temperature. γ_s is the friction coefficient due to electronic stopping and γ_p is the friction coefficient due to electron-ion interactions. At relatively low atomic temperatures only the electron-phonon interaction plays a role and γ_s is accordingly set to zero. For an atomic velocity v_{i}, the interaction between electrons and phonons will change above or below a cutoff velocity v_0.

$$\gamma_{\mathrm{i}} = \gamma_s + \gamma_p \text{ for } v_{\mathrm{i}} > v_0 \tag{3.3}$$

$$\gamma_{\mathrm{i}} = \gamma_p \text{ for } v_{\mathrm{i}} \leq v_0 \tag{3.4}$$

The electronic degrees of freedom transport heat and the TTM solves the heat diffusion equation for the electronic temperature at discrete grid points throughout the atomic structure. Our electron grid was set with one point for each two lattice spacing units in each direction and was found to be well converged. At each grid point the heat diffusion equation is coupled between the atoms and the electrons:

$$C_{\mathrm{e}}\rho_{\mathrm{e}}\frac{\partial T_{\mathrm{e}}}{\partial t} = \nabla(\kappa_{\mathrm{e}}\nabla T_{\mathrm{e}}) - g_p(T_{\mathrm{e}} - T_a) + g_s T_a^{'} \tag{3.5}$$

where C_{e} represents the electronic heat capacity, ρ_{e} the electronic grid density, and κ_{e} the electronic thermal conductivity. For our relatively low temperature simulations we ignore friction effects from electronic stopping by setting γ_s to zero. In turn, the coupling parameter for electron-stopping, g_s, is zero and the electron-phonon coupling parameter, g_p, is:

$$g_p = \frac{3Nk_b\gamma_p}{\Delta V m} \tag{3.6}$$

where N is the total number of atoms in the electronic grid, k_b is the Boltzmann constant, ΔV is the electronic grid volume, and m is the atomic mass. The related inverse relaxation time, or the coupling coefficient, χ, is:

$$\chi = \frac{\gamma_p}{m} \tag{3.7}$$

The TTM implementation used was extended to allow vacuum layers within laser ablation simulations [92], [93]. This was important for the simulation to ensure recrystallization along a single front (see Fig. 3.2). Finally, we note that electronic properties depend on the local atomic structure. For example, the thermal conductivity would be reduced in the amorphous regions. In this first study, we ignore such effects.

TTM verification tests and input parameters

A series of TTM MD simulations under adiabatic conditions (NVE ensemble) were performed for verification purposes and to test input parameters. A system of 4,000 atoms (40x5x5 unit cells) was created and equilibrated at 300 K using NPT. A series of TTM

simulations were carried out to assess the effect of electronic specific heat, Fig. 3.3(a), and electron-phonon coupling constant, Figure 3.3(b), on the equilibration of the electronic and ionic temperatures by setting the initial temperature of the electronic subsystem to 600 K.

Role of electronic specific heat

Fig. 3.3(a) shows the time evolution of electronic and ionic temperatures for simulations with three different electronic heat capacities: 0.3, 3, and 3 k_b/atom. All three simulations use a coupling constant of $\gamma_p = 10$ g mol^{-1} ps^{-1}, which results in a coupling coefficient of $\chi = 0.17$ ps^{-1}. The results of the simulations are as expected. Recall that within the classical harmonic approximation, the heat capacity of a solid in 3D is 3 k_b/atom. By setting the electronic specific heat to an unrealistically high value of 3 k_b/atom both sub-systems equilibrate at a temperature approximately half way between their initial values. However, an electronic heat capacity one-tenth of the atomic subsystem shows a more realistic process where electrons equilibrate to a temperature close to the initial ionic temperature [84], [94]. The heat capacity of electrons is temperature dependent, but to simplify the numerical aspects of our simulations we use a constant value 0.3 k_b/atom for the remainder of this work. This is a reasonable approximation for our problem of interest since the electronic heat capacity of Ni varies from 0.29 to 0.48 kb per atom in the temperature range 300-1500 K [94].

Electron-phonon coupling

Electron phonon coupling constants can be determined experimentally [95], [96], as well as from first principles [94]. The electron-phonon coupling constant for Ni has been reported with a range of $\chi = 0.10 - 0.29$ ps^{-1} from DFT simulations [97]. Estimates of the electron-phonon coupling coefficient through numerical analysis resulted in $\chi = 1.0$ ps^{-1} for Ni and 0.05, 0.03, and 1.5 ps^{-1} for Ag, Cu, and Fe, respectively [98]. Fig. 3.3(b) characterizes the process of equilibration between electrons and ions with coupling constants varying by three orders of magnitude and a fixed heat capacity of $C_e = 0.3$ k_b/atom. The results show the expected behavior for comparable values to those above and demonstrate that for Ni, equilibration occurs with a timescale of approximately one ps.

Electron thermal conductivity

Another important parameter in the TTM is the electronic thermal conductivity, given by the product of the electron grid density ρ_e, the electron heat capacity C_e, and the electron thermal diffusivity D_e. For the TTM simulations D_e= 2 cm^2 s^{-1} was used [93]. This corresponds to an electron thermal conductivity for Ni of $\sim$ 74 W m^{-1} K$-$1. Electron thermal conductivity accounts for nearly 90% of transport in bulk and at metal interfaces [81]. The experimental thermal conductivity for Ni is 90 W m^{-1} K^{-1} [99] and is the sum of both electronic and phonon thermal contributions. MD calculations of the phonon contribution range from 2-10 W m^{-1} K^{-1} [100]. Using the sum of the calculated electronic and phonon thermal conductivities our input for the TTM matches experiments well.

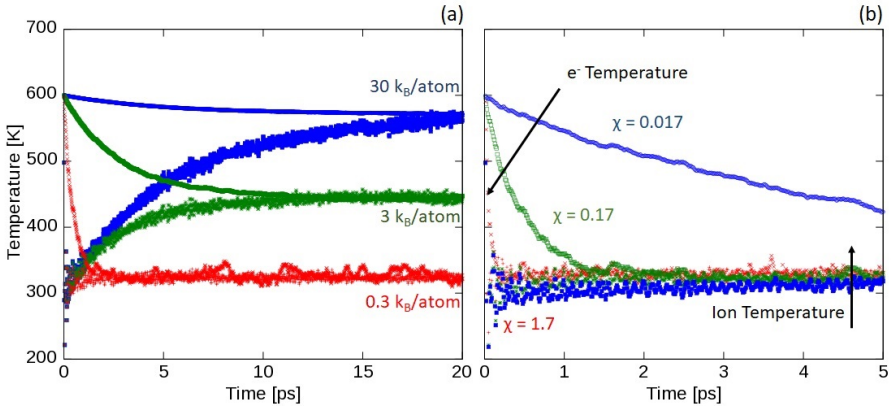

Figure 3.3. Temperature equilibration between atomic (T_0=300 K) and electronic (T_0=600 K) subsystems. (a) Varying electronic heat capacity with χ = 0.17 ps^{-1}. (b) Varying electron-ion coupling constant with C_e = 0.3 k_b/atom.

3.2.4 Recrystallization without electronic effects

We begin with standard MD recrystallization, without added electronic effects and under NPT conditions. Fig. 3.4 shows the atomistic structure during recrystallization for a representative system at 1000 K. The crystalline seed acts as a heterogeneous nucleation site for the amorphous metal and the recrystallization front moves across the system.

Throughout all simulations, the velocity is calculated from the slope of the linear recrystallization front position as a function of time, as shown in Fig. 3.5. Here we show the

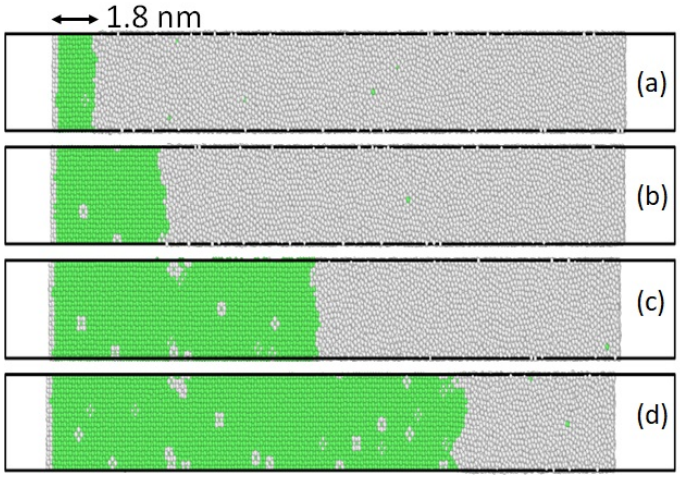

Figure 3.4. Snapshots of MD simulation of recrystallization of amorphous Ni at 1000 K at increasing time (a) 10, (b) 100, (c) 300, (d) 500 ps. Atoms in green are FCC coordination, while white atoms are unidentified.

recrystallization front as a function of time for systems with and without added electronic degrees of freedom. The velocity of the reaction front is fitted from this profile. We note a significant difference in the velocity of systems with and without electronic degrees of freedom. To ensure that we capture steady state we only fit the velocity to the middle portion of recrystallization. This ensures that we do not capture secondary transient effects of initial nucleation, and final recrystallization of the entire system.

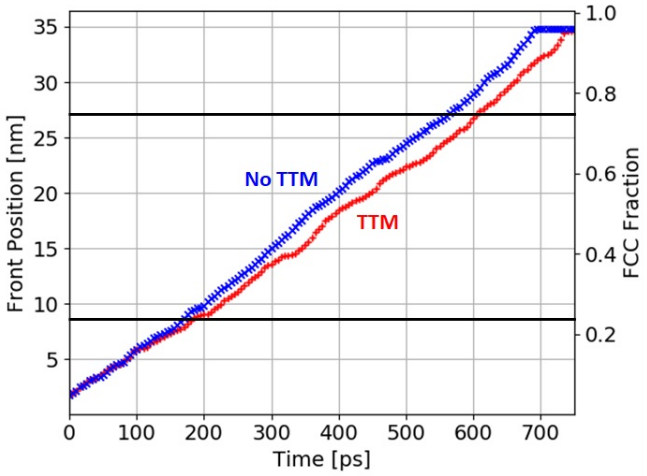

Figure 3.5. Recrystallization front as a function of time for representative MD simulations with and without TTM electronic effects. We fit velocities within the range of 25-75% conversion of amorphous to crystalline phases (solid black lines). These limits ensure that steady state recrystallization has been achieved and that we do not measure near full conversion.

Under isothermal conditions, the crystallization velocity vs. temperature exhibits a maximum at a value below the melting temperature that compromises between driving force for crystallization and kinetics. At lower temperatures kinetics dominate and reduce crystallization front speed. At higher temperatures the thermodynamic driving force for crystallization (the free energy difference between the amorphous and crystalline phases) decreases. Fig. 3.6 shows the average recrystallization front velocity as a function of temperature (black circles).

The velocity increases with increasing temperature up to 1200 K. Beyond 1200 K, a sharp drop in velocity is observed; this is consistent with previous MD simulations of Ni recrystallization [64]. Isothermal conditions are artificial as they remove or add energy to

the system in non-physical ways. In addition, a global thermostat couples with the total kinetic energy of the system and is unaware of local temperature variations, such as the local heating around the crystallization front. Adiabatic conditions represent experimental conditions for fast recrystallization processes more accurately. Thus, constant pressure and enthalpy (NPH ensemble) simulations were conducted and red squares in Fig. 3.6 show the resulting front velocity as a function of the initial seed temperature. As expected, we observe a weak dependence of recrystallization velocity on initial seed temperature as the system uses the heat generated from recrystallization to continue the process and evolves towards a steady state. This behavior is the focus for the remainder of this section.

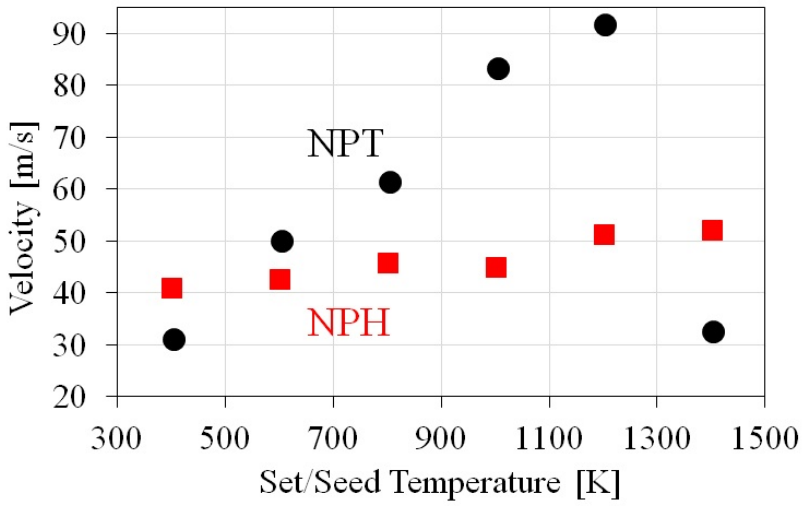

Figure 3.6. Crystallization velocities for isothermal-isobaric (NPT) and adiabatic (NPH) simulations. Temperatures shown represent the set temperature for the entire system in NPT, and the initial seed temperature in NPH simulations.

The details of the process under adiabatic conditions can be seen in Fig. 3.7(a) where we show temperature profiles at various times. At 10 ps, the crystalline seed has been heated to its set point temperature while the amorphous region was held at 300 K. By 100ps, the temperature of the crystallized region has dropped to 600 K, as some heat has diffused away from the recrystallization front to the rest of the amorphous sample. Following this initial stage, we observe steady state propagation of the recrystallization front where the vertical

lines in Fig. 3.7 indicate the location of the front at each time. Throughout the simulation the hottest part of the sample is within the recrystallized region. Heat builds up at the reaction front and only slightly increases the temperature of the amorphous region. Without a sufficient thermal conductivity to carry it away a large amount of thermal energy is provided for recrystallization.

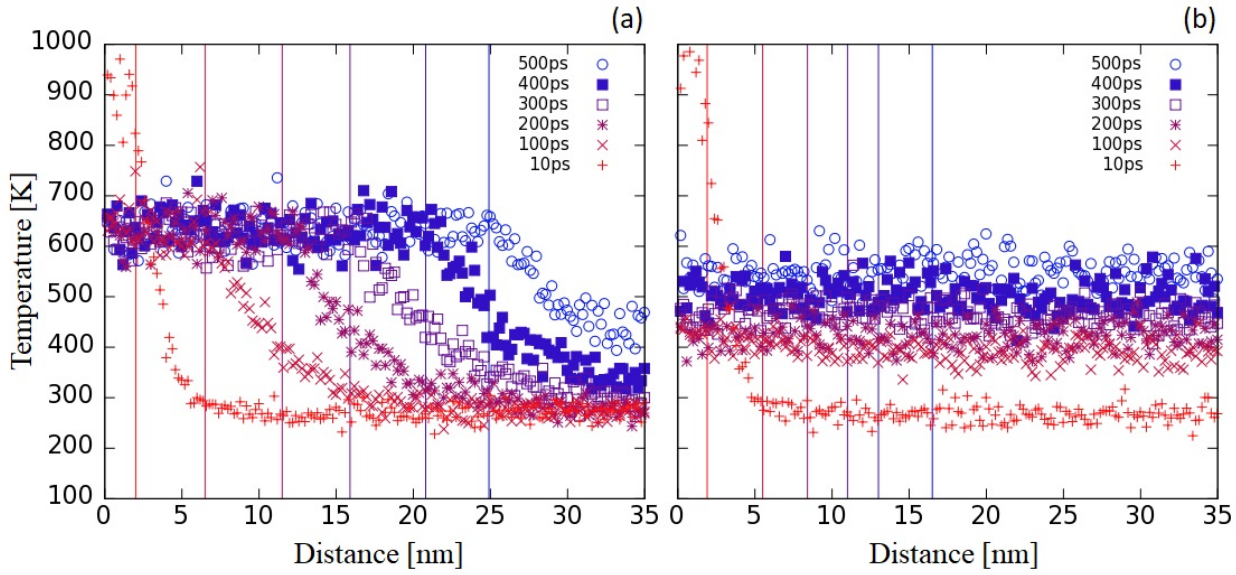

Figure 3.7. Local temperature along the recrystallization direction for (a) standard MD recrystallization and (b) TTM MD recrystallization with $\chi = 0.17$ ps^{-1}. Colored vertical lines indicate recrystallization front position for each time.

3.2.5 Role of electronic degrees of freedom on recrystallization

Adding the thermal role of electrons using the TTM, with parameters for Ni discussed in Sec. 3.2.3, to the adiabatic recrystallization simulations results in the temperature profiles shown in Fig. 3.7(b). The initial localized temperature in the crystalline seed dissipates through the material in under 100 ps. Instead of a stepped temperature profile that moves with the recrystallization front, the system exhibits a nearly homogenous temperature profile. As the reaction proceeds, the exothermic reactions increase the overall temperature of the system. To better understand the effect of electronic thermal transport we study how the

coupling coefficient, χ, and the electronic thermal conductivity, κ_e, affect the recrystallization process and front velocity.

Role of coupling strength

The effect of electron-phonon coupling on front recrystallization velocity is shown in Fig. 3.9(a). For low coupling rates, the energy of recrystallization does not couple to the electronic DoFs sufficiently fast to affect the process. We observe propagation velocities, Fig. 3.9(a) and temperature profiles, Fig. 3.10(a), similar to those corresponding the standard MD simulations without TTM, but we do note a shallower temperature gradient. As the electron-phonon coupling rate is increased, the recrystallization velocity smoothly decreases and then saturates. Once the coupling is strong enough to maintain an equal temperature associated with the two sets of degrees of freedom, an increase in coupling has no effect, leading to the converged velocity. Fig. 3.10(b) shows the effect of electron transport on the temperature profile for a higher coupling, showing earlier times from Fig. 3.7(b). Notably, as electronic transport plays a role in recrystallization, front velocities become length dependent, shown with two simulation cell lengths in Fig. 3.9(a). As the heat is quickly diffused and spread to more atoms, the larger system will be in average colder, and the overall recrystallization velocity reduced. Although the temperature profiles and reaction velocities are notably different with varied electronic coupling, the atomic features of the reaction front remain similar. As shown in Fig. 3.8, we observe no noticeable difference in the thickness and shape of the reaction front between simulations without electronic coupling and different levels of TTM coupling. In all structures, we find a relatively sharp transition between phases with roughness on the order of 1 nm or a few atomic layers. If there were inherently different shapes, or effects of disorder on our interface, it could complicate our linear approximation of crystal growth.

Role of electronic thermal diffusivity

Similarly, a sweep of electron thermal conductivity was performed, shown in Fig. 3.9(b). As with the coupling parameter, we find two regimes in the electronic thermal conductivity

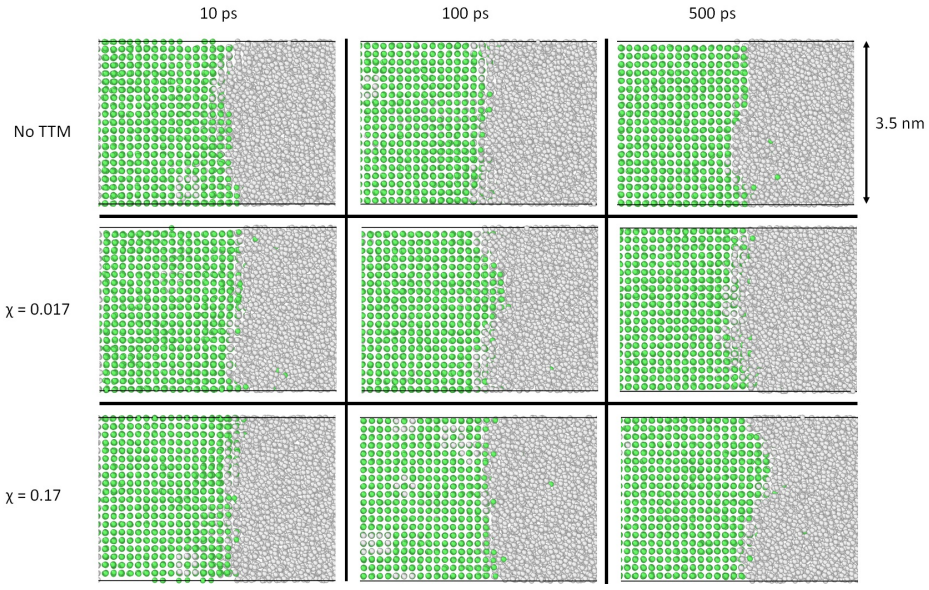

Figure 3.8. Representative atomic structures at the recrystallization interface. Snapshots taken at 20, 100, and 500 ps for simulations without coupling and with TTM electronic coupling of 0.017 ps^{-1} and 0.17 ps^{-1}. The interfaces are similarly sharp, with roughness up to approximately 1 nm

with a smooth transition in between. For small electronic thermal diffusivities (orders of magnitude smaller than the typical values) electronic degrees of freedom have no effect on recrystallization velocity, as phonons are still more effective at distributing heat. Once the electron thermal conductivity is on the same order of magnitude as the phonon thermal conductivity, about 1 W m^{-1} K^{-1}, there is a large reduction in recrystallization rate. At saturation the increased electronic conductivity cannot be used effectively by the system as the energy is distributed throughout the entire sample.

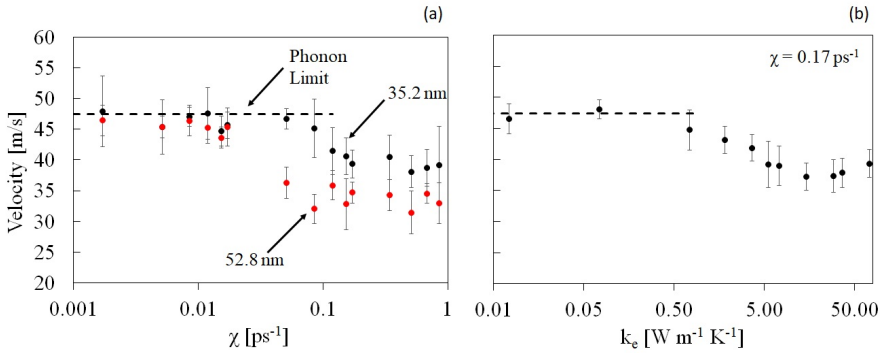

Figure 3.9. Recrystallization velocity as a function of (a) coupling coefficient χ for multiple lengths (Black = 35.2 nm, Red = 52.8 nm) and (b) electron thermal conductivity. Phonon limit (standard MD) represented by dashed line.

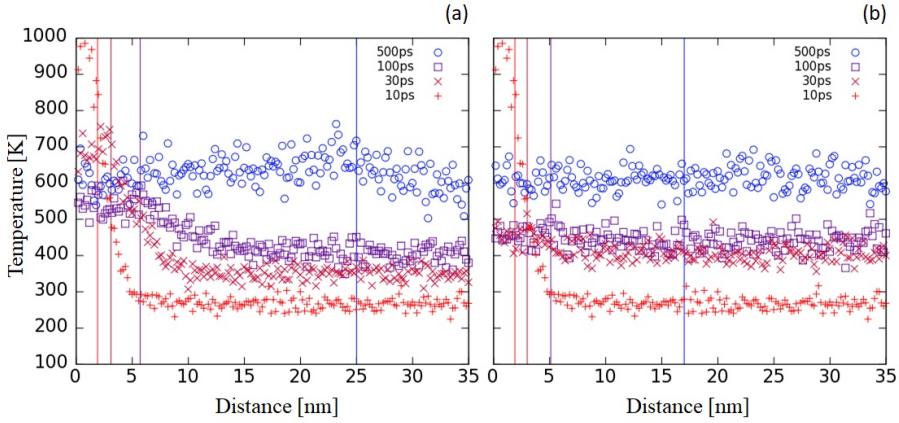

Figure 3.10. Local temperature along the recrystallization direction for (a) χ = 0.017 ps^{-1} and (b) χ = 0.17 ps^{-1}. Colored vertical lines indicate recrystallization front position for each time.

Size effects

Since electrons can spread the heat generated over the entire sample, one can expect size effects on the recrystallization velocity. To characterize such effects, a range of sample lengths were generated and recrystallized. Fig. 3.11 compares simulation results with and without electronic degrees of freedom to previously reported experimental values. As noted before, the standard MD with phonon dominated thermal transport has no length dependence on recrystallization velocity due to its low thermal transport. As expected, increasing the sample size for TTM MD simulations results in a reduction of the crystallization velocity due to the reduction in overall temperature. Fig. 3.11 enables an estimation of the crystallization velocity for bulk samples including the role of electrons. The resulting value of 23 m/s is closer to an experimental value of 10-20 m/s for amorphous Fe-Ni foils [78], and 2-17 m/s for amorphous Si [76], [101]. It should be noted that additional mechanisms such as interatomic diffusion in the Fe-Ni foils will contribute to recrystallization velocities. However, the alloy system provides a reference point for expected kinetics of metallic recrystallization. We note that experimental recrystallization of amorphous Ni powders has been reported at 0.3 mm/s, significantly lower than our predictions and the other experimental data. However, these Ni samples were agglomerates of nanoparticles with large porosity (∼80%) and oxide layers [64], a nanostructure very different from our current MD model.

3.2.6 Conclusions

In this introductory exercise we characterized the role of the thermal transport by electrons on the recrystallization of amorphous Ni. This is done using a two-temperature model coupled to molecular dynamics simulations. The simulation setup uses a heated crystalline seed to start the process of adiabatic recrystallization. We find that the increased thermal conductivity, resulting from the incorporation of electronic degrees of freedom, reduces the propagation velocity of the recrystallization front provided the coupling between ionic and electronic degrees of freedom is fast enough. For weak electron-ion couplings, the recrystallization velocity is not affected by the electronic degrees of freedom as heat from the exothermic recrystallization process is not transferred to the electrons within

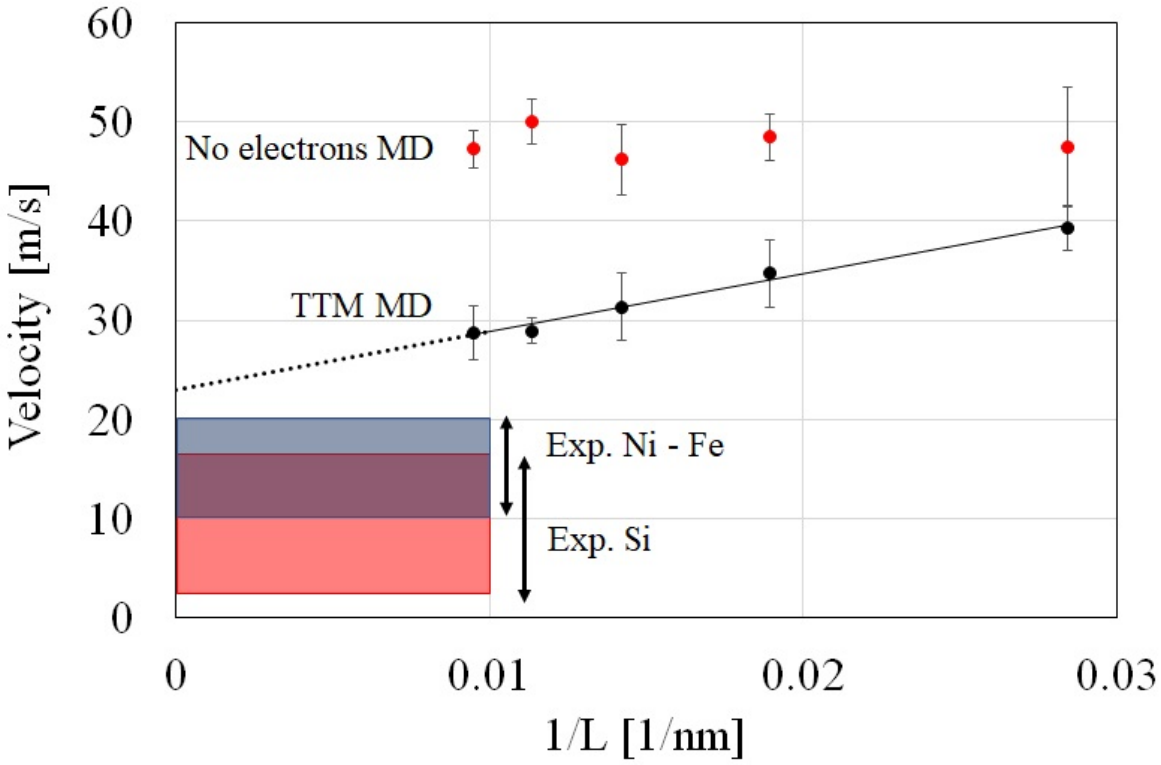

Figure 3.11. Recrystallization velocity extrapolated to bulk sample lengths from MD without electrons and TTM MD, compared to experimental data.

a characteristic timescale. For coupling coefficients between 0.01 and 1 ps^{-1} electrons can dissipate enough heat from the reaction zone to reduce the effective temperature of recrystallization and, consequently, velocity. When electronic effects are added to thermal transport we find significant length effects on recrystallization velocity as electrons diffuse heat rapidly through the system resulting in a homogeneous temperature profile. Previous estimates of recrystallization in amorphous Ni from standard MD greatly overestimated the recrystallization kinetics; adding electronic contributions to recrystallization results in an estimation of a bulk recrystallization velocity for Ni of 23 m/s.

Moving past single component metals, we pursued work on binary alloys with complex microstructures, and the impact of structure on functionality of devices. Building on the work from MD-assisted thermal transport assumptions we acknowledge the need for well defined electron interaction for complex nanostructures. In the following section we will highlight the use of both MD and DFT methods to generate phonon and electron interactions in a system.

3.3 Reactive metal multilayers

Multilayers are ubiquitous across the materials landscape being used to tune all manners of properties. Functionally, the properties of multilayers emerge from the spacing and materials composing the many interfaces implicit in layered solids. These boundaries can affect the wave nature of the energy carriers themselves while also impacting their scattering. Changes in the energy carriers (e.g., photons, phonons, electrons)—and their scattering—necessarily dictate properties. For example, in optical coatings, the spacing and refractive indices of material layers are chosen to create constructive or destructive interference of photons over a selected wavelength range. A similar goal is aspired to when using multilayers in thermoelectrics. In this case, the objective is to minimize phonon transport through scattering across the entire phonon spectrum. To this end, multilayers exhibiting extremely low thermal conductivity have been achieved by leveraging Van der Waals bonds [102], nanolamination [103], and nanostructuring [104].

Regardless of approach, the reported reductions in thermal conductivity evolve from scattering at interfaces consistent with the view that heat transport obeys a series resistor

model where the total thermal resistance is proportional to the number of material boundaries. This view is not strictly valid, however. The paradigm presumes that every interface between two materials offers the same amount of thermal resistance. Reality is not so simple. Variations in disorder [105]–[108], composition [109], [110], and changes in bonding [111], [112] between interfaces can all affect transport. Each of these characteristics, in turn, is subject to processing, environment, and even the spacing between boundaries. Taken together, thermal transport through a multilayer will therefore be intimately tied to interfacial structure and its evolution.

Despite numerous studies for phonon dominated systems where interfacial disorder has been shown to both [112] increase [107], [113] and decrease [81], [105], [106] boundary conductance, the impact of interfacial structure on thermal transport in electron dominated systems has not been addressed in nearly as much detail. Within several metal multilayers, a thin ($\sim$ 10 nm) compositionally mixed volume exists at each boundary [114]. This intermixed interlayer—termed a complexion [115]—profoundly impacts a wide breadth of properties ranging from grain growth and fracture strength to ionic conductivity and electron mobility [116]. The impact of complexion on thermal transport, especially within electron-dominated systems, remains largely unexplored, however. To address this gap, we investigate here the influence of periodicity and interfacial structure on the thermal resistance of metal multilayers to highlight the predominant role of complexion, and more generally disorder, on heat transport within electron-dominated multilayers.

The thermal conductivity of Al/Pt metal multilayers with varying bilayer period thickness was measured using time-domain thermoreflectance (TDTR) and simulated under a nonequilibrium Green's functions (NEGF) framework in which the multilayers are described through a combination of molecular dynamics (MD) and density-functional theory (DFT). Due to the preference of Al and Pt to be fully alloyed, a 10-nm region of amorphous intermixing (i.e., a complexion) exists at each material transition. This intermixing has a significant effect on the thermal transport, most notably inducing a nonmonotonic trend in thermal resistance with the bilayer period thickness. While this observation cannot be explained using traditional approaches such as the electron diffuse mismatch model (eDMM), it evolves due to the changing complexion of the multilayer as its periodicity is altered.

3.3.1 Material and Methods

Multilayers were grown by sputtering Al and Pt with a period thicknesses ranging from 6 to 800 nm [128 to 1 bilayer period(s)] atop a silicon substrate. All films had a total thickness of $\sim$ 800 nm. Multilayer thin films were deposited by direct current, magnetron sputtering using a cryopumped, Unifilm Co. PVD-300 system–base pressure $< 2.6 \times 10^5$ Pa (2.6×10^7 Torr). Film deposition onto clean, 300 µm-thick Si(100) substrates, involved the sequential sputtering of 99.9995 at.% Al and 99.95 at.% Pt targets using ultra-high purity Ar gas. Substrates were rotated and moved under sputter targets in order to precisely establish uniform layer thicknesses. Argon flow rate was controlled to maintain a process pressure of 1.33 Pa (10.0 mTorr) as measured by a capacitance manometer. Deposition rates were calibrated prior to multilayer growth using a DEKTAKTM surface profilometer and scanning electron microscope employing focused ion sectioning. Films made for TDTR experiments consisted of an 800 nm thick multilayered volume and an 80 nm Al capping layer. Demonstrated in Fig. 3.12, the multilayered portion is composed of alternating Al and Pt layers wherein the period of a given sample is kept constant. The period, also referred to as bilayer thickness, was changed in successive depositions. Thus, it is an experimental design variable. Al and Pt layer thicknesses were chosen to establish an overall equimolar stoichiometry within each multilayer volume. This stoichiometry was achieved by setting the Al:Pt thickness ratio equal to 1.09, which accounts for the different densities of Al (5.84×10^{22} at./cm^3) and Pt (6.41×10^{22} at./cm^3). The densities of single constituent films were determined separately using X-ray Reflectivity (XRR) by employing a Scintag PAD-X diffractometer equipped with a sealed tube source (Cu K-α radiation), an incident beam mirror optic, and a Peltier-cooled Ge solid state detector. Reflectivity scans were collected over typical 2θ ranges from 0.2 to 3^o with a count time of 1 to 4 s. Scattering density, thickness, and roughness were modeled using Parratt software for fitting. All multilayer depositions started with an Al layer and ended with Pt.

The same cross-sectioned samples were evaluated for structure using electron diffraction and high resolution TEM methods. Thick layers were generally textured, polycrystalline metal having expected face-centered cubic (FCC) structures for Al and Pt. Thin layers

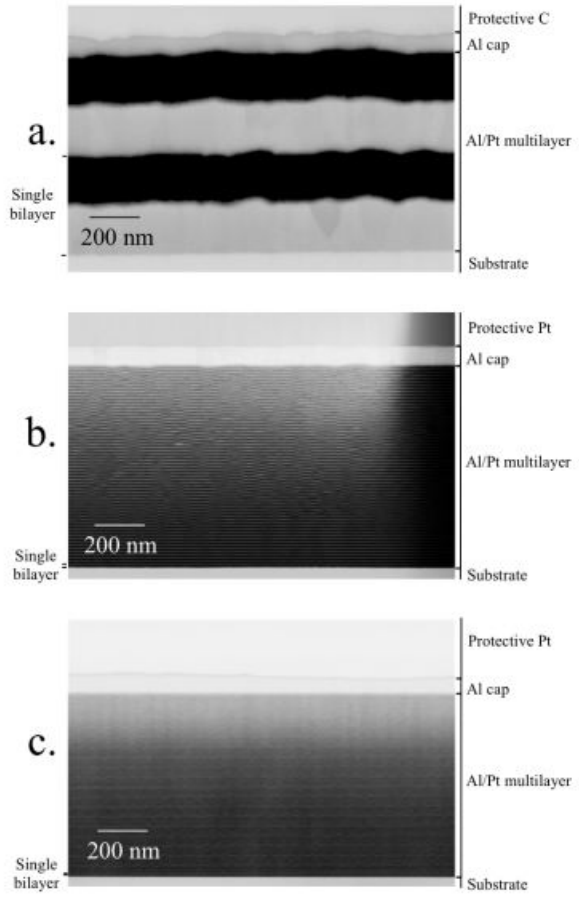

Figure 3.12. (a) Cross-section, bright field electron images of several multilayered samples evaluated by TDTR. Micrographs (a), (b), and (c) show samples with 2, 64, and 96 bilayer periods, respectively. Aluminum layers appear bright, and Pt is dark

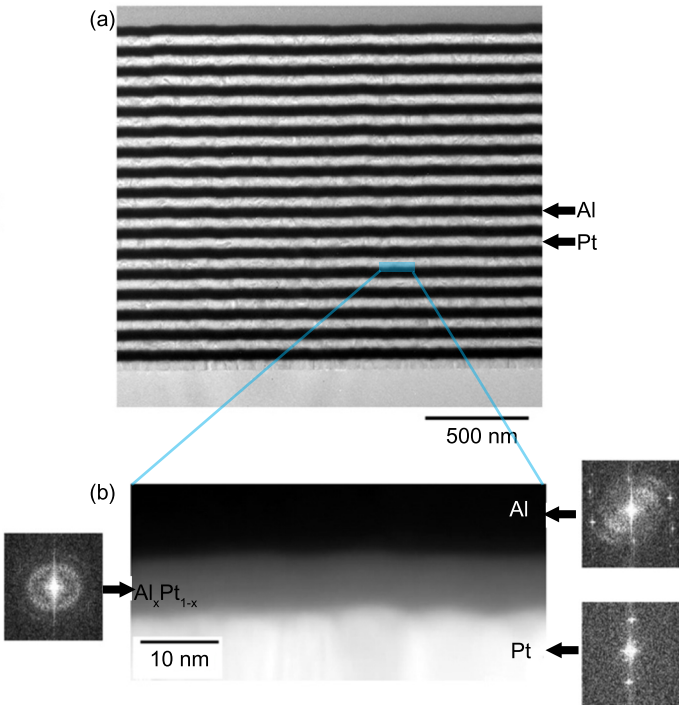

Figure 3.13. (a) Bright-field TEM image of Al/Pt metal multilayer with (b) zoom in of region denoted by box in (a). An approximately 10 nm interphase region exists at each boundary and is amorphous as determined by the selective area electron diffraction images accompanying the TEM.

of amorphous, mixed Al and Pt were generally present at metal-metal interfaces. These interfacial layers [114] are 10 nm thick, on average, although alternating layers varied in dimension between 8 and 12 nm as described elsewhere [117]. These interfacial complexions likely form as a result of interdiffusion at the deposition temperature ($\sim$ 335 K). In contrast, the 96 and 128 period multilayers were mostly amorphous although isolated, nanoscale crystallites were evident by high-resolution TEM. Importantly, isolated, amorphous interfacial layers are not distinguished in multilayers having periods $<$ 10 nm. In all cases, the Al capping layers were polycrystalline.

Energy dispersive spectroscopy (EDS) mapped film composition through the thickness of produced films. Compositional characterization involved positioning a cross-sectioned specimen in the microscope such that two of the silicon drift detectors (SDDs) were aligned to look down at sample interfaces. This geometry helps minimize X-ray absorption by adjacent layers of metal (e.g., absorption of soft Al X-rays by bounding Pt). The Al K-α (1.486 keV) and Pt M-α (2.046 keV) emissions were selected for quantitative study [118]. A Cliff-Lorimer k-factor [119] of 4.56 was obtained from separate, blended, AlPt thin film standards fabricated by co-deposition methods (described below). These standards, of known composition, were made to thicknesses that were similar to other cross-sectioned specimens. Thus, the thickness-resolved k-factor should account for X-ray absorption effects in addition to differences in ionization threshold and X-ray production.

EDS showed that films were composed of Al and Pt. There was no evidence of impurities consistent with previous depth-profiling Auger electron spectroscopic investigations.S5 In addition, multilayers made of relatively few periods ($<$ 64) were comprised of pure Al and pure Pt layers along with the aforementioned 10 nm-thick amorphous complexion. EDS also indicated that the 64- and 128-period films were more blended although a nanometer-scale compositional modulation was observed. As demonstrated in Fig. 3.14, these particular multilayers are comprised of alternating Pt-rich material and mixed Al_xPt_{1-x} (with x $>$ 0.5).

The distinct 10 nm thick nanolayer complexion exhibits a compositional variation through their thickness. Applying the aforementioned k-factor, EDS reveals a range of composition from 50 to 80% Al as shown in upper right-hand plot of Fig. 3.14. Outside of this compositional range the intermixed material retains the FCC structure.

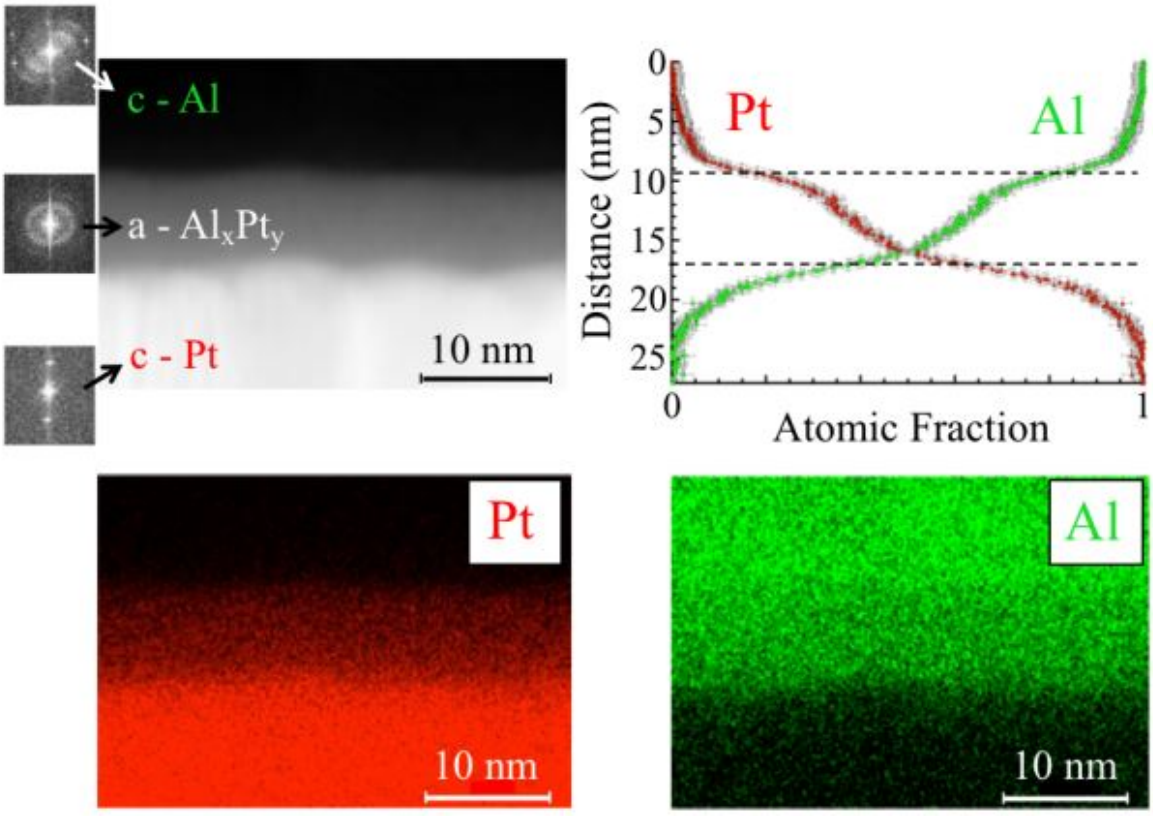

Figure 3.14. Cross section STEM image and composition maps (Pt, Al) of a mixed interfacial volume and surrounding material. Also, included in upper right is a plot of composition derived from the maps shown below. A thickness-resolved k-factor = 4.56 (derived from calibration standards) was used for EDS analysis. Letter "c" and "a" refer to crystalline and amorphous, respectively. The included diffraction patterns are Fast Fourier Transforms (FFT) of complementary high-resolution TEM images

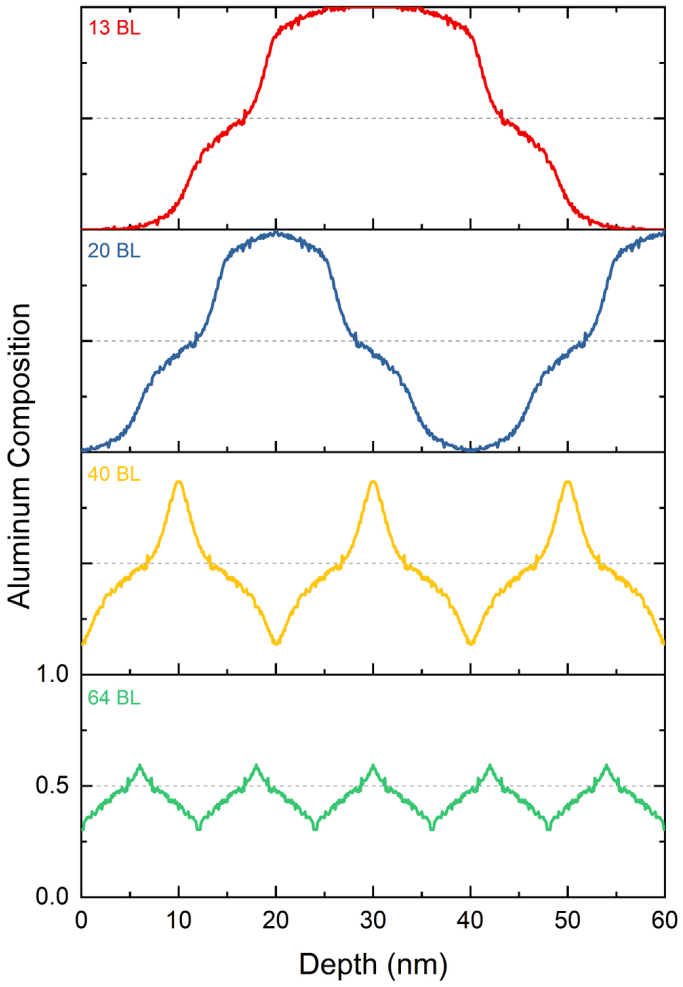

Figure 3.15. EDS derived compositional maps of Al/Pt multilayers possessing varying interfacial density. The composition range decreases with increasing interfacial density.

Additional 800 nm-thick Al_xPt_{1-x} thin films were fabricated by co-deposition methods to create blended volumes of Al and Pt without compositional modulation or distinct interfaces. Two sets of such films were synthesized. First, 800 nm-thick homogeneous Al_xPt_{1-x} films were made to different compositions and then capped with 80 nm of Al for subsequent TDTR characterization. Second, separate blended samples were made without an Al cap for use as EDS standards (described above) and for independent confirmation of the compositional range associated with amorphous Al_xPt_{1-x}. Blended films were made by simultaneously co-depositing Al and Pt from individual sources using a turbopumped Kurt J. Lesker PVD 200 sputter system. The composition was tailored by independent control of two, dedicated, DC magnetron power supplies. This high vacuum apparatus maintained a base pressure of 9.3×10^6 Pa (7×10^{-8} Torr) thus enabling pure film growth. Ultra-high purity Ar was again utilized at 1.33 Pa (10.0 mTorr) for sputtering. X-ray diffraction (XRD) determined the structure of co-deposited films (without caps) as reported in Table 3.1. XRD involved a Bruker D2 Phaser diffractometer equipped with Cu K-α radiation (30 kV, 10 mA) and a LynxEye silicon strip detector. Crystalline phases were identified using archived diffraction patterns of known Pt-Al intermetallics available through the International Center for Diffraction Data (ICDD) [120]. For example, the crystalline film described in Fig. 3.16 exhibits sharp diffraction peaks that index well to the known reflections of Pt_3Al. Polycrystalline diffractograms are distinctly different than those generated from amorphous films as the latter are characterized by broad, diffuse, diffraction peaks. The composition of each co-deposited AlxPt1-x film (without Al cap) was determined by Wavelength Dispersive Spectroscopy (WDS). A JEOL Inc. JXA-8530F field emission electron probe microanalyzer utilized a 7 keV, 20 nA spot beam for characterization. Again, the Al K-α and Pt M-α X-ray emissions were selected for analysis, thus avoiding spectral interference. A ZAF correction was applied to account for atomic number differences (Z), effects of absorption (A), and fluorescence (F). Taylor metallurgical standards enabled quantification. Measurement uncertainties are estimated to be $\pm$ 2 at.%.

Due to the lower potential energy of the fully mixed phase compared to the segregated multilayer, the Al and Pt layers spontaneously intermix to form a compositionally graded region approximately 10 nm thick at each boundary between them. This intermixed region is

Table 3.1. Summary of blended films used to map the compositional range of amorphous Al_xPt_{1-x}. Monolithic Al and monolithic Pt films are also included. Specified composition was determined by WDS. FCC = face-centered cubic. tr = trace.

Composition	Structure	Crystalline Phase(s)
Al	crystalline	FCC Al
$Al_{0.79}Pt_{0.21}$	amorphous	none
$Al_{0.70}Pt_{0.30}$	amorphous	none
$Al_{0.67}Pt_{0.33}$	amorphous	none
$Al_{0.62}Pt_{0.38}$	am.+cry. (*tr*)	Al_2Pt
$Al_{0.50}Pt_{0.50}$	crystalline	AlPt+Al_2Pt (*tr*)
$Al_{0.4}Pt_{0.6}$	crystalline	$AlPt_3$
$Al_{0.33}Pt_{0.67}$	crystalline	$AlPt_3$
Pt	crystalline	FCC Pt

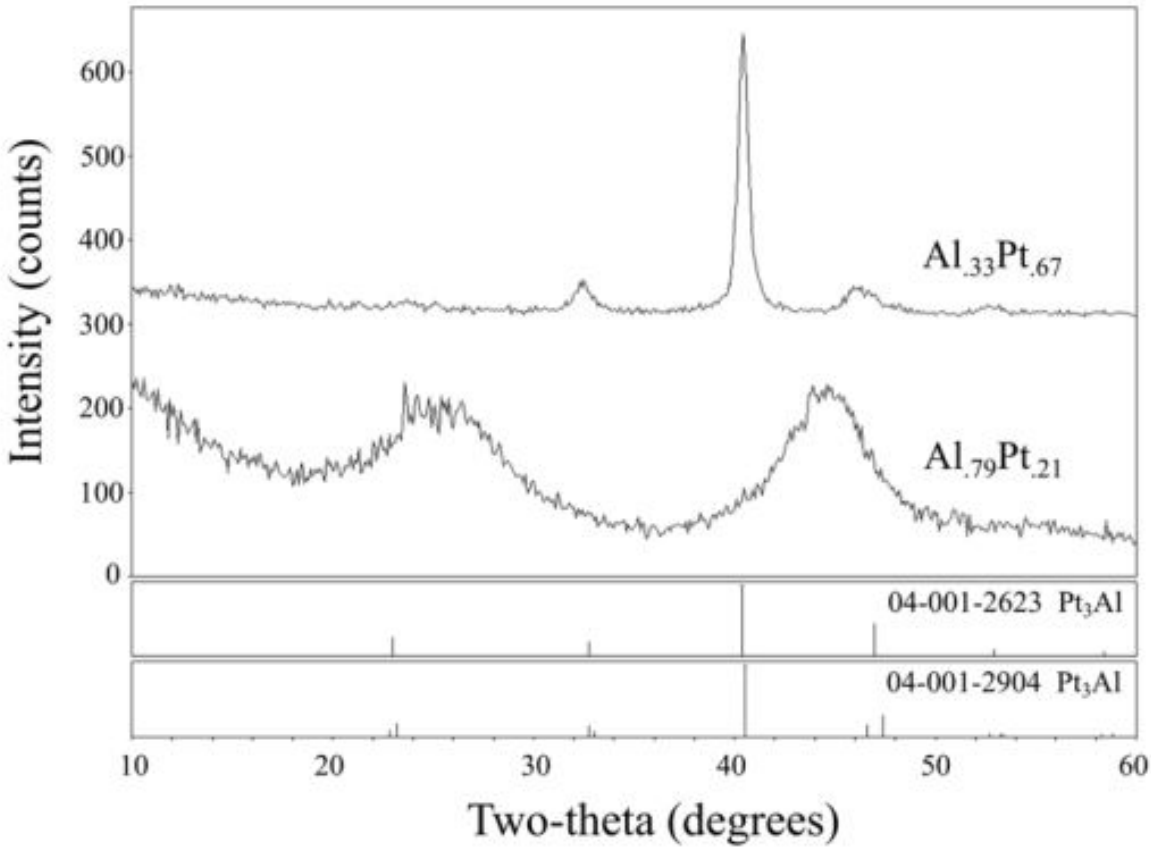

Figure 3.16. X-ray diffractograms obtained from co-deposited amorphous $Al_{0.79}Pt_{0.21}$ and crystalline $Al_{0.33}Pt_{0.67}$ films. Archived patterns[120] included below the graph identify the phase of the crystalline $Al_{0.33}Pt_{0.67}$ film (top diffractogram)

a complexion, which describes a finite layer forming at the boundary between two bulk phases [15]. To further specify, the Al/Pt system forms a nanolayer complexion under the Dillon-Harmer taxonomy due to its near constant 10 nm thickness irrespective of the bilayer period [27]. The complexion thickness is comparable to other bimetallic systems [15]. Hereafter, the terms complexion and interlayer will be used synonymously. Regardless of nomenclature, this interlayer, or complexion, is amorphous as is observed through electron diffraction and emerges when the minority elemental species exceeds $\sim 10\%$. Having a constant interlayer thickness, multilayers with periods below 20 nm are therefore effectively composed of only the complexion.

Energy dispersive x-ray spectroscopy (EDS) composition traces, meanwhile, indicate that not only is the crystallinity of the multilayer impacted by the periodicity but so too the composition of the films. With decreasing bilayer thickness, the composition no longer possesses regions of the pure metal layers but is instead an alloy of Al_xPt_{1x} that varies in its composition as is seen in Fig. 3.15. Due to this observation, monolithic Al_xPt_{1x} alloys were also synthesized via sputtering codeposition to thicknesses of ~ 800 nm. These monolithic films were pursued to both validate the computational approaches described in the proceeding sections and assess the intrinsic thermal resistance of the interlayer apart from the boundaries. All films, both those monolithic and multilayer, were capped with an $\sim$ 80 nm Al layer—denoted subsequently as the cap—to facilitate thermal characterization as is shown in Fig. 3.13.

3.3.2 Thermal measurements

Thermal conductivity measurements were performed using two color time domain thermoreflectance (TDTR) as described elsewhere [121], [122]. The ratio of the in-phase to out-of-phase response was fit to a three-layer model where the multilayer was assumed to be an effective medium [123]. The Al-cap/multilayer boundary conductance and thermal conductivity of the effective medium representing the multilayer were left as free parameters and fit to the acquired data. The total resistance of the multilayer film was then taken to be the sum of the Al to multilayer boundary resistance and that of the effective medium.

Pump and probe spot diameters (80 and 12 μm) were chosen to ensure sensitivity to only cross-plane transport. Laser power densities (30 and 20 mW) ensured DC heating below 10 K. A total of five TDTR scans were collected at differing sample locations using a modulation frequency of 3.37 MHz. A subset of samples was also analyzed using a modulation frequency of 11.3 MHz. Raw data was phase corrected following the procedure as described in Ref. [124] and analyzed from a delay time of 200 ps to 4000 or 6000 ps depending on signal to noise. TDTR traces were fit to a three-layer model where the multilayers were assumed homogeneous with an effective specific heat of 2.83 J/cm^{3}K. The model was insensitive to the back boundary conductance between the multilayer and silicon, GML/Si, and the silicon substrate thermal conductivity, κ_{Si}. These parameters were assumed to be $G_{ML/Si} = 150$ MW/m^{2}K and $\kappa_{Si} = 140$ W/mK. Therefore, the effective ML thermal conductivity, $\kappa_{ML,eff}$,and front boundary conductance between the Al transducer and multilayer, $G_{Al/ML}$, were left as free parameters. From the fits of these parameters, the total resistance of the multilayer was determined via:

$$R_{Tot} = \frac{1}{G_{Al/ML}} + \frac{d}{k_{ML,eff}} \tag{3.8}$$

where, d is the total ML thickness. Material layer thicknesses were determined from TEM cross-sections. Uncertainties were determined from the vector summation of measurement standard deviations, fit uncertainties calculated from the Jacobian and fit residuals, and uncertainties from constant model parameters (e.g., specific heats, thicknesses, laser spot sizes, etc.). Finally, thermal conductivity of the pure Al-film shown in **Figure. 3** was deduced from a 4-point resistivity measurement that was transformed to thermal conductivity by employing the Wiedmann-Franz law.

Results are tabulated in Table 3.2, where it is seen that thermal resistance does not monotonically increase with the number of periods. The causes for this nonmonotonic trend will be discussed in detail via modeling of the system as is described subsequently.

Table 3.2. Characteristics of examined Al/Pt multilayers accompanied by measured and NEGF+DFT predicted thermal resistances.

Number of periods	Period thickness (nm)	Exp. resistance (m^2K/MW)	NEGF resistance (m^2K/MW)
1	800	0.024	
2	400	0.024	
4	200	0.038	0.019
8	100	0.050	0.030
16	50	0.11	0.06
32	25	0.15	0.11
48	16.7	0.16	
64	12.5	0.12	0.10
96	8.3	0.14	0.103
128	6.25	0.10	0.102

3.3.3 Modeling and Simulation

To provide insight into the physical mechanisms behind the experimental observations, thermal resistance of the multilayers was predicted under two different computational paradigms. First, the traditional electron diffuse mismatch model (eDMM) was employed where boundaries are presumed abrupt and interfaces perfect with energy transmission determined based on density of states overlap between the two materials making up the boundary. [123] As will be described below, this simple model fails to capture the experimental observations, even for long-period superlattices. This is not surprising given the complexity of the structures involved. Thus, we developed a first principles-based multiscale approach to describe thermal transport in our multilayers capable of explicitly describing the interphase region and the contact resistance between the various phases. We use MD simulations to predict atomistic structures of the Al_xPt_{1-x} alloys of various compositions observed in the interface and to capture thermal ionic fluctuations. Thermal transport in these structures was then modeled using NEGF within the Landauer approximation with electronic structure obtained from DFT calculations.

In both the eDMM and the first principles calculations, the resistance of the interfaces were quantified and then the total thermal resistance of the multilayer film (R_{Tot}) determined using a series resistance model given by [125]:

$$R_{Tot} = R_0 + nR_{BD}, \tag{3.9}$$

where R_0 is the resistance of the constituent material layers ($R_o = R_{Pt} + R_{Al}$), n is the number of interfaces, and R_{BD} is the resistance of the interphase region. The resistance of Al and Pt was determined using separate NEGF+DFT simulations of pure crystalline layers using methods previously employed. The methodology to determine the interphase resistance, meanwhile, differed based upon the technique employed as is described below.

eDMM

The eDMM is based on free electron theory and can be written as [81]:

$$\frac{1}{R_{BD,eDMM}} = \frac{1}{4}\zeta_{Al\to Pt} C_{e,Al} v_{F,Al}, \tag{3.10}$$

where $C_{e,Al}$ is the electronic heat capacity, v_F is the Fermi velocity. The specific heat is given by:

$$C_{e,Al} = \frac{\pi^2}{3} D(\epsilon_{F,Al}) k_B^2 T, \tag{3.11}$$

where k_B is Boltzmann's constant, T is temperature, and $D(e_{F,Al})$ is the density of states at the Fermi energy. The transmission coefficient is:

$$\zeta_{Al\to Pt} = \frac{D(\epsilon_{F,Pt}) v_{F,Pt}}{D(\epsilon_{F,Pt}) v_{F,Pt} + D(\epsilon_{F,Al}) v_{F,Al}}, \tag{3.12}$$

where $Al \to Pt$ denotes transport from Al to Pt. Since the interface is an alloyed metal, thermal transport was assumed to be dominated by electron transport as is implied by this approach. The resulting Kapitza resistance derived by the eDMM was calculated to be 0.23 m^2K/GW. The material parameters used for this calculation are presented in Table 3.3

Table 3.3. Aluminum and platinum material properties used in the eDMM to calculate Kapitza conductance.

	Al	Pt
$D(\epsilon_F)$ (m^{-3})	1.26×10^{47}	7.05×10^{47}
v_F (m/s)	1.3×10^{6}	0.3×10^{6}

NEGF

The eDMM assumes pristine, atomistically sharp, interfaces. However, there is significant intermixing at the interfaces of the multilayers studied here. As shown in Fig. 3.13 and 3.15, this region has both finite thickness and variable composition. Therefore, a more

detailed description of the electronic transport through the interphase and the associated contacts must be considered. Given the smooth variation in composition observed experimentally, our approach involves computing the thermal resistivities of Al_xPt_{1-x} alloys of various compositions from first principles and combine them in series to predict thermal transport in the alloy. Importantly, the simulations provide information about the contact resistance involved in transitioning between Al or Pt and the interphase region; this additional resistance is added to the model. To capture the complexity of the structures we use MD simulations to melt and quench the systems of interest and obtain relaxed structures for thermal transport calculations from a simulation at room temperature to capture thermal ionic vibrations. The structures are designed for transport calculations and contain perfect Al and Pt contacts separated by the alloy of the desired composition, see Fig. 3.17. These structures are then used within the NEGF/Landauer transport formalism using electronic structures obtained from DFT calculations (using the generalized gradient approximation of Perdew [126]. For each composition we characterize transport in samples obtained from an MD simulation at 300 K to capture elastic phonon-electron coupling and of different channel length from which we can extract thermal resistivity [127].

The atomistic structures for transport contain and Al lead and a Pt lead separated by a Al_xPt_{1-x} channel. The leads consist of perfect FCC crystals oriented along [100], replicated 2x2x2 with an averaged lattice parameter between Al and Pt of 3.98 Å. The averaged lattice parameter was used to provide minimal strain on the contacts between the leads and channel. Varied compositions of amorphous devices were created by extending a channel of 50/50 Al/Pt and randomly swapping Al or Pt atoms.

The leads are separated by a channel with the desired alloy. We characterized Al_xPt_{1-x} alloys of seven different compositions and channel lengths varying from 1.4 to 3.0 nm. The channel was melted at T=3000 K via isochoric isothermal MD simulations and then quenched to room temperature in 10 ps while maintaining the leads fixed. All MD simulations were carried out with LAMMPS [40] using an effective medium theory (EMT) potential developed by Jacobsen et al. [128]. A representative atomic structure is shown in Fig. 3.17.

An effective medium theory (EMT) potential for Al/Pt, parameterized by Jacobsen et al. [128] and accessed through the OpenKIM potential repository [129], [130], was used

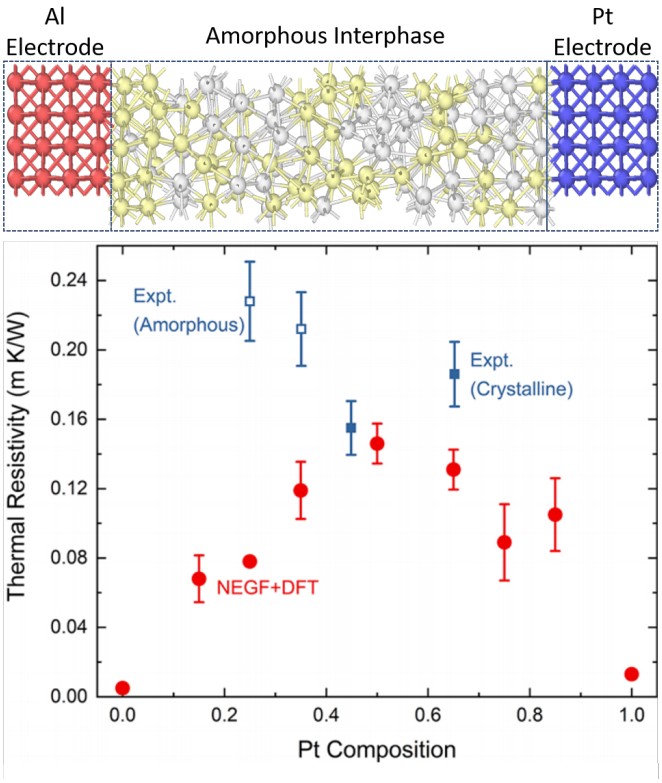

Figure 3.17. (top) Representative atomic structure of simulation domain realized through molecular dynamics. (bottom) Comparison of computationally derived thermal resistance of amorphous Al_xPt_{1-x} alloy compared to experimental results of monolithic films measured using TDTR.

in our simulations for relaxation and amorphous structure generation. All MD simulations were performed using the LAMMPS software package from Sandia National Laboratories [40]. Visualization of atomic structures and structure identification was performed with OVITO [85]. A timestep of 1 fs was used throughout, with damping constants of 0.1 ps for the Nose-Hoover thermostat.

The electronic structure required for the NEGF calculations were obtained from MD-derived atomic structure using DFT. All electronic transport computations were carried out using the TranSIESTA code [131], implemented in the SIESTA package [132]–[134]. The core electrons are replaced by pseudopotentials, and the valence electrons are represented using numerical atomic-like basis set. The exchange-correlation potential is calculated within the generalized gradient approximation [135] functional. Double zeta plus polarization (DZP) numerical orbital basis sets were used for all atomic species.

The self-consistent field calculation was terminated when an energy tolerance of 10^4 eV is reached. k-mesh of 16×16×50 was used for the computation of the leads self-energy, 16×16×3 for the device self-energy, while the transmission spectra calculations was carried out on a 30×30×1 k-grid to ensure converged values.

The DFT-derived electronic structure was employed using NEGF within the Landauer approximation to deduce resistance to thermal transport by simulating electrons' movement from a crystalline aluminum contact through the interphase and to a Pt contact. By simulating a series of films possessing identical compositions but of varying length, the thermal resistance was deduced through the slope of resistance versus film thickness. [127] The intercept of this linear fit, meanwhile, provides an estimate of the combined Al/interphase and interphase/Pt Kapitza resistance, which is also dependent upon composition.

As discussed above, the simulations of the two-probe devices consist of a left-lead (L), a central-region (C), and a right-lead (R), where the self-consistent Kohn-Sham potentials and Hamiltonian matrices of the leads come from a separate, fully periodic DFT calculation. The retarded Green's function is only calculated in the central region, which is treated as an open system, using the Hamiltonian of the central region (H_C) and the lead self-energies (Σ_L and Σ_R) according to:

$$G^{ret} = [(\varepsilon + \mathrm{i}\eta_+)S - H - \Sigma_L(\varepsilon) - \Sigma_R(\varepsilon)]^{-1}, \tag{3.13}$$

where S is the the overlap matrix and η_+ is an infinitesimal positive number. The left and right density matrix contributions are calculated as:

$$D_{L,R} = \int \rho_{L,R}(\varepsilon) n_F(\varepsilon - \mu_{L,R}) d\varepsilon, \tag{3.14}$$

where the n_F is the Fermi function, $\mu_{L,R}$ the electrochemical potential of the left or right lead, and the spectral function $\rho(\varepsilon)$ is given by:

$$\rho_{L,R}(\varepsilon) = \frac{1}{2\pi} G^{ret}(\varepsilon) \Gamma_{L,R}(\varepsilon) G^{adv}(\varepsilon). \tag{3.15}$$

In Equation 3.15, $G^{adv}(\varepsilon)$ is the advanced Greens function defined as $G^{adv}(\varepsilon) = (G^{ret}(\varepsilon))^\dagger$. And $\Gamma_{L,R}(\varepsilon)$ is known as the broadening function of the left and right lead given by:

$$\Gamma_{L,R}(\varepsilon) = \frac{1}{\mathrm{i}}(\Sigma_{L,R} - (\Sigma_{L,R})^\dagger). \tag{3.16}$$

The electron density of the central region is calculated from the total density matrix $D = D_L + D_R$ as:

$$\mathcal{N}(\mathbf{r}) = \sum_{\mathrm{ij}} D_{\mathrm{ij}} \phi_{\mathrm{i}}(\mathbf{r}) \phi_{\mathrm{j}}(\mathbf{r}). \tag{3.17}$$

where ϕ are numerical basis orbitals. And the total transmission is computed as:

$$T(\varepsilon) = G^{ret}(\varepsilon) \Gamma_L(\varepsilon) G^{adv}(\varepsilon) \Gamma_R(\varepsilon). \tag{3.18}$$

The resistance is calculated according to:

$$G = \frac{1}{R} = G_0 \int T(\varepsilon) \left(-\frac{\partial f}{\partial \varepsilon} \right) d\varepsilon, \tag{3.19}$$

where G_0 is the quantum of conductance given as $2e^2/h$ with e as the electron charge and h as the Planck's constant, and $\partial f/\partial \varepsilon$ is the derivative of the Fermi function (Fermi window) at a certain temperature.

The electronic contribution to the thermal conductance is given by:

$$\kappa_{el} = \frac{1}{R_{el,th}} = \frac{2}{hT}\left(K_2 - \frac{K_1^2}{K_0}\right), \tag{3.20}$$

where K_n is defined as:

$$K_n = \int (\varepsilon - \varepsilon_F)^n T(\varepsilon)\left(-\frac{\partial f}{\partial \varepsilon}\right) d\varepsilon. \tag{3.21}$$

Fig. 3.17 shows the calculated resistivities of the Al_xPt_{1-x} alloys and a representative device of a Al_xPt_{1-x} alloy at 50/50 composition. End points in the graph are fully crystalline Al and Pt devices used as validation and were calculated at 185 W m^{-1} K^{-1} and 76 W m^{-1} K^{-1} for Al and Pt respectively. Thermal resistance of Al_xPt_{1-x} alloys calculated via the NEGF+DFT approach compare well to experimental measurements of monolithic films as seen in Fig. 3.17 lending further credence to the methodology employed. The correlation to within a factor of three is notable considering the differences in that simulated versus that measured. The model, for instance, does not account for any impurities that may exist in the actual films. The co-deposited films, meanwhile, begin to crystallize at higher Pt compositions as was confirmed using XRD whereas the simulation cell maintains an overwhelmingly amorphous structure similar to that observed within the interphase region of the multilayers. Taken together, the computational approach therefore provides a reasonable description of transport through the interphase. Specifically, it accounts for its compositional dependence as is seen by the "bathtub" shape of Fig. 3.17.

These compositionally dependent thermal resistances were then used to deduce the total interphase resistance ($R_{BD,NEGF+DFT}$) in a manner that included both the variation in composition and the finite width of the intermixed region. This was realized by first discretizing the compositional maps shown in Fig. 3.15 in 0.1 nm steps. Using the average composition within a given step, the resistance was calculated for this discretized region using values supplied from the simulations of Fig. 3.17. The total interphase resistance was then realized via a series sum of the discrete elements while also including the Kapitza

resistance existing between the metallic end members and the interphase. Mathematically, this is described as shown in Eq. 3.22 below,

$$R_{BD,NEGF+DFT} = \sum_{\mathrm{i}}^{n} \frac{d_{\mathrm{i}}}{\kappa(x_{\mathrm{i}})} + R_{Kapitza} \tag{3.22}$$

where, d and $\kappa(x_{\mathrm{i}})$ are the thickness and composition dependent thermal conductivity of the i^{th} element, respectively, and $R_{Kapitza}$ is the total Kapitza conductance accounting for boundaries between the interphase and both Al and Pt. Quantitatively, a compositional average of 0.28 $\mathrm{m^2K/GW}$ derived from the NEGF+DFT was employed for this value.

3.3.4 Results

Bulk electronic thermal transport values were calculated from the length dependent thermal conductance of each material according to the frozen phonon approach to electron-phonon interactions [127]. This method takes the Born-Oppenheimer approximation for relative movement of phonons to electrons. We assume that by taking multiple snapshots of a phonon population and averaging their effects we capture the collective scattering effects from electron-phonon interactions. We take an average of the transmission spectra calculated at each frozen phonon snapshot for Al in Fig. 3.18. N=5 was sufficient for averaging and capturing noise around the Fermi window. We calculate the resistance of the device by integrating the transmission function with the Fermi smearing function, and adjusting for the modes of thermal conductance shown in Fig. 3.19.

From the slope of the resistance vs. length we calculated bulk thermal resistivities for Al and Pt.

The same method was applied to varied compositions of amorphous Al/Pt devices. However, due to the inherent randomness of the amorphous structure an increased number of samples are required for statistical averaging. For each amorphous sample N=10 frames were used for transmission spectra averaging.

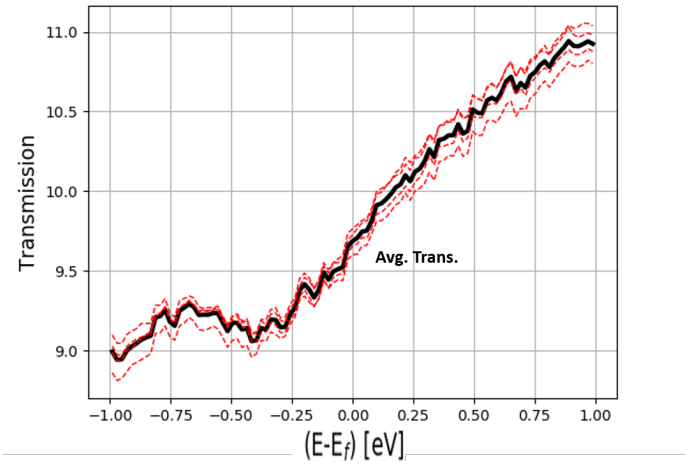

Figure 3.18. Transmission spectra of pure crystalline Al device. Individual transmission spectra shown as red dashed lines, and averaged spectra as solid black. N = 5 samples used for averaging.

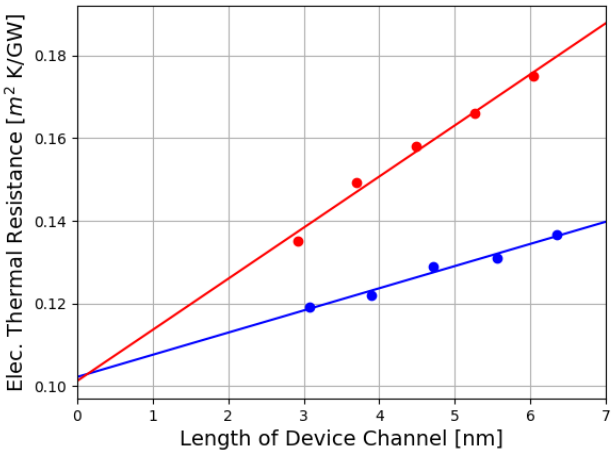

Figure 3.19. Length dependent resistance for Al (blue) and Pt (red) at 300 K. Calculated thermal conductivity of Al = 185 W m^{-1} K^{-1} and Pt = 76 W m^{-1} K^{-1}.

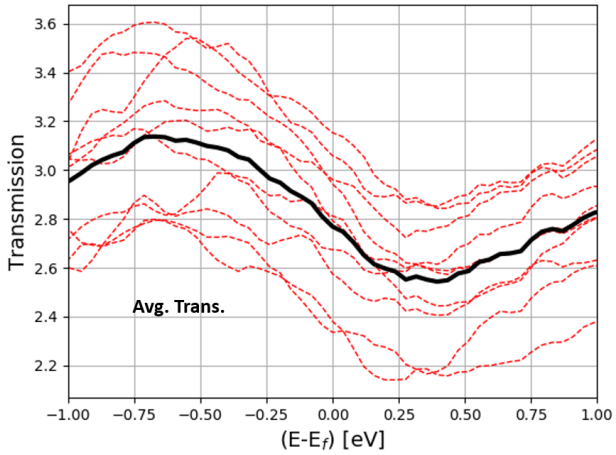

Figure 3.20. Transmission spectra of amorphous 50/50 Al/Pt device. Individual transmission spectra shown as red dashed lines, and averaged spectra as solid black. N=10 samples used for averaging.

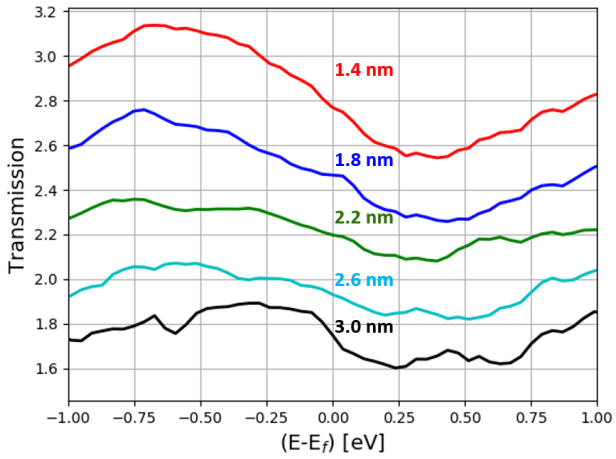

Figure 3.21. Length dependent transmission spectra of amorphous 50/50 Al/Pt device. Averaged transmission for each length shown.

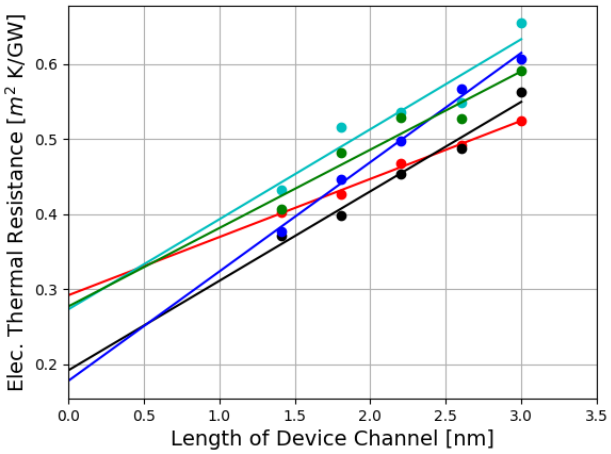

Figure 3.22. Length and temperature dependent resistance for varied composition of amorphous Al/Pt at 300 K. 25% Pt = red, 35% Pt = black, 50% Pt = blue, 65% Pt = cyan, 75% Pt = green. Only 25-75% Pt channels shown for clarity.

As mentioned in the main text, we make use of the contact resistance obtained from the length dependent resistance calculation as a lumped parameter to account for crystalline Al/Pt to amorphous transitions, as well as gradient effects across the device.

The contact resistance is the cost of injecting electrons into the device from the pure crystalline left lead, to the channel, and the right lead. We take the contact resistance as the y-intercept in Fig. 3.22.

We note that while the contact resistance is compositionally dependent, when modeling across the amorphous interphase we take an averaged value of 0.28 m^2K/GW as shown in Fig. 3.23

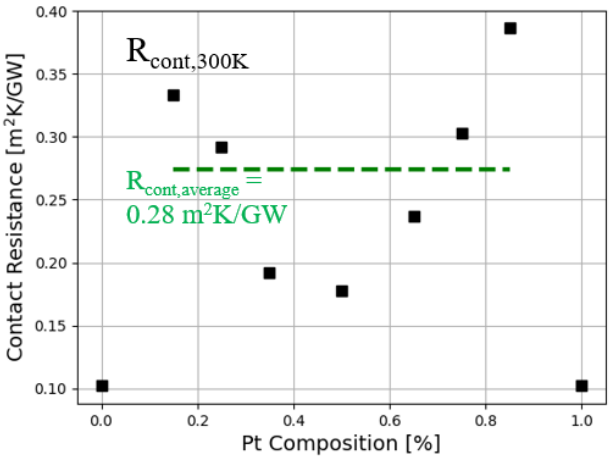

Figure 3.23. Compositional dependence of contact resistance. End points are fully crystalline Al/Pt devices with identical atom type leads.

As shown with experimental results of co-deposited Al/Pt films, Al-rich films tend to stay amorphous, while Pt-rich films tend to order and recrystallize.

We first used the polyhedral template matching algorithm [86] implemented in Ovito [85]. The method sorts basic crystal structures based on local atomic environments. In Fig. 3.24 we sort by FCC (green atoms) and 'Other' (white atoms) which we take as amorphous. The content of FCC vs. amorphous material should not be taken as an absolute value. The algorithm has a tendency to over characterize local crystallinity, and even shows high degrees of FCC content in systems being run at 3000 K. However, even with the over characterization

of FCC content in the system, it is clear to the eye that there is a larger degree of plane stacking and ordering in the Pt-rich MD simulations. This is encouraging since it aligns with experiments, but we further verified the amorphous nature of our MD devices using a radial distribution function (RDF) calculation.

In Fig. 3.25 we show the RDF for each composition during the initial configuration, melted, and quenched phases. We clearly show that in the melted phase we no longer have any long range order, and we maintain this broadened second peak across all compositions even after the quench. However, we do see a small peak forming with the Pt-rich compositions which is consistent with both experimental results and the PTM analysis of our structures.

Each composition profile was analyzed during the melt and quench stages of processing, and FCC RDFs were overlayed for peak identification and clarity. FCC RDFs were generated from an MD simulation of Al at 300 K using the Jacobsen potential [128]. In the Al-rich devices we see consistent peak broadening in the melted phase, followed by little peak heightening near the long range FCC peaks at ∼5 Å. For Pt-rich systems we see consistent peak broadening in the melted phase, but we do some some peak definition around FCC peaks. However, the degree to which these peaks remain smoothed is encouraging for the stability of our amorphous samples.

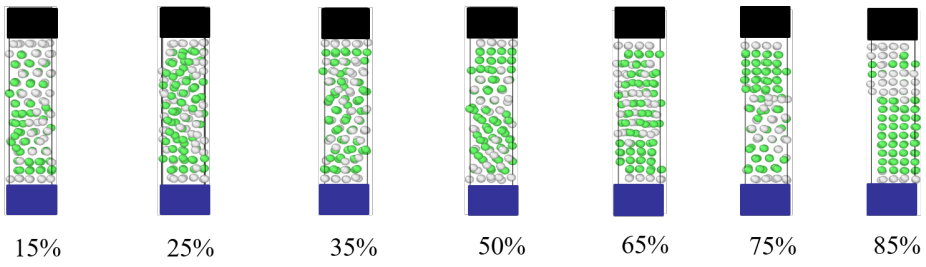

Figure 3.24. Atomic visualizations of quenched Al/Pt alloy devices ranging from 15-85% Pt content. PTM algorithm used with green atoms at FCC type structure, and white atoms as amorphous 'Other'.

Fig. 3.27 (a) plots the total thermal resistance of the multilayers as a function of the number of interfaces within the multilayer. The resistance initially increases linearly at low interface density whereupon it rolls over after reaching a maximum resistance at about 97 interfaces, corresponding to a bilayer thickness of 16.7 nm. Subsequently, the resistance

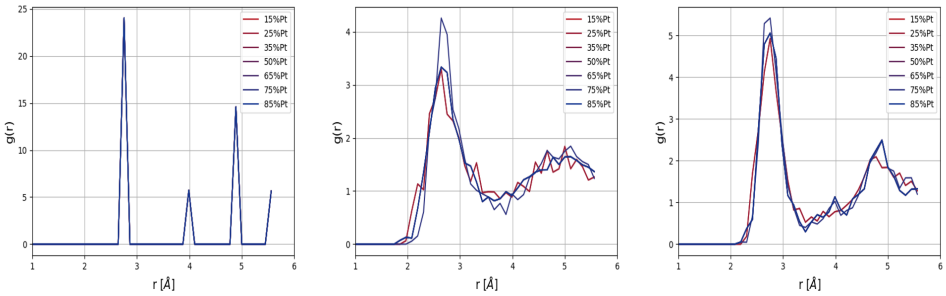

Figure 3.25. RDF calculations of initial configurations (left), melted structures at 3000 K (middle), and quenched structures (right) for all compositions simulated.

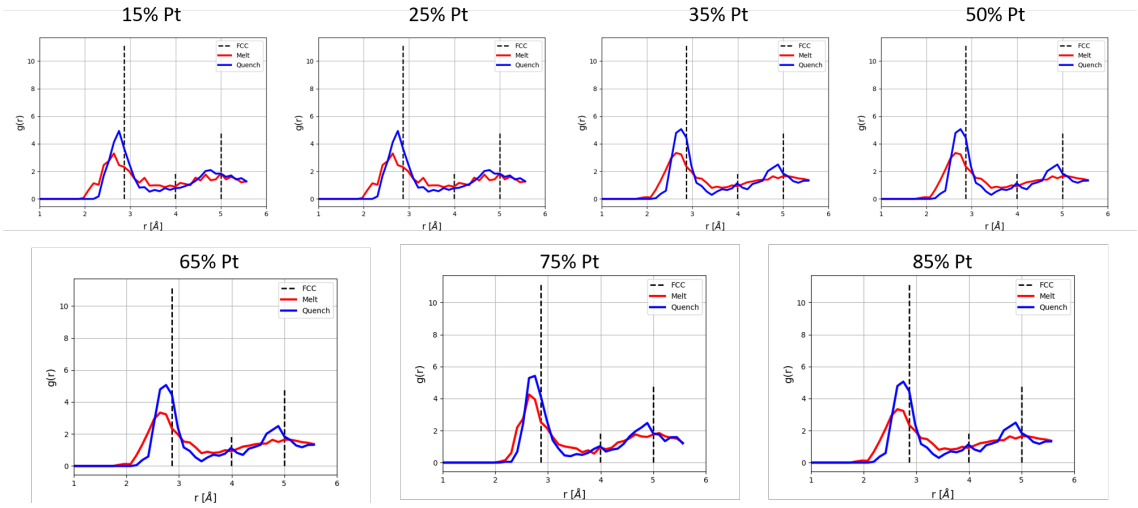

Figure 3.26. RDF calculations of individual compositions of Al/Pt alloys based on Pt%. FCC RDFs were gathered from MD simulations on pure Al and as full crystalline comparison.

decreases to a value that is on par with the resistance of the monolithic alloyed films whose resistances were shown in Fig. 3.17. While the rollover in thermal resistance is qualitatively similar to that seen in phonon dominated systems previously, [136] its causation is of wholly different character. Here, the rollover in thermal resistance originates due to a transition from interface dominated scattering to alloy-mediated point defect scattering. The following provides evidence supporting this assertion.

Predictions of thermal interface resistance of electron dominated systems have overwhelmingly used the eDMM, which presumes a perfect interface resulting in a monotonically

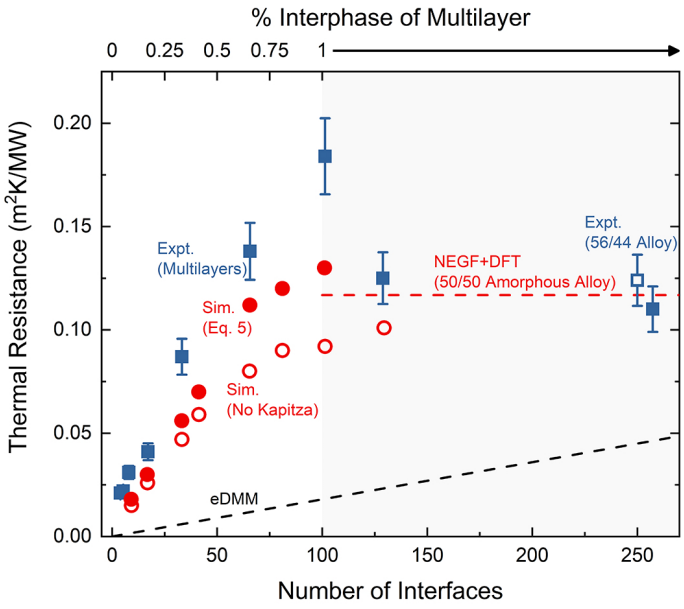

Figure 3.27. Thermal resistance of Al/Pt metal multilayers as measured by TDTR and simulated using both eDMM and NEGF+DFT approaches. Measured resistance reaches a peak at 97 interfaces and then decreases. Unlike the eDMM which assumes a perfect interface, the NEGF+DFT approach captures this behavior by accounting for both the finite width of the interphase region and its compositional dependence. Note that with increasing interface density the multilayer film approaches the values measured and modeled for an amorphous alloy of approximately 50/50 composition.

increasing trend in total resistance stemming from Eq. 3.9. As seen from Fig. 3.27, this approach is contrary to observation both qualitatively and quantitatively despite estimating a Kapitza resistance that is within a factor of 2 of that derived using the NEGF+DFT approach. The NEGF+DFT approach, in contrast to the eDMM, does predict the observed non-linear trend in thermal resistance while also comparing quite well to experiment. Since the Kapitza resistances are quite similar between the techniques, the discrepancy points to the the centrality of the disordered amorphous interphase on thermal transport for which only the NEGF+DFT approach "sees."

Examining alterations in the interphase with interface density therefore provides a means to understand the causation of the non-monotonic thermal resistance trend. Mechanistically, the interphase changes with interface density—thereby modifying the thermal transport—owing to the rather constant ($\sim$ 10 nm) thickness of intermixing implicit with Al/Pt synthesis. This is in contrast to the the thickness of Al and Pt layers themselves, which continually decrease in thickness with higher interface density. Quantitatively, the finite thickness of intermixing means that that the volume percentage of the interphase within the multilayer increases with the number of interfaces as is highlighted in the upper axis of Fig. 3.27.

The evolution in the relative amount of interphase with increasing interface density impact not only the structural character (crystalline to amorphous) of the multilayers but their composition profile as well. This is shown definitively by examining the composition of the multilayers with interface number (see Fig. 3.15). With increasing interface density, the percentage of "pure" metal continually drops eventually resulting in full interdiffusion where the alloy makes up the entirety of the film. Quantitatively, this occurs when the period thickness becomes less than twice the interphase thickness—about 17 nm at 97 interphases—where pure Al is first consumed owing to its more rapid diffusion.

The interphase therefore induces two major alterations in the multilayer with increasing interface density. First, it changes its elemental composition. As the interphase has a resistance that is heavily dependent upon composition as exhibited in Fig. 3.17, this will necessarily impact thermal resistance. Second, the amount of interphase making up the multilayer increases, which first acts to "soften" the material contrast at the boundaries and

then removes the boundaries altogether as the "multilayer" becomes effectively a $Al_{50}Pt_{50}$ alloy. In the language of thermal transport, the increasing amount of interphase in the multilayer acts to reduce and then eventually remove the Kapitza resistance. [107]

To highlight this fact, NEGF+DFT derived thermal resistances were calculated apart from the contribution of the Kapitza resistance stemming from the boundaries between amorphous alloy and the crystalline Al and Pt (i.e. , $R_{BD,NEGF+DFT}$ in Eq. 3.22). These resistances are represented by the open circles of Fig. 3.27 where the values smoothly approach the resistance of both the modeled and measured monolithic 50/50 alloy. The simulations accounting for the Kapitza resistance (closed circles of Fig. 3.27), meanwhile, exhibit a much steeper trend with interface density while also comparing favorably to the measured results for the region where boundaries persist between the pure metals and the interphase. The differences between the two curves, in turn, are therefore indicative of the reduced role that boundary scattering plays with increasing interface density. We therefore stipulate that the observed rollover in thermal resistance occurs because scattering transforms from being mediated through boundary scattering at the lower interface densities to one in which point-defect scattering implicit in alloys dictates transport. The disorder mediated by the interphase has acted to not just change the structure of the film but so too the character of transport.

At a higher level, this observation also underscores the sometimes counter intuitive role that interfaces play on thermal transport. More is not necessarily better. Increasing the number of interfaces is not inextricably linked to reduced thermal transport even in electron dominated systems. Rather, as the number of interfaces increases, the physical processes belying the movement of energy can change. This change can actually increase the efficiency by which heat is moved. Here, disorder at the interfaces actually removed their ability to limit heat transport even as their number was magnified.

3.3.5 Conclusions

The thermal resistance of metal multilayers does not necessarily scale directly with interface density. This was deduced by observing a non-monotonic trend in thermal resistance

of Al/Pt multilayers as the number of interfaces increased. To understand this reduction in thermal resistance with increasing interface density, structural characterization of the films was combined with simulations leveraging density functional theory in concert with non-equilibrium Green's functions to describe the transport. Taken together, these efforts show that intermixing occurs at the Al/Pt boundary leading to the creation of a $\sim$ 10 nm amorphous "interphase" region at every interface. This interphase dominates transport within the multilayers facilitating a transformation from boundary- to point-defect scattering that acts to reduce thermal resistance. Like that seen in phonon-dominated systems, decreasing the spacing between material boundaries in electron mediated systems does not necessarily reduce the efficiency by which heat is moved.

3.4 Final Remarks

In this chapter we highlighted the role of both DFT and MD simulations in bridging gaps between experimental certainties, and how they can be used in tandem to complement each other for materials simulations. Particularly interesting are the properties that we derive for thermal transport in metallics with atomistic modeling, and the impact of proper assumptions for verification of results. In MD simulations, given that the electrons are implicit and ions interact through phonon vibrations it is reasonable to assume that thermal conductivity contributions from electrons are minimized. While this reduction in fidelity may be acceptable for some dynamic simulations it is critical to understand the impact. In the case of multi-layers the electronic effects are highly relevant and must be modeled explicitly. Full device simulations with *ab initio* or trained machine learning model potentials would be of interest for future work in this area.

4. MACHINE LEARNING AND DATABASE SOLUTIONS FOR MATERIALS EXPLORATION

ADAPTED FROM:

Zachary D. McClure & Alejandro Strachan. Expanding Materials Selection Via Transfer Learning for High-Temperature Oxide Selection. JOM, 73, November 2020. ©2020 The Minerals, Metals & Materials Society. [137]

David E. Farache, Juan C. Verduzco, **Zachary D. McClure**, Saaketh Desai, Alejandro Strachan. Active learning and molecular dynamics simulations to find high melting temperature alloys. https://arxiv.org/abs/2110.08136. October 2021. [138]

Zachary D. McClure ... Feature selection and interpretation for machine learning models: reducing the dimensionality of complex concentrated alloys. TBD.

4.1 Introduction: Chemical Specie Featurization

Building on the tools we developed in Ch. 3, we will extend our use of domain knowledge to build descriptors for machine learning models to expedite material screening and active learning procedures. We will first highlight the state of the art in high entropy alloy development, and address the gaps that need to filled within the sphere. Addressing the 'Inverse Problem' of machine learning we develop rapidly acquirable descriptors for oxide screening, and combine databses with accessible APIs to predict macroscopic properties for over 10,000 materials.

The methods for feature generation and selection were transferred to similar work where a functional software tool was developed to leverage molecular dynamic simulations and acquisition functions to seek out high performing alloys. The simulations were picked by trained machine learning models, and parameters of new search spaces were found with an active learning framework. Assessment of uncertainties both within the random forest models and statistical mechanics simulation were catalogued for discussion.

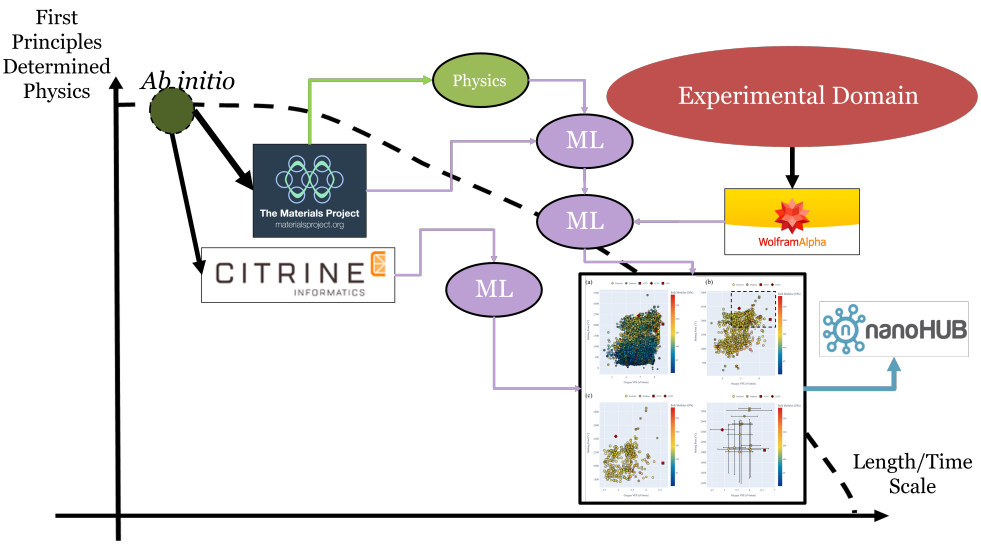

Figure 4.1. Flowchart for feature generation and material screening with machine learning. Features for chemical screening can be obtained from varied levels of fidelity, and models used to infer additional data with proper uncertainty calibration.

Finally, to demistify the 'black-box' nature of some of these models we will use model explainer methods to investigate the value of our features, and to understand what information we can provide a machine learning model for best likelihood of success.

4.2 Transfer learning for oxide selection

Materials capable of operating at high temperatures are critical for applications ranging from aerospace to energy [139] and increasing their operating envelope over the current state of the art is highly desirable. For example, increasing the operating temperature of land-based turbines by 30 °C would result in an approximate 1% efficiency increase and can translate into sector-wide fuel savings of $66 billion with significant environmental impact over a 15 year period [140]. In addition, high temperature metallic alloys can enable rotation detonation engines for hypersonic vehicles [141]. In all of these applications, high-temperature mechanical integrity or high strength are required, and so is oxidation resistance. The latter can be achieved either by the formation of a protective oxide scale during operation [142] or by the incorporation of a protective oxide (often sacrificial) during fabrication [143], [144]. This

article combines existing experimental, first principles data, and physics-based models with data science tools, including uncertainty quantification, to create a comprehensive dataset of potential oxides and the physical properties relevant for materials selection.

In recent years, complex concentrated alloys (CCAs, multi principal component alloys that lack a single dominant component) and the closely related high-entropy alloys (HEAs) [145], [146] have attracted significant attention as they have been shown to exhibit properties not possible with traditional metallic alloys [147]. Particularly interesting for high-temperature applications are refractory CCAs (RCCAs), [148] which have emerged as an attractive alternative to current superalloys. While RCCAs exhibit high-temperature strength surpassing the state of the art, their oxidation resistance is far from ideal. For example, the mass gain at T=1000°C for TiZrNbHfTa during 1 hour in air is 65 mg/cm^2, almost an order of magnitude higher than the Cr_2O_3-forming wrought Ni-based superalloys [149], [150]. Thus, efforts are underway to design RCCAs capable of growing effective oxide scales at temperatures above 1000°C [149], [151]. Beyond RCCAs, high-temperature protective oxides are required in a range of applications. Carefully engineered oxide scales can be used to prevent further oxidation and embrittlement of alloys in high temperature applications [149], [150], corrosion resistance in adverse environments [150], or as a protective coating during aerospace re-entry applications [143].

Desirable properties in these oxides include high-melting temperatures, good thermodynamic and mechanical stability to facilitate their formation over competing oxides, and low oxygen ion and cation mobility to slow down oxidation kinetics. Other properties are also desirable: a coefficient of thermal expansion (CTE) matching that of the substrate, a Pilling-Bedworth ratio (defined as $\frac{V_{oxide}}{V_{metal}}$ with V the molar volume) near one, and good adhesion to the substrate [152]. Designing RCCAs with desirable oxide scales presents additional challenges since the large number of metallic elements results in various possible, competing, oxides and complex multi-layer scales [149]. The design of RCCAs with appropriate high-temperature oxidation resistance and the selection of oxide coatings that can be added to structures would benefit enormously from an extensive database of all possible high-temperature oxides and their properties of interest.

Unfortunately, the required information is not available for the majority of the tens of thousands of stable oxides known. To date, over 60,000 metastable oxides have been studied by the Materials Project (MP) via first principles calculations [14]. Of these, about 11,000 are either the ground state or low-energy metastable structures at zero temperature. Elastic constants are known for a small sub-set, totaling roughly 1,000 oxides in the MP database [14]. However, melting temperatures are known for an even smaller subset. In this paper we use data science tools including machine learning to generate materials property information that can used for materials selection for the majority of known oxides. We build on the fact that some of these properties are correlated to each other due to similar underlying physics and can be used to create physics-based surrogate models of the quantities of interest to address the challenge of small data sets.

Cyber-infrastructure for materials data Motivated by the need for faster and less expensive materials discovery and deployment cycles [9], great strides have been made in the development of cyberinfrastructure for materials science and engineering over the last decade. Examples of this infrastructure include open and queryable repositories with first principles data, such as MP or the Open Quantum Materials Database (OQMD) [14], [153], open repositories of materials properties such as Citrination [154], and even published interatomic potential models for atomistic simulations [130]. In addition to data and models, platforms for online simulations and data analysis such as nanoHUB [12] and Google Colab [155] lower the barrier of access to simulation and data science tools for research and education [156]. These repositories are making strides towards making data findable, accessible, interoperable, and reproducible (FAIR) [11]. Data can be queried through online user interfaces or via application programming interfaces (APIs) for rapid querying and analysis of data.

Transfer learning for materials selection. Materials selection requires access to data and often involves a multi-objective optimization [157], [158]. This was traditionally done with existing experimental data, sometimes combined with simple models [159]. More recently, *ab initio* electronic structure calculations have been incorporated to such efforts [160] and progress in multiscale modeling is providing additional tools to materials design and optimization [161]. In addition to such data, machine learning tools are being used to assess the current state of knowledge and make decisions. In our application, it would be tempting

to use machine learning to develop models to predict our quantities of interest (QoI), such as melting temperature, from composition using the available data; these models could then be used to explore the properties of a wide range of oxides. Unfortunately, the limited set of known melting temperatures precludes such an approach as standard ML methods require vast amounts of data, on the order of 10^3-10^6 datapoints. The lack of data is common in materials applications and several approaches have been developed to address it.

Many of the most relevant studies have successfully been able to use traditional machine learning approaches, albeit on minimal datasets ranging from 10^1-10^3, through the use of careful descriptors and transfer learning approaches. This has been used, for example, to screen billions of compositions for Li-ion conductivity [162], using preliminary *ab initio* data to extend 101 compounds into a space of 54,779 [163], and showing that by distinguishing between descriptors such as energy difference between phases accurate predictions can be made on datasets as little as 82 samples [164]. These methods compensate the lack of large amounts of data with domain expertise, physics, and chemistry. These approaches are not unique to the field of materials. In fact, they have been extensively used in chemistry for polymer selection [165], and design of chemical compounds [166] for decades.

One such method is to enhance the information fed to the model by adding surrogate properties as inputs. These surrogate properties should be both easy to obtain and be expected to correlate to the quantity of interest. In this paper, we use the oxide stiffness (easily computable via *ab initio* simulations) and melting temperature estimates using Lindemann's law [167] as additional inputs to the model to predict melting temperatures. Lindemann's estimates can be easily obtained from available properties and can be expected to serve as good surrogates based on prior studies in oxides [168] and minerals [169]. We note that stiffness and melting temperature are both governed by the strength of the inter-atomic interactions, there is a correlation between these properties and adding stiffness as an input to the models results better accuracy.

4.2.1 Currently available data

The design or selection of protective oxide scales would benefit from access to materials properties for all possible oxides that are either stable or metastable at the operating temperatures. As discussed above, a large number of oxides structures are known, but high-temperature data, including melting temperatures are known for a small subset. Thus, we start from all known oxides and combine existing data with machine learning to provide information about structures for which we lack experimental data. This section explores the relevant data available in online repositories and Sec. 4.2.2 discusses the use, combination, and extension of the data.

As discussed above, several materials data repositories focusing on various types of data and materials classes are available today. We leveraged the MP database, Citrination, and WolframAlpha [170]. The MP is a database with density functional theory (DFT) results including crystal structure data, relative stability to the ground state, elastic constants for select materials, calculated X-ray diffraction (XRD) and X-ray photoelectron spectoscopy (XPS) spectra, and even T=0 K phase diagrams for compounds. MP has information about a majority of known oxides and we start our search within this list. The properties in the MP can be accurately calculated from first principles calculations; however, properties like melting temperature are computationally too intensive for high-throughput DFT calculations. Therefore we turn to repositories with extended datasets for additional information like melting temperature and oxygen vacancy formation energy (VFE).

The Citrination database [154] is an open repository where researchers can upload their own data and share it with the community at large. At the time of writing, citrination contains 454 public databases curated by public users and Citrine staff. Databases previously curated through research efforts have been published in their database and are freely available for download and use. For our efforts we turned to Citrination for databases of oxygen VFE.

We were unable to find an electronic database with melting temperatures for oxides, most reported melting points exist within individual papers, collected handbooks, or commercial databases. However, we were able to find some of these properties in WolframAlpha, a general

purpose, queryable, compute engine. Using WolframAlpha, we generated a list of melting temperatures for a subset of the oxides queried from MP with elasticity data.

The Materials Project: basic oxide data

We accessed the MP database using the Pymatgen API [15] and analyzed every oxide available. MP contains information about 60,000 distinct oxides (differing either by composition or crystal structure). All these structures were obtained by energy minimization using DFT within the generalized gradient approximation and additional details of the calculations can be found in work by Jain et al. [171]. Confirming the metastability of these structures would require positive phonon frequencies and elastic constants to discard local energy maxima; these quantities are not available for all these oxides. To address this challenge we first filtered the data to retain only structures that are 1 meV above the convex hull (i.e. the predicted ground state for that composition). We note that the energies resulting from energy minimization correspond to a temperature of T=0 K (minus zero point energy) and phases with free energy higher that the ground state at 0 K can be stabilized at higher temperatures due to entropic contributions to the free energy. Furthermore, many metastable structures are long-lived and used in applications. After this stability constraint, we are left with ∼11,000 possible oxides. However, elastic constants are documented in MP (from DFT calculations) for a subset of 855 oxides. To illustrate the available data, Figure 4.2 compares two properties of the available oxides after filtering by energy stability and elastic constants. We plot the ionic packing fraction (defined as the total atomic volume assuming hard spheres with the corresponding ionic radius in the unit cell divided by the cell volume) vs. density obtained from the crystal structure data. Red points indicate oxides with at least one element that is found in RCCAs: we select Ti, V, Cr, Zr, Nb, Mo, Ru, Hf, Ta, W, and Re as well as Al, Cr, and Si since they are useful additives. Of the 855 oxides with elasticity data, 235 of them contain an element pertaining to an RCCA or additive compound. The figure also highlights common protective oxides Al_2O_3 and Cr_2O_3; as expected these oxides have high packing fractions (which correlates in low ionic diffusivity). Interestingly, there are a number of potential compounds with comparable properties to both these common oxides.

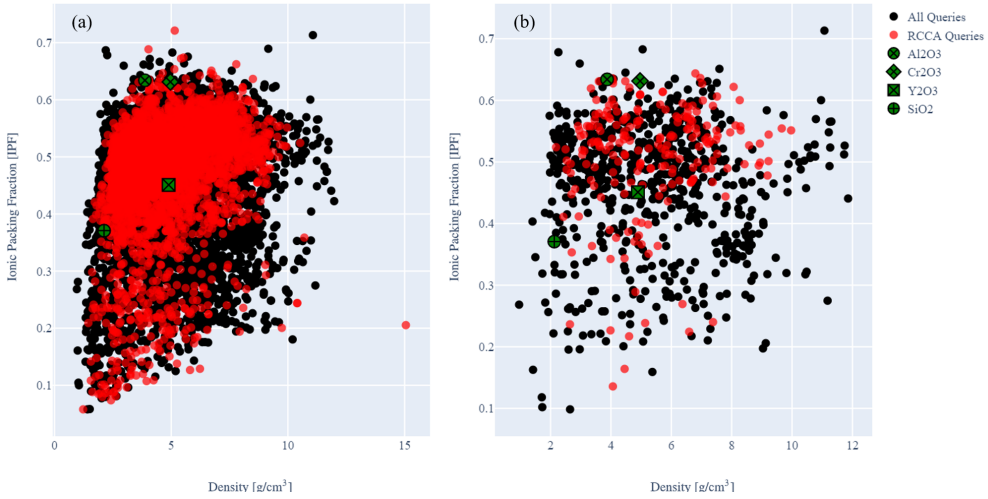

Figure 4.2. Calculated ionic packing fraction of oxides and their queried densities. a) Database curated post energy stability filtering. b) Database curated post elasticity filtering

With this basic information at hand, we now focus on the remaining properties: melting temperature and ionic mobility. While these properties can be obtained, in theory, from first principles they are computationally very intensive and they are not included in the MP. Therefore we turn to other repositories for additional DFT and experimental data.

WolframAlpha: melting temperatures

At the time of writing, melting temperatures of the oxides of interest were not available in materials-specific online repositories. A single curated inorganic melting point database on Citrination exists, but many of the values are not oxides, and do not overlap where we have existing elasticity data. Fortunately, WolframAlpha provides an API for data exploration. Through a series of string queries we obtained and curated melting points of 158 oxides into our database. Since this data is significantly less abundant than the elasticity data from Materials Project we will consider the melting point to be our harder to acquire, or more expensive, set of data. Improvements to this dataset can be made through literature searches, or analyzing phase diagram textbooks, but our goal is to illustrate a rapid acquisition of data rather than the traditional task of searching through physical copies of information. Figure 4.3 shows the results of the melting point query with respect to density and IPF properties. RCCA contaning oxides are highlighted to guide the eye, and as expected a number of them have comparable properties to common oxides such as Al_2O_3 and Cr_2O_3.

Citrination: vacancy formation energies and thermal expansion

Ionic mobility is another critical material property in the design of protective oxides, unfortunately, ionic mobility (oxygen or cation) is not widely available. However, since oxygen mobility is mediated by vacancies, the vacancy formation energy is a good surrogate for ionic transport: the higher the vacancy formation, the lower the vacancy concentration and oxygen ion mobility. Citrination includes a database of nearly 2,000 charge neutral vacancy formation energies of oxides based on first principles approaches originally published by Deml et al. [172]. Of this dataset 1,200 were unique oxide compounds.

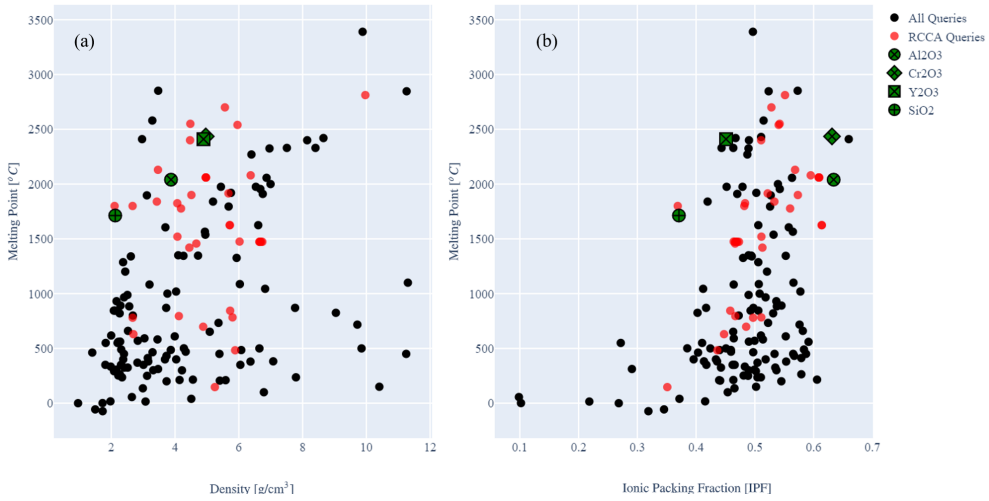

Figure 4.3. Queried melting points from WolframAlpha with bulk modulus (a) and IPF (b) properties.

Another database of importance is available based on work from by Shick et al. [173] containing 69 average coefficient of thermal expansion (CTE) values obtained from anharmonic phonon calculations. We will consider the use of this database in future work, but at this time the limited dataset provides a great challenge to accurate predictions outside the selected compounds via machine learning methods. A transfer learning approach could be used, but our focus here will be the melting temperature.

4.2.2 Extending oxide data via data-driven transfer learning

In summary, from online accessible databases we were able to extract 11,000 stable and metastable oxides from an initial 60,000 query on MP. Of these 11,000 possible oxides roughly 1,000 have existing elasticity data. From the list of 1,000 oxides with elasticity information only 162 melting points were obtained through queries. In addition, we have VFE values for 1,200 cases and CTE for 69. Ultimately our goal is be to build models with each of these properties, and use the information leveraged from each to extend a materials search into the original 11,000 stable and metastable oxides.

Since filling in the gaps in data discussed in Section 4.2.1 via first principles calculations or experiments would be prohibitive in terms of time and cost, we will explore using data-driven machine learning tools like neural networks [174], [175] and random forests [176]. One could attempt to train models that relate composition to the final QoI (e.g. melting temperatures) from the existing data. However, standard ML approaches are not applicable directly due to the scarcity of the data. This is a common challenge in the field of materials. Feature engineering, which involves feeding additional data to the model that can be easily obtained from the raw input data, can be used to address this challenge. For example, we could include electronegativity and ionic radii of the elements as inputs to the model, thus, we include information about bonding and packing. In addition to such *periodic table* data, one can further increase the information fed into the model by adding physics-based modeling results or material properties that are easy to obtain and that are expected to correlate with it. It has been shown that even with limited training data physics-based descriptors have had a significantly higher impact than models that only rely on raw volumes of data

[162], [177]. Here we will build on two surrogate pieces of information. First since melting temperature and stiffness are both governed by similar physics, stiffness (available from first principles calculations for 1,000+ oxides) is added as an input parameter. Second, we can use Lindemann's melting temperature model to estimate values of the QoI from basic properties and add that estimate as an input. Lindemann's melting law is based on the approximation that melting occurs when atomic oscillations reach a critical value relative to the materials lattice parameter. The amplitude of atomic oscillations can be easily obtained using statistical mechanics and the resulting expression for the melting temperature is:

$$T_m^{'} = \alpha(4\pi^2 k/9h^2 N^{5/3}) f^2 \overline{M} V^{2/3} \Theta_D^2 \tag{4.1}$$

where, $T_m^{'}$ is the Lindemann melting temperature, α is a structural factor generally taken as 1, k, h, and N are the Boltzmann, Planck constants respectively, $\overline{M}$ is the mean atomic mass, V the molar volume, and Θ_D the Debye temperature.

The Debye temperature is taken from the approximation proposed by Blackman [178] using the expression:

$$\Theta_D = (h/k)(3N/4\pi)^{1/3} V^{-1/3} \nu_m \tag{4.2}$$

with ν_m as the average acoustic velocity given by:

$$3/\nu_m^3 = 1/\nu_p^3 + 2/\nu_s^3 \tag{4.3}$$

with ν_p and ν_s the P and S wave velocities calculated from the root of the bulk and shear modulii with their densities (ρ):

$$\nu_p = \sqrt{\frac{K + \frac{4}{3}G}{\rho}} \tag{4.4}$$

$$\nu_s = \sqrt{\frac{G}{\rho}} \tag{4.5}$$

Descriptors

As mentioned above, a common way of building physics into ML models is to use periodic table data of the elements involved as inputs. We primarily use the *composition featurizer* from Matminer [179] to generate a variety of properties with composition as the only input. As shown in previous work by Ward et al. [180], statistical descriptors based on the chemical formula are useful for machine learning features.

The descriptors we use are described as the follows:

1. A stoichiometric calculation of fractions of elements without considering the actual composition. This calculation includes number of elements in the compound and normalizations of the respective fractions.

2. Periodic table type descriptors including mean, mean absolute deviation, range, minimum, maximum, and mode of elemental properties. These values include maximum row on the periodic table, average atomic number, and the difference in atomic radii in all elements present.

3. As previously shown by Meredig et al. [181], electronic structure attributes with averages of s, p, d, and f valence shell electron concentrations are useful as descriptor inputs.

4. Assuming that the ionic species in the oxide can form a single oxidation state, an adaption of the fractional ionic character of a compound can be used based on an electronegativity-based measure [182].

5. The fraction of the transition metal elements.

6. The cohesive energy per atom using elemental cohesive energies.

7. An estimation of the band gap center based on electronegativity.

8. Number of available oxidation states in the compound.

9. For mechanical properties models we also extend descriptors to include properties queried from the MP database like density, space group number, and calculated ionic packing fractions. Importantly, we added two descriptors for the meting temperature models: predicted stiffness properties and estimate of the melting temperature according to Lindemann's law discussed above.

These descriptors are able to characterize the output properties for VFE sufficiently, and we do not see evidence of over parameterization of the models. For stiffness we add additional descriptors queried from MP, and for the melting point we use the full knowledge of composition descriptors, queried MP properties, predicted stiffness, and Lindemann law melting predictions.

Predictive models for melting temperature using random forests

Random forests (RFs) approach regression methods through a series of decision trees [44], [183] whose outputs are averaged. This averaging is done to overcome the limitation of individual tree predictions which may have difficulty assessing noise or non-linearities in the data. Importantly, progress has been made in the quantification of uncertainties in RFs by Efron [184] and Wager et al. [185], and more recently by Ling et al. [176] with the addition of an explicit bias term to the uncertainty. Neural networks, often outperforming random forest predictions, were considered for this study, but quantification of uncertainty in their outputs is still an active field of research [186]. Due to the accessibility of uncertainty quantification we choose to implement random forest models with the state of the art uncertainty calibration proposed by Ling et al [176]. It involves sample-wise variance defined as the average of the jackknife-after-bootstrap and infinitesimal jackknife variance estimates with a Monte Carlo sampling correction. The RF models implemented in this study are available in the Lolo scala library [187].

We set the maximum number of trees to match the number of samples collected in each model for our RFs while allowing for uncrestricted maximum depth. While saturation of averaged prediction can occur beyond 200 or more trees [188] the uncertainties in Lolo will not be well calibrated. The maximum depth parameter cutoff is defined by the nodes

increasing until the leaves become pure, or until the all leaves contain less than two samples. This is the default parameter for Lolopy.

As is common practice, each descriptor descriptor is normalized by standard normalization and data was split into 80% training and 20% testing to evaluate performance. Assessment of the model was performed for each material property by reshuffling the dataset 10 different times, and taking an aggregated MAE.

When assessing uncertainty estimates for an individual output x, the residuals, $r(x)$, of the prediction when normalized by the uncertainty $\sigma(x)$ ($N = \frac{r(x)}{\sigma(x)}$), should have a Gaussian distribution with zero mean and unit standard deviation. This metric can help quantify if the random forest uncertainty predictions are well calibrated with respect to the inherent error predicted.

Using the set of descriptors and architecture detailed in Sec. 4.2.2, we implement random forest models to predict the set of desired properties using databases from MP, WolframAlpha, and Citrination. All reported MAE values are taken as an aggregate mean after shuffling the training and testing sets 10 times.

Random forest performance for VFE

Using the curated Citrination dataset we developed a RF model for VFE. Composition based descriptors obtained via Matminer were used for model predictions. For 10 shuffling samples we report an aggregated MAE of 0.17 eV/atom.

Random forest performance for stiffness

In addition to the Matiminer featurizers described above we added additional descriptors such as IPF and space group number since these were easily queried. Figure 4.5 shows a parity plots and normalized residuals for bulk and shear moduli. An aggregated testing MAE score of 18 and 10 GPa for bulk and shear modulus, respectively, was obtained after 10 shuffling of samples.

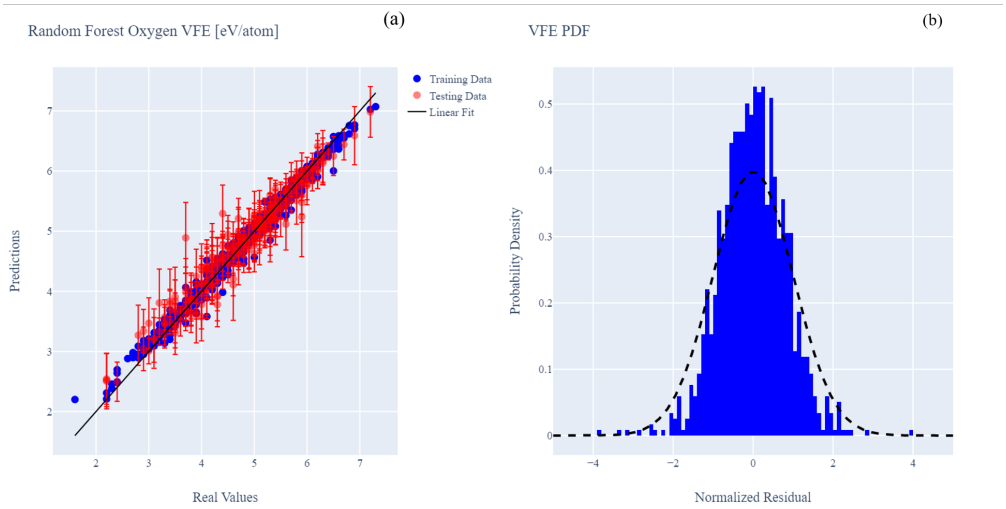

Figure 4.4. a) Parity plot diagram for predicted and real values of oxide VFE. Values directly on the line are a perfect match. b) Normalized residuals for VFE with Gaussian like distribution.

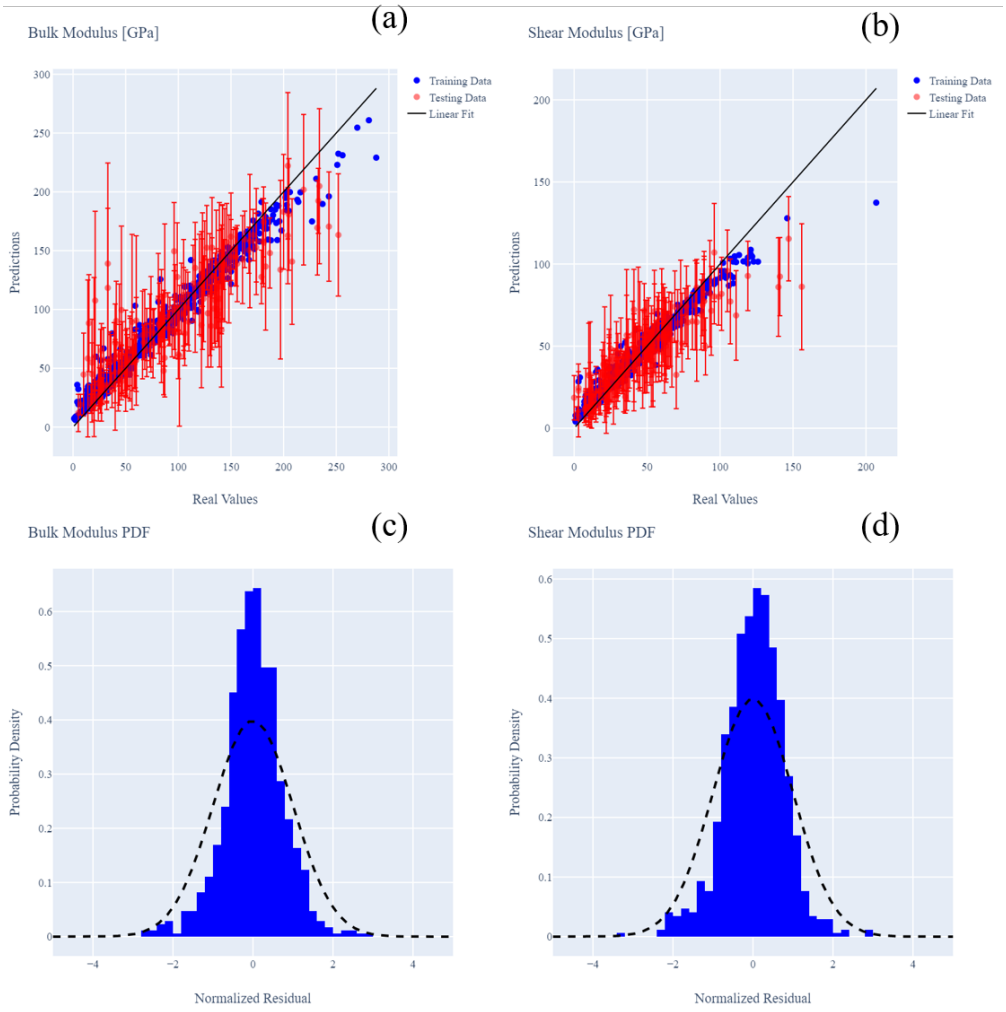

Figure 4.5. a) Parity plot diagram for predicted and real values of oxide bulk modulus, b) shear modulus. Included are normalized residual calculations for c) bulk modulus, and d) shear modulus.

Random forest performance for melting temperature

Our dataset of 162 melting points with corresponding elasticity data was used to create a predictive model for varied oxides. Fig. 4.6 shows the performance of both the training and testing data before and after adding stiffness and Lindemann properties into the model. As we can see adding stiffness and Lindemann's law as descriptors improves the accuracy of the model to a significant degree with a reduction of the MAE from 368 to 303 ^{o}C. While uncertainties are not negligible, these models are promising for an initial sweep of potential oxides. A noticeable reduction in uncertainty can be seen between the Fig. 4.6(a) and Fig. 4.6(b), and after adding stiffness and Lindemann's law fewer points lie outside the linear fit in the parity plot. After training the model we use identical descriptors for the remaining compounds that we were unable to easily obtain melting points for and extend our predictions using the information gained from stiffness and melting point models. In Sec. 4.2.4 we will assess some the sensitivity of this prediction with varied UQ methods. In the outlook section of this paper we will discuss the implications and results of extrapolating our predictions to other oxide melting points outside of our initial query with WolframAlpha.

4.2.3 Materials selection for protective oxide scales

Using the models above we begin to extend our search space of potential oxides from our initial query of 162 melting points and 855 points with elasticity data and move into the space of the remaining 11,000 stable oxides from MP. First, we predict the elasticity data of the remaining 11,000 oxides that did not have this data to begin with. Then we use those descriptors to expand our melting point database from 158 queried data points to nearly 11,000 data points: a two order of magnitude increase.

Figure 4.7(a) shows the 11,000 oxides and their respective properties. We show melting temperature and the oxygen VFE, bulk modulus is shown as the color of the symbol, and the IPF is represented by size. Figure 4.7(b) filters radioactive elements and lanthanides out, and also removes bulk and shear modulus values below 125 and 25 GPa respectively. The plot highlights common and effective protective oxides. As expected, Cr_2O_3, Al_2O_3, and SiO_2 are among the top performers. However, our study reveals other oxides predicted to perform

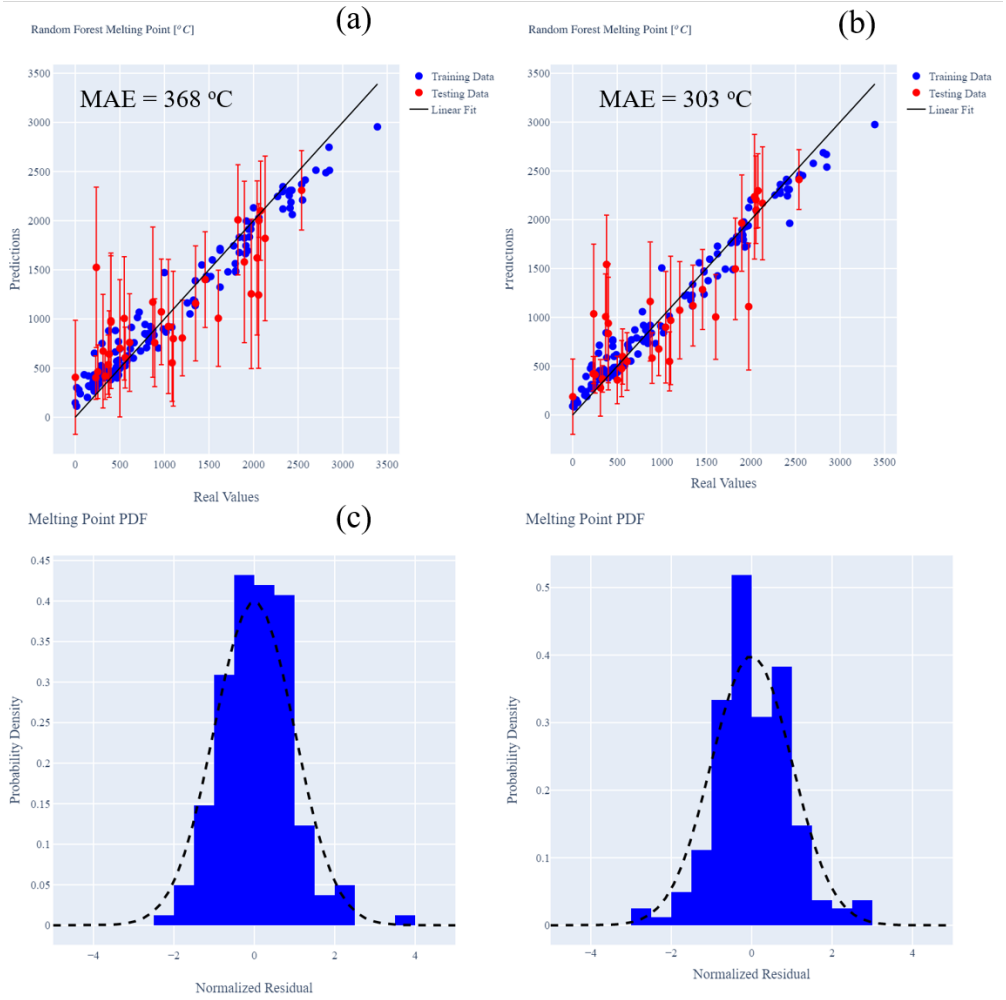

Figure 4.6. a) Parity plot diagram for predicted and real values of oxide melting temperature. Values directly on the line are a perfect match. b) Adding stiffness and Lindemann melting law properties to the model causes a decrease of 65 ^{o}C with respect to MAE and a noticeable decrease in uncertainty, c) and d) show normalized residuals for models trained without and with additional descriptors.

equally well or outperform them. Fig. 4.7(c) shows the final filtering of outlier properties such as low VFE, low melting point, and IPF values below 0.4 and Fig. 4.7(d) shows these final points including the uncertainties in the RF model. Data points with a cross represent materials with existing melting temperatures from WolframAlpha and empty symbols are predictions. Values without error bars in either direction indicate database collected values.

The top oxides identified and their properties are summarized on Table 4.1. In the case of the design of refractory CCAs, HfO_2 (T_{melt} = 2812°C, VFE = 5.9 eV/atom) is an attractive candidate since Hf is a common element. Our results indicate that the addition of Y as a dopant to RCCAs could result in the formation of $Y_2Hf_2O_7$, $YTaO_4$, $Y_3Al_5O_{12}$, or Y_6WO_{12}. While many of these have lower predicted melting points (in the 1900-2000° range), they may stabilize as complex oxides between the outer scale and substrate. Each of these oxides coupled with the RCCA substrate could be engineered to form a stabilized complex oxide of one or more of these structures. Quite interestingly, even though the Y containing compounds in Table 4.1 were not present in our initial database of melting points, they have been investigated as promising candidates for thermal barrier coatings [189] and scales in high temperature applications. Synthesis routes have been discussed in the literature. Reported melting temperatures include 2300°C for $YTaO_4$ [190], the well studied yttrium aluminum garnet (YAG) compound melting at roughly 1940°C [191], stability of single phase under ablation temperatures of over 2000° for $Y_2Hf_2O_7$ [192], and finally Y_6WO_{12} melting at 2360°C [193]. Our predicted values of the melting point through random forests with predicted uncertainties fall very close to experimental results which is encouraging for potential extrapolation of other materials for RCCA protective scales.

Two of the oxides predicted to be of interest for high temperature applications lack experimental melting temperatures. Also containing Y in their structure, $Y_3Al_3Cr_2O_{12}$ and Y_3ReO_8 at this time do not have reported melting points. Each of them have predicted melting points exceeding 1900^{o}C, and each have elements that could be used as base components in RCCA applications. Already proven to be an excellent candidate for scale formation and physical properties is Cr, but more interesting is the presence of Re in Y_3ReO_8. In the 2nd generation of Ni-based superalloys Re was proven to be an excellent dopant for extending the creep lifetime in alloys [194]. While expensive, the addition of such an element to RCCA

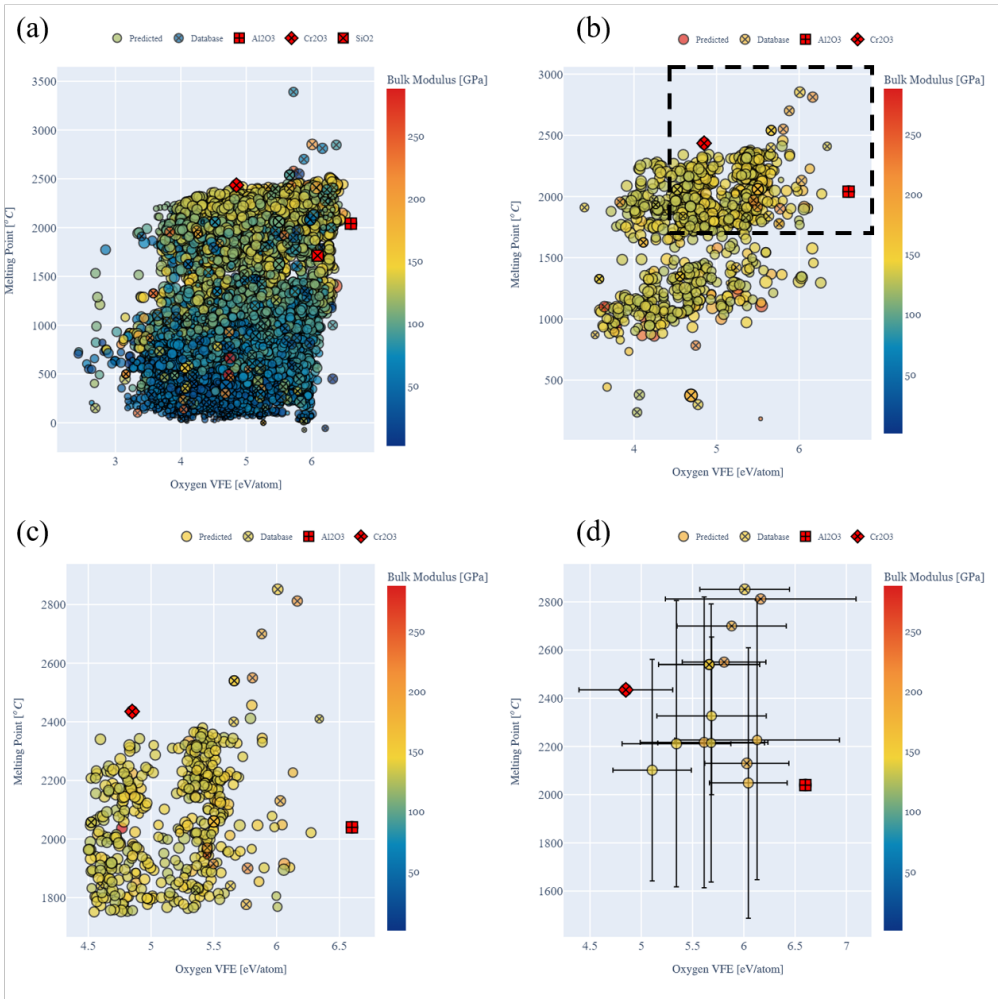

Figure 4.7. Comparison of melting point and vacancy formation energy of oxide compounds. Coloring corresponds to stiffness of the material, and marker size indicates IPF where larger markers are a higher IPF. Points with an 'x' are melting points collected from queryable sources where open circles are predicted values. a) Predicted results for original 11,000 query. b) Results filtered to remove radioactive and lanthanide compounds, and bulk and shear modulus values below 125 and 25 GPa respectively c) Additional filtering of properties with remaining values including IPF > 0.4, $T_{melt} > 1750$°C, and VFE > 4.5 eV/atom. d) Selected compounds for final application consideration. These compounds are listed in Table I as potential complex or native scale formers. Values that have database values do not show error in respective direction. Note the slightly different scales in the filtered figures.

Table 4.1. Final compounds with uncertainties. If zero uncertainty is reported the value was obtained from database results. Experimental validation of results for predicted values is shown in middle column. Entries with '?' indicate an unknown melting point at this time in literature.

	RF T_m [^{o}C]	Exp. T_m [^{o}C]	Oxygen VFE [eV/atom]
Al_2O_3	–	2040†	6.6±0
Cr_2O_3	–	2435†	4.91±0.48
HfO_2	–	2812†	5.99±0.89
MgO	–	2852†	6.0±0.43
$MgAl_2O_4$	–	2130†	6.05±0.38
ZrO_2	–	2700†	5.79±0.52
$ZrSiO_4$	–	2550†	5.71±0.39
$BaZrO_3$	–	2540†	5.63±0.48
$SrZrO_3$	2326±327	2610 [195]	5.62±0.49
$Y_2Hf_2O_7$	2226±579	2000* [192]	5.66±0.77
Y_6WO_{12}	2214±576	2360 [193]	5.66±0.45
$YTaO_4$	2217±603	2300 [190]	5.59±0.55
$Y_3Al_5O_{12}$	2048±561	1940 [191]	6.04±0.39
$Y_3Al_3Cr_2O_{12}$	2101±459	?	4.39±0.42
Y_3ReO_8	2211±594	?	6.89±0.44

† Experimental value from queried dataset
* Ablation study

type materials could have potential for oxide scale formation as well as modifying the overall mechanical properties of the material.

Other notable oxides that we found with excellent properties include: MgO (T_{melt} = 2852°C, VFE = 6.0 eV/atom), $MgAl_2O_4$ (T_{melt} = 2130°C, VFE = 6.05 eV/atom), ZrO_2 (T_{melt} = 2700°C, VFE = 5.79 eV/atom), $BaZrO_3$ (T_{melt} = 2450°C, VFE = 5.63 eV/atom), $ZrSiO_4$ (T_{melt} = 2550°C, VFE = 5.70 eV/atom), and $SrZrO_3$ (T_{melt} = 2326°C, VFE = 5.6 eV/atom). Of these other promising candidates we note the experimental melting point for $SrZrO_3$ to be recorded as 2610°C [195], well within the predicted value with random forest uncertainties. We would like to stress that additional variables need to considered in the design of oxide scales, such as processability and kinetics; these are not considered in this first study.

4.2.4 Uncertainty propagation on the melting temperature calculation

As in any decision-making exercise, uncertainties are critical in materials selection and optimization. Several sources of uncertainties must be accounted for in workflows such as the one used here. These include uncertainties in the ML models and in the input and output data fed to them. An additional challenge in our approach is the combination of experimental (e.g. melting temperatures) and first principles data (e.g. elastic constants). One could expect for systematic errors in surrogate data, like our DFT stiffness values, to be of relatively low importance as they are only used to *help* the ML models. For example, if gradient corrected exchange and correlation functionals used in DFT tends to underestimate binding; the ML models should be able to easily compensate for such discrepancies. The effect of non-systematic errors in input data like Lindemann's melting law predictions are harder to estimate. Here we focus on a specific kind of uncertainty that originates in transfer learning, i.e. how uncertainties are propagated across models in the transport process.

When using RF-predicted values for bulk and shear modulus as input descriptors to the melting point model, it is critical to assess how the uncertainties in elastic constants affect the predicted T_m. We note that the majority of the compounds in the final list of oxides selected in Section 4.2.3 had first principles elastic constant data. One exception is $BaTi_2O_5$ so we use this material to study uncertainty propagation. The predicted mean melting point for this specific compound was 2144±435^{o}C, this was obtained with mean bulk and shear moduli. Since the elasticity models yield mean and the associated deviations, we can assess how sensitive the predicted melting point is to uncertainties in the moduli parameters. Our trained random forest models predict mean values of 142±27 and 75±16 GPa for bulk and shear modulus, respectively.

To propagate uncertainties in elastic constants through the melting temperature model, we use a brute force random sampling of the Gaussian distribution for each stiffness property. The resulting distribution from 10,000 samples is shown in black in Fig. 4.8(a). The predicted distribution shows a sharp peak at 2150 ^{o}C, very close to the mean prediction, and extends towards lower values with a second peak at 1950 ^{o}C, and a third smaller distribution centered at 1700 ^{o}C. The predicted RF distribution of melting temperatures with mean stiffness values is

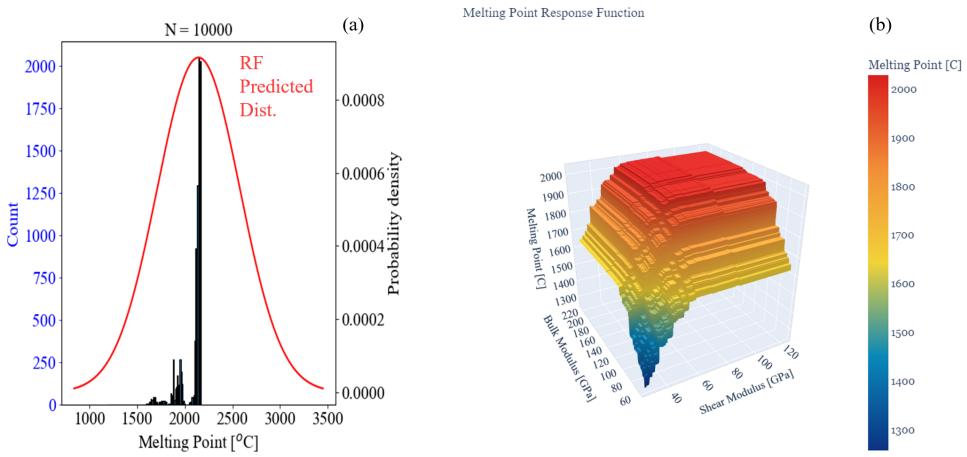

Figure 4.8. a) Histogram results for Monte Carlo (MC) sampling of bulk and shear modulus compared to original random forest (RF) predicted distribution. b) Response surface for shear (x-axis) and bulk modulus (y-axis) with respect to melting temperature (z-axis).

shown in red. Importantly, the uncertainties originating from the propagation of uncertainties in the stiffness are small compared with the intrinsic uncertainties in the prediction of melting temperature. This is, perhaps, not surprising since the melting temperature model has larger uncertainties that that for stiffness. The multi-peak nature of the distributions indicates large non-linearities in the T_m model. To assess this, we plot the melting temperature as a function of shear and bulk modulus in Fig. 4.8(b) with all other paramters fixed to those of $BaTi_2O_5$. We find that melting temperature drops quite significantly for low values of shear and bulk moduli. This is not surprising given the positive correlation between stiffness and melting temperature, but such extrapolations using machine learning models should be done with care.

4.2.5 Summary and outlook

We showed that by leveraging queryable open repositories and the use of machine learning tools with infused physics one can greatly expand the information available for materials design or selection. Our specific goal was to find oxides for high temperature

applications with high melting temperature, high oxygen vacancy formation energy (to minimize O transport) with coefficient of thermal expansion and stiffness as secondary design variables.

Machine learning models with physics insight built in via feature engineering and surrogate properties enables us to take sparse existing data and fill-in gaps in knowledge. Specifically, we found that by adding elastic constants (known for a relatively large number of oxides) as an input descriptor, and easily calculated Lindemann melting laws we could develop accurate models for melting temperature which are harder to obtain and exist for relatively small number of materials.

Through transfer learning we were able to expand an initial query of 162 melting points to over 10,000 compounds. The effort resulted in several candidate oxides with properties comparable to those of the protective scales of high-temperature metals such as Al_2O_3 and Cr_2O_3 with respect to melting point, VFE, IPF, and stiffness. Candidate materials include: HfO_2, $Y_2Hf_2O_7$, $YTaO_4$, $Y_3Al_5O_{12}$, Y_6WO_{12}, $Y_3Al_3Cr_2O_{12}$, Y_3ReO_8, MgO, $MgAl_2O_4$, ZrO_2, $BaZrO_3$, $ZrSiO_4$, and $SrZrO_3$.

Quantifying uncertainties in such efforts is critical in materials selection and optimization efforts. In this paper we focus on the uncertainties propagated through predicted stiffness parameters, and the uncertainties in the random forest. Additional work on uncertainties originating from combining information from different sources (e.g. DFT and experiments) would be very valuable.

Contribution to these databases remains key, and we intend to supplement the Materials Project Database with more calculations for elasticity from first principles, and the curated datasets from this project will be made available on Citrination for public use. Additional extensions of queried data to supplement our models from the previously mentioned databases is a continued area of work.

The models built and developed in this paper can be accessed through the nanoHUB tool High Temperature Oxide Property Explorer [196]. Final curated data can be downloaded, and models can be modified at the leisure of the user

Gathering the initial information from materials informatics platforms is a key step in our workflow and many similar efforts. This is enabled by recent progress on materials

cyberinfrastructure and community contributions to these repositories remains key. In the case of oxides, additional elastic constant calculations and melting temperatures would be beneficial. The models developed in this paper can be accessed online through the US National Science Foundation's nanoHUB [12]. The tool High Temperature Oxide Property Explorer [196] includes live Jupyter notebooks with all models and data. The final curated data and models can be downloaded but the can also be modified and executed online.

A limitation to the process included here is the availability of dense data across a range of compositional spaces. While we were fortunate to have a baseline database with the Materials Project, it is often the case that we must curate our own from scratch, or with limited initial information. Another method that can be used to expedite the screening of material systems is active learning. With the methods described above, but attaching acquisition function algorithms on the fly experiments or simulations can be performed with the instruction of a trained model. In the following section we will highlight how using random forests with limited initial information can begin to build their own database where one previously is unavailable. The construction of a new database with iterative learning protocols is shown for the driving of a Sim2L for MD melting point simulations.

4.3 Active learning for high melting temperature alloys

4.3.1 Introduction

The combination of experiments, physics-based simulations, and data science tools has been shown to have the potential to accelerate discovery of novel materials with optimized properties [163], [176], [197]–[206]. Active learning (AL), a subset of machine learning in which models learn dynamically, has gathered significant interest, both from the point of basic science [9] and for commercial applications [207]. AL models analyze existing data and formulate queries to acquire additional information towards a design goal. AL workflows start with a model trained with an initial set of data and evaluate the expected gain towards an objective function of all possible new experiments within a design space [208]. Top candidates are characterized (e.g. by performing an experiment) and the outcome is added to the existing dataset. Given this additional information, a new cycle is started. With this iterative process,

illustrated in Figure 4.9, the model becomes more accurate in regions of interest within the design space. In order to identify the next best query, AL uses selection strategies known as information acquisition functions. These strategies differ from each other in the relative balance between exploitation and exploration included in their mathematical formulation. Exploitation favors cases expected to maximize the objective function while exploration focuses on areas of high uncertainty.

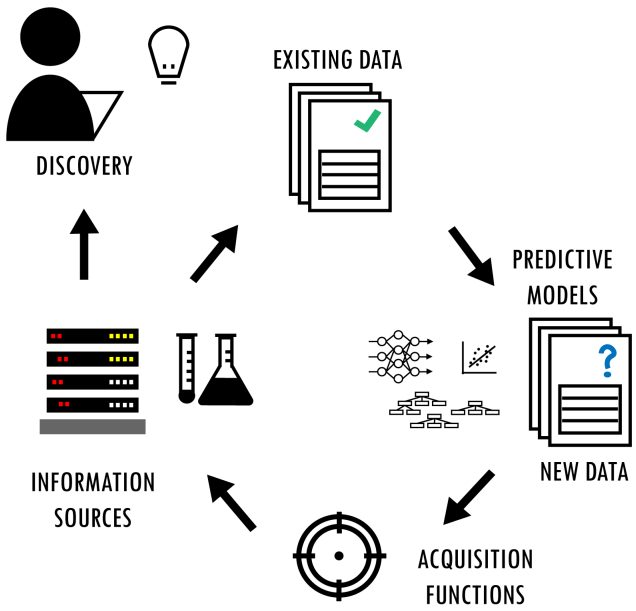

Figure 4.9. Schematic representation of the AL iterative optimization. Starting from existing data, a predictive model is trained with machine learning. The model is then applied to all candidate materials in the design space to assess their performance, including uncertainties. An acquisition function is used to select the next material(s) to be tested (either via experiments or simulations). If the new material does not satisfy the design conditions, it is added back to the existing data set and the cycle re-initiated.

AL has been used for a wide range of applications from natural language processing [209], reaction screening for pharmaceutical applications [197], and multiscale modeling [198], [210]. In materials science, AL has been used to accelerate discovery of materials with desired properties by coupling it with experiments [200], [211]–[215] and physics-based simulations [163], [201], [202]. In addition, AL workflows paired with existing closed data sets has shown

the ability of these models to reduce the number of queries needed to identify the best candidate [176], [204]–[206].

Kusne et al. [199] developed a closed-loop autonomous algorithm for materials discovery and used it to explore the Ge-Sb-Te ternary system in search for an optimal phase-change material with the highest optical bandgap difference between phases. With an iterative approach using experimental ellipsometry data, they identified a new composition, $Ge_4Sb_6Te_7$, with nearly three times the optical bandgap of $Ge_2Sb_2Te_5$, a material used in random-access memory devices. Xue et al. [200] demonstrated an adaptive design strategy for the search of NiTi-based shape memory alloys with low thermal hysteresis. Starting with 26 materials synthesized their optimization algorithm was used to screen a space of ∼800,000 candidate materials. This screening resulted in 36 alloys queried and tested using differential scanning calorimetry (DSC) with several showing lower thermal hysteresis than the starting training set. For additional examples, see Refs. [212]–[215].

Coupling AL workflows with physics-based simulations can further reduce the costs and time required to discover new materials by minimizing the number of experimental trials required. Tran and Ulissi [201] explored a high dimensional space of density functional theory (DFT) results on intermetallics in search for optimized catalyst materials for CO_2 reduction. Their approach produced a set of 54 promising materials from ∼1499 candidates in the Materials Project [202]. Out of these materials, some have been since confirmed experimentally. Seko et al. [163] reported a virtual screening of 54,779 materials looking for compounds of low lattice thermal conductivity. They used AL to select appropriate descriptors for a Gaussian process regression (GPR) model used to get information on the large dataset. Their screening yielded 221 materials with very low conductivity. After some constraints were imposed, 19 of them were characterized using DFT calculations for possible applications as thermoelectric materials.

Work involving the combination of AL or similar optimization methods and materials simulations has been mostly limited to DFT calculations and to properties that can be obtained from relatively small simulation cells without noise. While DFT often offers a good balance between accuracy and computational efficiency, it remains limited to a subset of materials properties. Fortunately, materials models across scales are available [161] and

significant progress in their coupling across scales has occurred over the last few decades. Examples range from crystal plasticity models [205], [216] to interatomic potentials from ab-initio simulations [38], [210], [217]–[220]. Physics-based materials models across scales open the possibility of significantly expanding the reach of AL approaches.

Molecular dynamics (MD) is an important rung in multiscale modeling that connects *ab initio* physics to the meso- and macro-scales. MD has been paired with optimization methods like reinforcement learning, including Monte Carlo tree search (MCTS) [221], [222]. For example, Patra et al. [223] paired MCTS with MD to find a copolymer compatibilizer, and Loeffler et al. [224] applied the methods to study defect structures in metal dichalcogenides.

In this work, we couple AL with MD simulations to find multiple principal component alloys (MPCAs) with high predicted melting temperature for a model CrCoCuFeNi system. The main objective of this paper is to demonstrate the potential of coupling of AL workflows with MD simulations and characterizing the effect of the stochastic and noisy nature of MD results on AL workflows. We find that even under the noise in relatively short MD simulations, acquisition functions that balance exploration with exploitation are capable of finding the desired alloy efficiently. This paper is organized as follows: Section 4.3.2 introduces the overall approach, design space, and simulation details. Section 4.3.3 describes the AL runs exploring the effects of (a) different information acquisition functions and (b) different MD simulation times. It also discusses how uncertainty of the model evolves as the active learning workflow progresses. In Section 4.3.4 we summarize and draw final conclusions of the paper.

4.3.2 Problem statement

MPCAs are a new class of materials, which includes high entropy and complex concentrated alloys, where four of more elements are combined in nearly equal atomic percentages. Interest in these materials has grown due to several desirable properties. Properties such as high strength at elevated temperatures, radiation resistance [147], and high melting temperatures make some MPCAs attractive for high temperature and extreme applications [225]–[227]. Experimental determination of this melting temperature is challenging because of their complex processing [228] protocols, chemical reactions, and phase separation [229], [230].

For these reasons, using computational methods to determine the melting temperature of these alloys is highly desirable. Fortunately, melting temperature can be accurately obtained from first principles, and the first calculations for MPCAs are emerging [231].

The high-dimensional space of possible compositions prevents brute-force approaches even for relatively cheap simulations. Therefore, an efficient way to explore this space is essential. In this work, we pair AL with MD simulations to find MPCAs with the highest possible temperature in a model system. AL in the context of materials design seeks to reduce the number of experiments needed to find an optimal candidate out of a pool of untested contenders in an unexplored space. Within the AL scheme, predictions of the surrogate machine learning model, along with sample-wise uncertainties, are used to identify the next candidate to be tested, via acquisition functions. In this work, we make use of the FUELS (Forests with Uncertainty Estimates for Learning Sequentially) framework proposed by Ling et al. [176] based on random forests and a supervised machine learning algorithm.

Initial and candidate space

We explore 5-component FCC alloys incorporating Cr, Co, Cu, Fe, and Ni. Iterative AL models require an initial subset of entries to be evaluated to serve as the prior for the ML model. Our initial set of alloys was selected to be around the equiatomic composition $Cr_{20}Co_{20}Cu_{20}Fe_{20}Ni_{20}$. Alloys included in this initial set deviate by 10 at.% from this composition, such that each element composes 10 at.% to 30 at.% of the alloy. Using steps of 10 at. %, we get 39 data points. Melting temperatures were calculated via MD simulations, as described below, using the two-phase coexistence method. This initial set was the starting point for all of our AL runs.

For our exploration space, we relaxed the constraints from the initial set to allow more deviations from the equiatomic composition such that any of the elements is between 0 at.% and 50 at.%. This allows for exploration of binaries, ternaries and quaternaries. Using composition steps of 10 at. %, this results in 554 possible alloys that are not included in our initial subset. This grid spacing may seem coarse, but it is appropriate for an initial exploration and could be refined in a subsequent step if needed. This refinement was not

necessary as will be shown below. Our iterative AL model starts with a training step on our initial space, and the candidate space will serve as the pool from which it will be drawing candidates.

MD simulations of melting

To predict the melting temperature of FCC MPCAs, we made use of MD simulations and the two-phase coexistence method [88], [232]. To describe atomic interactions, we used a many-body, embedded-atom-method interatomic potential developed by Farkas and Caro [233]. The potential was designed for FCC Cr-Co-Cu-Fe-Ni alloys with near equiatomic compositions. Its parameterization focused on reproduction of elastic constants, vacancy formation energy, stacking fault energy, surface energies and relative phase stability to ensure that FCC is the stable structure for all five components. However, this potential has not been trained or tested for melting temperature calculations.

Several techniques exist to compute the melting temperature of materials using MD simulations. The most direct approach involves heating a crystalline sample until it melts and then cooling the melt until it recrystallizes. This results in overheating and undercooling, requiring a-posteriori corrections [234]. To avoid these issues, we predict melting temperatures using a phase coexistence method [88], [232]. In this method, a temperature at which the crystal and liquid phases of the material of interest coexist is established.

The first step is to generate an initial sample with a liquid and solid phase in contact, we achieve this with the use of two thermostats during the sample preparation step. We start from the four-atom FCC unit cell with a lattice parameter of 3.56 Å and replicate it 8 x 8 x 18 to create a periodic simulation cell. The third direction (z) is normal to the solid-liquid interface. Atom types are assigned based on the desired composition randomly to each lattice site. We note that this approach ignores possible short-range order in these alloys. This is an acceptable compromise since our goal is to demonstrate the coupling of active learning with MD simulations and, thus, the accuracy of the model predictions is of secondary importance. This random alloy assumption was also made by the original authors of the interatomic potential [233] and, in a subsequent paper, Farkas and Caro showed that the

short-range ordering effects were negligible with a similarly parametrized EAM potential for the Cr-Co-Al-Fe-Ni system.[235]. The entire structure is relaxed via energy minimization with respect to cell parameters and atomic positions. To create initial liquid/solid structure the cell is then divided in halves along z and two thermostats are used to create the two phases. The two regions are equilibrated at their initial temperatures for 10 ps using isothermal, isobaric MD simulations (NPT ensemble) with a Berendsen thermostat [236] and Nose-Hoover barostat [237]. This first step seeks to create an initial cell with both liquid and solid and the two temperatures need to chosen appropriately (the temperature of the liquid half should be higher than the melting temperature). A snapshot of the system after this initial step is shown in Figure 4.10. Clearly, this initial sample is not in equilibrium, and we follow the initial step with an isoenthalpic, isobaric (NPH) simulation where the liquid and solid regions are allowed to exchange energy freely and come to equilibrium. After steady state is achieved and the simulation cell has a uniform temperature, if the sample contains both liquid and solid the simulation temperature corresponds to the melting temperature of the system. The use of the NPH ensemble allows the temperature of the system to evolve towards the melting temperature. If the sample is initially too hot the liquid phase will grow at the expense of the solid, the heat of fusion will automatically cool the sample down towards the melting temperature; the reverse happens if the sample is initially too cold.

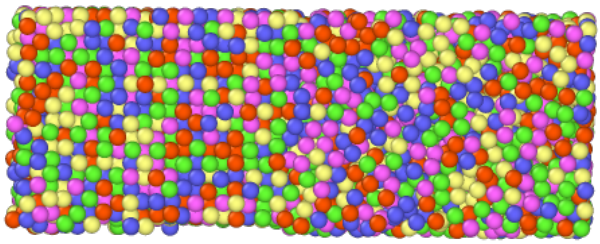

Figure 4.10. MD snapshot of the simulation cell divided into a liquid and solid region for an equiatomic alloy of Cr,Co,Cu,Fe and Ni. Each color represents a different element.

A successful run requires the coexistence of liquid and solid after steady state has been achieved. These conditions are checked by analyzing the atomic structure and the time evolution of the instantaneous temperature of the system. Phase fractions of solid and liquid

are calculated using the polyhedral template matching (PTM) algorithm from OVITO [86], [238]. Coexistence is defined as the fraction of liquid being between 35 and 65 at.% and the FCC crystal also between 35 and 65 at.% in the final snapshot of the simulation. We also analyze thermodynamic data from the last 60% of the simulation to assess steady state, we require the absolute value of the slope of the instantaneous temperature vs. time to be less than 1 K/ps. Finally, we also compute the 95% confidence interval on the temperature calculation. Melting temperatures for our initial set ran with 100ps of simulation time.

Uncertainties and noise in the MD predictions

The description of interatomic forces is at the heart of MD simulations and determines the predicted melting temperatures. The potential used in this work was not developed or widely tested for melting temperature calculations and we are unaware of any potential that has. This limits the accuracy of our melting temperature predictions. However, as stated above, our focus is on understanding the use of MD in ML workflows and not in accurate predictions of melting temperatures. Our work can be easily extended to MD simulations using higher-accuracy interatomic potentials [239]–[241] or density functional theory calculations [242]. Nevertheless, for completeness, Figure 4.11 compares the predicted melting temperatures with experimental values [243] for various alloys and single-elements [244]. Our simulations assume random FCC alloys since that is what the AL workflow uses, this does not match the experiments in all cases. We included experimental melting temperatures for single elements for comparison. All MD simulations all arrange structures as FCC. For iron, we show the predicted melting temperature of FCC Fe since the value is relevant for the high-melting temperature alloys found by the AL workflow. Interestingly, for Cobalt (Co), the MD simulation with this potential achieved coexistence by creating a solid BCC structure rather than FCC. We find that there is a consistent overestimation of melting temperatures, except for Cu and Cr.

It is important to note that the generation of the initial atomic structure is done stochastically, and temperature is obtained as the time-average of a fluctuating quantity. Thus, the resulting melting temperature from independent runs will vary due to sample-

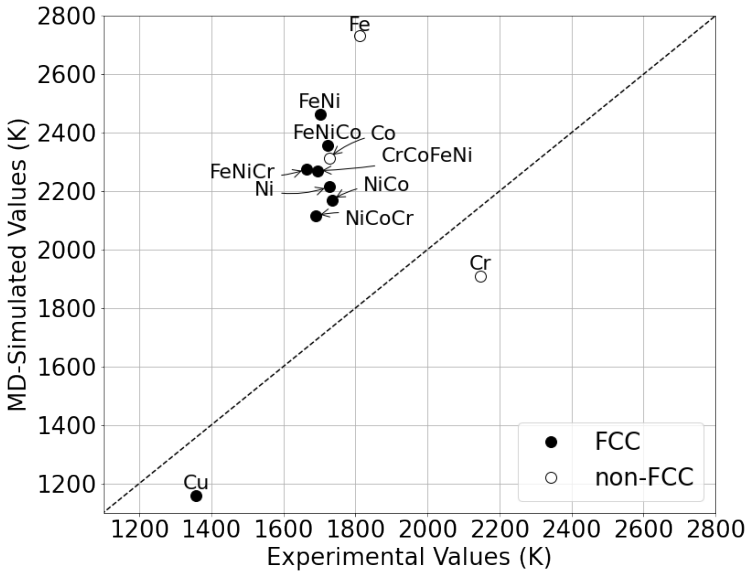

Figure 4.11. Comparison for MD simulated and experimental melting temperatures for alloys composed of elements included in the potential. Dashed line indicates a match between the values. Experimental values for MPCAs taken from Wu et al [243]. Experimental values for single-elements taken from NIST [244].Filled symbols represent alloys reported as FCC crystal structure, open symbols represent non-FCC crystal structures.

to-sample variability (which can be reduced increasing the simulation cell size) and time averaging of instantaneous temperature (which can be reduced by running longer simulations and also by increasing system size). One of the goals of this paper is to assess what level of noise can be tolerated in an AL workflow.

To quantify the variability in predicted melting temperatures, we chose one of our optimal alloy candidates. We performed 36 simulations with the same overall composition but independent atomic configurations and initial velocities. Figure 4.12 shows the resulting distributions of predicted melting temperatures for the three simulation times used in this study, the inset shows the mean and uncertainty estimates. Figure 4.12 shows that, as expected, repeated experiments result in mean values relatively independent of the simulation time (2470-2480 K) and reducing the simulation time results in broader distributions (remember that temperature is associated with the time-averaged kinetic energy per degree of freedom).

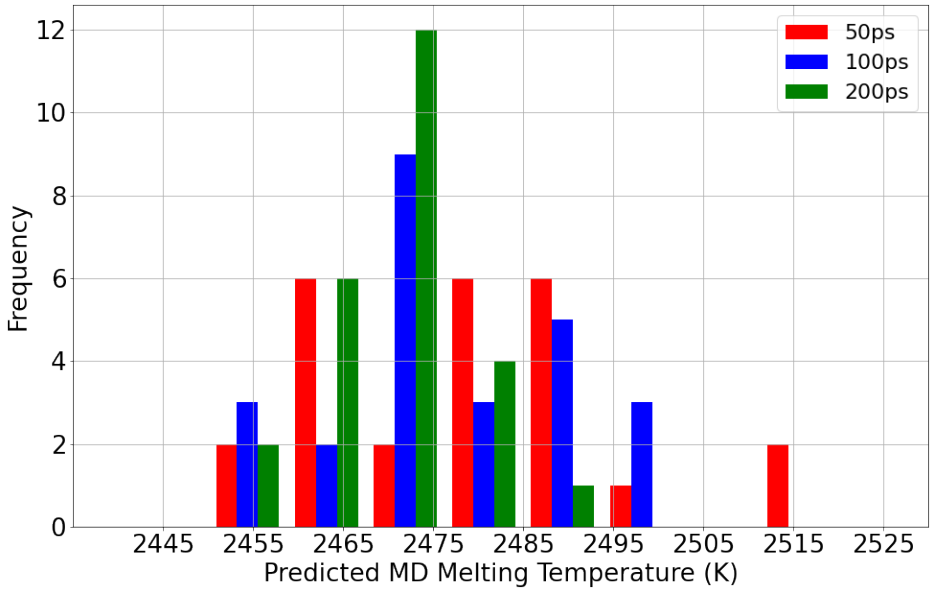

Figure 4.12. Distributions of MD simulated melting temperatures for a MPCA of 50% Fe and 50% Co at different simulation times. Distributions indicate a mean temperature around 2473 K and narrow down as the simulation time increases.

Table 4.2. Mean and uncertainty estimates for the distributions in Figure 4.12 of MD simulated melting temperatures for a MPCA of 50% Fe and 50% Co at different simulation times. These distributions were constructed by running 36 simulations with combinations of 6 different atomic arrangements and 6 different atomic velocities initializations.

Time Per Simulation (ps)	Mean MD Simulated T_M (K)	Standard Deviation MD Simulated T_M (K)
50	2479	17
100	2476	13
200	2470	8

Random forests and uncertainties

Several regression models, including neural networks, gaussian processes, and random forests have been used in AL workflows. Neural networks have stark advantages in image recognition and other high-dimensionality problems and have seen widespread adoption in a myriad of problems and disciplines. However, uncertainty quantification on their applications for regression and classification is remains an active area of research [186]. Random forests, on the other hand, are often preferred when dealing with tabulated data with well-defined and limited descriptors [245]. Random forests also have the ability to yield descriptor importance for interpretability [44] and there has been significant work on quantification of uncertainties [176], [184], [185].

In this work, we selected random forests as uncertainties are critical to make use of information acquisition functions within AL and to deal with noisy data present within our simulations. We use the lolopy library [187], [246] to develop RF models for melting temperature of MPCAs using composition and derived descriptors as inputs. The models provide an expectation value and sample-wise uncertainties that are fed to various acquisition functions to select the next experiment. Uncertainty estimates are obtained based on the recalibrated bootstrap forest variance as formulated in lolopy version 1.1.1 [246].

Materials Descriptors and Model Hyperparameters

For the RF models described in subsection 4.3.2 to predict melting temperature, it is vital to represent our alloys with descriptors that can serve as inputs. A brute-force approach would use composition as the sole inputs but, given the relative scarcity of the data in materials applications, enhancing the inputs with descriptors or features expected to correlate with the outputs is highly beneficial [247]. Previous work by Zhang et al. [248] in which they used a classification algorithm to predict the phases of alloys outlined a materials descriptor space and a genetic algorithm strategy to down select a subset of the pool of properties. The complete set included 70 properties based on the molar average value and mismatch value for each elemental property of the alloy. Their work showed that RFs outperform other models in the prediction task, but their algorithm selects subsets limited to four descriptors. Similarly,

work by Zhou et al. [249] shows that machine learning models can help the exploration of phase design of high entropy alloys. Their model uses 13 descriptors based on the average and standard deviation of properties like atomic radius, melting temperature, mixing enthalpy, mixing entropy, electronegativity, bulk modulus and valence electron concentration.

In this work, atomic properties for elements contained in the alloys are queried from properties available online on Pymatgen[250] and used to generate a unique fingerprint of each alloy using a weight-average rule. The final set of descriptors was based on the strongest correlation with melting temperature of our initial set, measured with Pearson correlation coefficients. Our feature set aligns well with the one proposed by Zhang et al. [248] as it includes properties such as: melting point, electronegativity, boiling temperature, atomic mass, atomic radii, and density. Another subset of the highly correlated descriptors relates to mechanical and thermal properties and includes Young's modulus, Poisson's ratio, hardness, and coefficient of thermal expansion. Finally, electrical resistivity was added to the descriptors as it showed a high correlation to melting temperature.

To test our initial model, we split the initial set in an 80/20 train/test split. In Figure 4.13 we present a parity plot of the melting temperatures predicted by our RF model and the melting temperatures obtained via MD simulations.

Given the small initial dataset, large deviations in MAE can occur due to the random selection of the testing set. To properly determine the model's accuracy, the dataset was shuffled, split, and rerun 30 times through the random forest. The average of the MAE for the RF predicted melting temperature with respect to the MD simulated temperature was 36.44 K with an uncertainty estimate of 12.2 K. A 10-fold cross validation analysis was used to create a normalized residual distribution to determine calibration of the uncertainty. Models with well calibrated uncertainties would produce a Gaussian distribution with a mean of zero and unit standard deviation as seen in Figure 4.14.

We optimized the number of estimators (trees) to 350 trees at the maximum depth of each estimator, using the mean absolute error (MAE) as a metric to measure the accuracy of our model predictions. Increasing the number of trees reduces variance, while increasing the maximum depth can help reduce bias. Our optimization stopped based on the idea that the larger quantity of either would improve the RF until we observed diminishing returns.

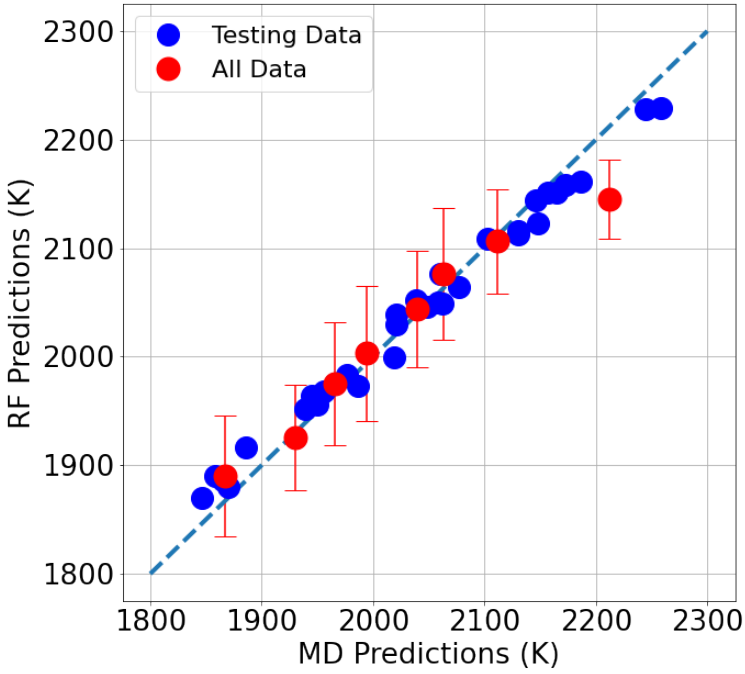

Figure 4.13. Comparison for MD simulated and RF predicted melting temperatures for the 39 MPCAs compositions in our initial set.

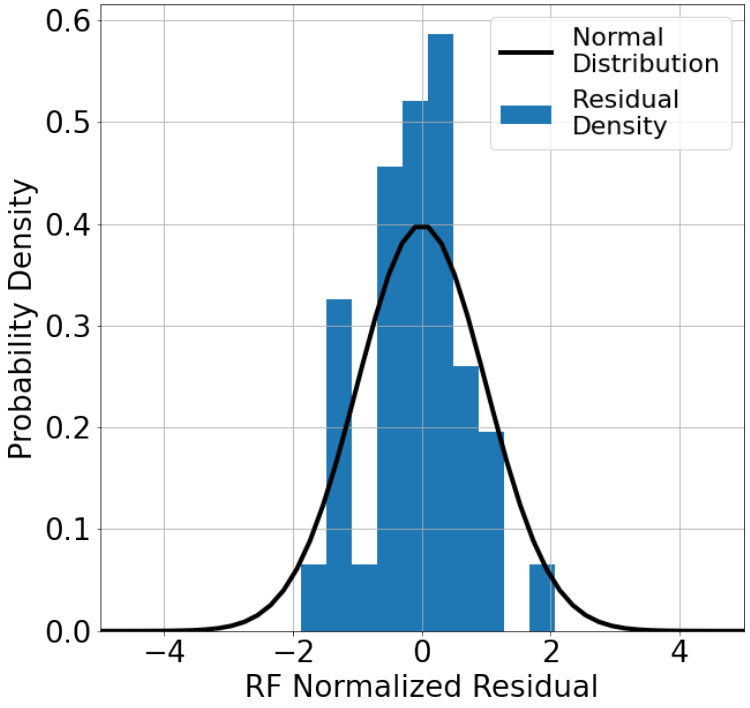

Figure 4.14. Probability densities of normalized residuals of RF model computed through tenfold cross-validation. Solid line is perfectly calibrated uncertainties.

Acquisition functions

Acquisition functions are mathematical expressions designed to select which of the possible tests (MD simulations in our case) to carry out next. This is done by considering the model predicted value and uncertainty estimate of the quantity of interest for all possible tests. For our case, we use the predicted value and uncertainty estimates provided by the RF model. Various acquisition functions have been proposed with different balances between exploitation and exploration. Exploration functions select areas of high uncertainty and, on the opposite end, functions based purely on exploitation select candidates with the highest expected values.

A purely exploitative approach would be maximizing the mean predicted value; this maximum mean (MM) function can be written as:

$$MM : x^* = argmax \;\; E[M(x_\mathrm{i})] \tag{4.6}$$

Such functions can easily get trapped in local maxima, and some degree of exploration is often desirable. Upper confidence bound (UCB) queries the sample with the maximum predicted value plus its uncertainty estimate. For this study, the adjustable parameter K is set to unity ($K = 1$).

$$UCB : x^* = argmax \;\; (E[M(x_\mathrm{i})] + K * \sigma[M(x_\mathrm{i})]) \tag{4.7}$$

Maximum likelihood of improvement (MLI) chooses the sample with the highest probability of surpassing the current best previously evaluated material. This probability can be computed as the Z-score for predictions that are normally distributed. Therefore, it is represented by the difference between the expected value of the prediction and the value of the current best case, x_{best}, over the uncertainty estimate.

$$MLI : x^* = argmax \;\; \frac{E[M(x_\mathrm{i})] - E[M(x_{best})]}{\sigma[M(x_\mathrm{i})]} \tag{4.8}$$

Maximum expected improvement (MEI) [251] works by modeling our knowledge of the prediction as a normal distribution. It uses the model's mean prediction and uncertainty

estimate to draw a normal probability density function at each point. The improvement can then be understood as the probability that the function at this value surpasses our current maximum. Intuitively, this relates to the tail of the distribution crossing our current maximum. In the absence of uncertainties, the function maximum expected improvement (MEI) [251], [252], reduces to maximizing the expected values. We note that the MM function was incorrectly labeled as MEI in Refs. [176], [206].

$$MEI : x^* = argmax \;\; \rho\left(E[M(x_{\mathrm{i}})] - E[M(x_{best})], \sigma[M(x_{\mathrm{i}})]\right)$$
$$with \;\; \rho\left(z, s\right) = \begin{cases} s\phi'(\frac{z}{s}) + z\phi(\frac{z}{s}) & s > 0 \\ max(z, 0) & s = 0 \end{cases} \tag{4.9}$$

Finally, maximum uncertainty (MU) is at the extreme end of exploration and only focuses on the candidates with the highest uncertainty in their predictions. MU has little incentive to find the top performer and finds its best use when trying to fine-tune ML models for surrogate-based optimization.

$$MU : x^* = argmax \;\; \sigma[M(x_{\mathrm{i}})] \tag{4.10}$$

In these Equations, x^* marks the composition selected by the function to be tested next. x_{i} are the possible experiments to run (our search space). x_{best} is the current best performer in the training set. $E[M(x_{\mathrm{i}})]$ is the expectation value of the prediction at point x. This expectation value is equal to $\frac{1}{N}\sum_{\mathrm{j}}^{n_T} t_{\mathrm{j}}(x)$ where $t_{\mathrm{j}}(x)$ is the prediction of tree j for point x and n_T is the total number of trees. Within MEI, ϕ represents the Gaussian cumulative distribution function, and ϕ' is the Gaussian probability density function. Finally, "arg max" (arguments of the maxima) operation returns the sample material where the function is maximized for the quantity of interest.

4.3.3 Active learning: exploring high melting temperature alloys

As described above, we started with knowledge of the melting temperature of 39 alloys and use AL to search for the alloy with the highest melting temperature in a 5-dimensional compositional space with 554 possible alloys. The performance of these acquisition functions was assessed from AL workflows with a budget of 40 experiments, mimicking time and resource constraints in real world applications. This budget is somewhat arbitrary and chosen based on prior active learning efforts. Here an experiment is equal to the convergence of a composition of an alloy which may require several simulations. As previously described, convergence is defined when the simulation outputs an alloy that reached temperature equilibrium and the fraction of atoms in each of the phases needs to account for 35% to 65% of the system. At each step of the iterative process, we use the RF models to evaluate each acquisition function over all unexplored alloys and select the alloy that maximizes the selected function. A new MD simulation is launched with that alloy and the resulting data is added to the data set and the processes is restarted.

The MD simulation is run for the desired time and convergence, or lack thereof, is determined with the criteria in Section 4.3.2. Simulation inputs are box length n=18, MD simulation time, and temperatures of the liquid and solid regions. Additional inputs include seeds for pseudo-random number generation for initialization of the initial random structure and atomic velocities. The initial temperatures for the liquid and solid regions from the melting temperature predicted by the RF (T_m) as follows: $T_{liquid,t=0} = 1.25\ T_m$; $T_{solid,t=0} = 0.5\ T_m$. If a converged melting temperature calculation is not reached, the initial temperatures are adjusted, and a new simulation launched. The initial temperatures are increased or decreased by 5% if the resulting structural analysis indicated little liquid or solid left, respectively. This process continues until convergence is achieved. Once convergence is reached the value will be added into the known set, the model will make new predictions, and the process will repeat until our design goal is reached or our experiment budget is exhausted.

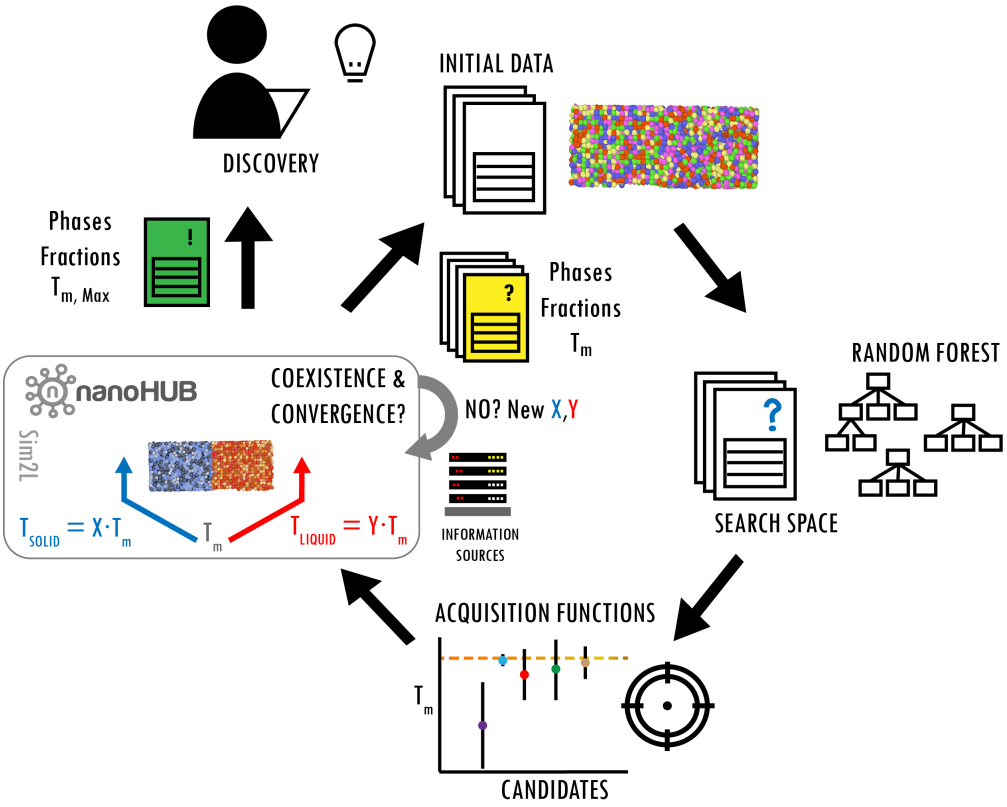

Figure 4.15. Schematic representation of the active learning iterative process outlined in this work. Initial data for random forest training is generated from MD simulations in a selected subspace. After training, an acquisition function is selected to provide a candidate composition for testing. Within the Simtool a new MD simulation is performed, characterized, and a secondary loop is performed to modify inputs to ensure final system convergence. Upon convergence, the data is collected, augmented to the training set, and the cycle continues until the budget is exhausted.

Active Learning results

To explore the combination of AL with MD simulations we tested six selection strategies (five information acquisition functions described in Section 4.3.2 and a random sampling baseline). To assess the effect the noise in the data we considered three simulation times 50, 100 and 200 ps. Each of the 18 runs used the same initial set as described in section 4.3.2.

Figures 4.16, 4.17 and 4.18 show the melting temperature, both predicted by the RF (filled symbols) and the one obtained from the subsequent MD simulation (open symbols), as a function of experiment number for the various acquisition functions and simulation times. The insets in Figures 4.16, 4.17 and 4.18, show the composition of the alloys with the highest predicted melting temperatures. Tables 4.3, 4.4 and 4.5 summarize the various AL outcomes: for the alloy with the highest melting temperature within the budget of 40 possible compositions for each case, we list the temperature predicted by the RF and the actual value obtained from the MD simulations, the uncertainty in the RF model, number of experiments required to locate the alloy with the highest melting temperature and its composition.

Table 4.3. Best results from information acquisition functions running simulations for 50 ps. RF prediction melting temperature and uncertainty taken when best composition selected for each acquisition function.

Acquisition Function	MD Simulated T_M (K)	RF Predicted T_M (K)	RF Uncertainty (K)	Experiment	Composition
MM	2472	2345	71	31	$Fe_{50}Ni_{50}$
UCB	2498	2416	63	10	$Co_{10}Fe_{50}Ni_{40}$
MLI	2502	2348	98	12	$Fe_{50}Ni_{50}$
MEI	2492	2403	82	12	$Co_{20}Fe_{50}Ni_{30}$
MU	2282	2137	118	-	$Co_{40}Fe_{20}Ni_{40}$
RAND	2382	2202	52	-	$Cu_{10}Fe_{50}Ni_{40}$

Quite interestingly, all acquisition functions that take exploitation into consideration (MLI, UCB, MEI, and MM) find the alloys with the highest melting temperatures in few experiments, regardless of the simulation time used. This is quite remarkable given the significant variability in output for the 50 ps long simulations. We find that highest melting temperature alloys consist of 50 at.% Fe with the remaining 50 at.% distributed between

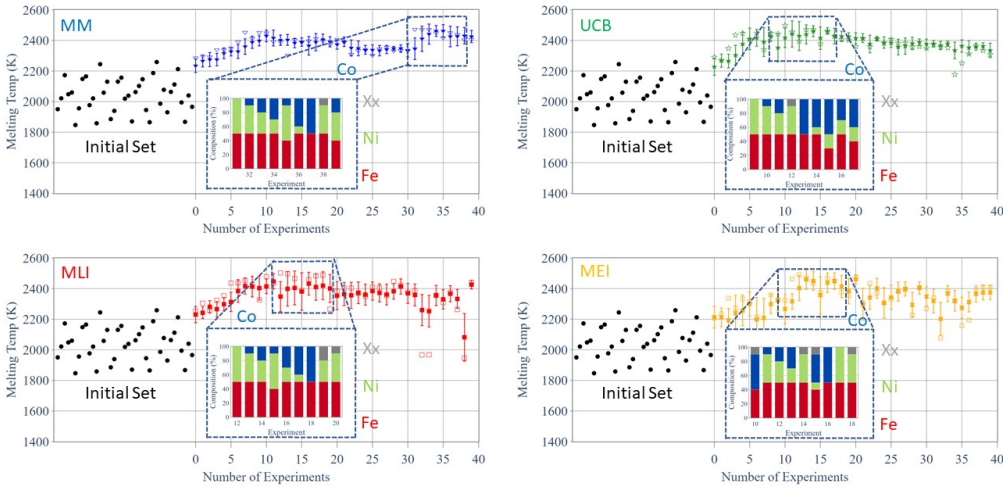

Figure 4.16. Performance of different acquisition functions in a 40-experiment budget AL run with 50 ps MD simulation time. All functions start from identical initial sets. Open symbols represent MD simulated melting temperatures. The filled symbols with error bars represent RF predicted melting temperatures. Xx represents other elements, Cu and Cr. The insert includes a close-up on the best performing MPCAs compositions. These compositions contain high quantities of Fe.

Table 4.4. Best results from information acquisition functions running simulations for 100 ps. RF prediction melting temperature and uncertainty taken when best composition selected for each acquisition function.

Acquisition Function	MD Simulated T_M	RF Predicted T_M	RF Uncertainty (K)	Experiment	Composition
MM	2489	2460	36	24	$Fe_{50}Ni_{50}$
UCB	2521	2422	78	15	$Fe_{50}Ni_{50}$
MLI	2539	2455	60	15	$Fe_{50}Ni_{50}$
MEI	2486	2415	87	12	$Co_{30}Fe_{50}Ni_{20}$
MU	2299	2160	117	-	$Co_{20}Cu_{20}Fe_{50}Ni_{10}$
RAND	2382	2202	52	-	$Cu_{10}Fe_{50}Ni_{40}$

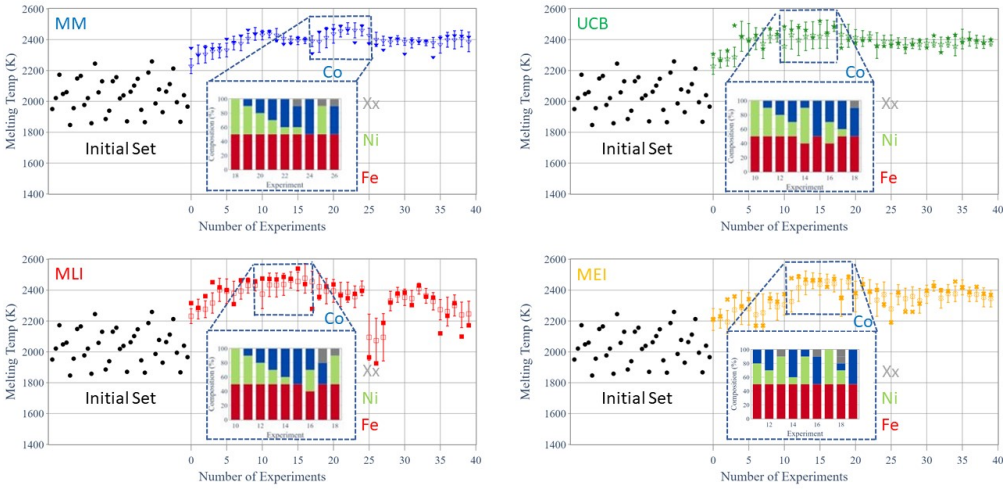

Figure 4.17. Performance of different acquisition functions in a 40-experiment budget AL run with 100 ps MD simulation time. All functions start from identical initial sets. Open symbols represent MD simulated melting temperatures. The filled symbols with error bars represent RF predicted melting temperatures. Xx represents other elements, Cu and Cr. The insert includes a close-up on the best performing MPCAs compositions. These compositions contain high quantities of Fe.

Table 4.5. Best results from information acquisition functions running simulations for 200 ps. RF prediction melting temperature and uncertainty taken when best composition selected for each acquisition function.

Acquisition Function	MD Simulated T_M (K)	RF Predicted T_M (K)	RF Uncertainty (K)	Experiment	Composition
MM	2485	2461	29	27	$Fe_{50}Ni_{50}$
UCB	2487	2417	81	17	$Co_{30}Fe_{50}Ni_{20}$
MLI	2480	2413	85	14	$Co_{10}Fe_{50}Ni_{40}$
MEI	2476	2418	91	9	$Co_{10}Fe_{50}Ni_{40}$
MU	2285	2160	118	-	$Co_{40}Cu_{10}Fe_{30}Ni_{20}$
RAND	2432	2360	80	-	$Co_{40}Fe_{40}Ni_{20}$

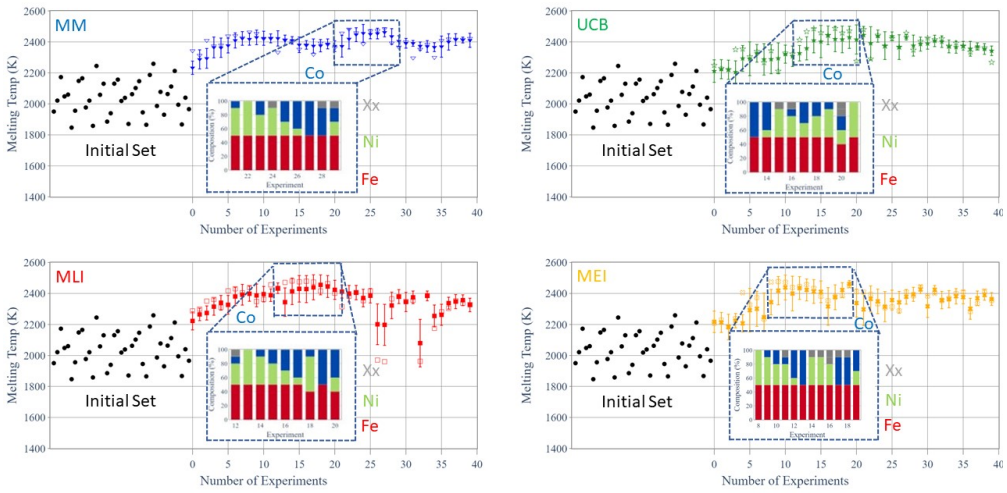

Figure 4.18. Performance of different acquisition functions in a 40-experiment budget AL run with 200 ps MD simulation time. All functions start from identical initial sets. Open symbols represent MD simulated melting temperatures. The filled symbols with error bars represent RF predicted melting temperatures. Xx represents other elements, Cu and Cr. The insert includes a close-up on the best performing MPCAs compositions. These compositions contain high quantities of Fe.

Ni and Co; this set of alloys are predicted to have nearly identical melting temperatures, see Figure 4.19. As expected, the MU function and random exploration do not explore high-melting temperature compositions within the 40 experiment budget. The graph seen for MM, UCB, MLI, and MEI has been made for MU and random and are in the SI as S3, S4 of the original published work [138].

We find that RF prediction for the highest melting temperature alloy tends to underestimate the MD result, see Tables 4.3, 4.4 and 4.5, this is mostly true during the entire active learning workflows themselves, see Figures 4.16, 4.17 and 4.18. This can be attributed to the nature of the RF models, which are based on decision trees that saturate when extrapolating to values outside of their training.

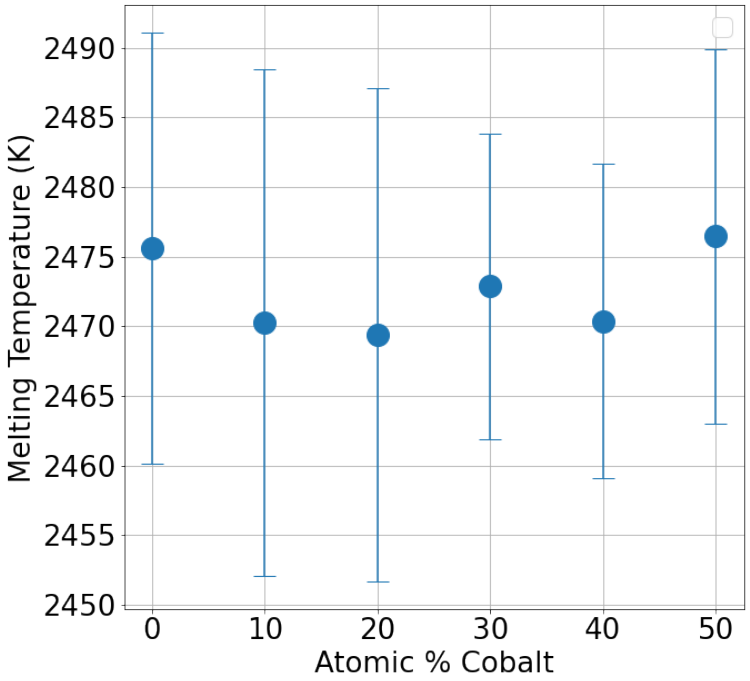

Figure 4.19. Predicted melting temperatures for high Fe ratio compounds. These alloys proved to be the highest found melting point for this interatomic potential in composition: $Fe_{50}Co_xNi_{50-x}$. Mean values and uncertainty estimates over 36 independent runs are shown.

In all our tests, we find that the acquisition functions that combine both exploration and exploitation (UCB, MLI, MEI) consistently outperform the purely exploitative function (MM), which can spend resources in local maxima. As seen in Tables 4.3-4.5 UCB, MLI,

and MEI find the optimal composition within less than 17 experiments while MM found it past 27. While in our example MM finds the family of best performer candidates, it takes approximately twice as many experiments as the other functions.

Noisy data and RF uncertainty

We find that the AL workflow can find best performer alloys in approximately the same number of iterations regardless of the simulation time use. Between 9 and 17 simulations were required for the UCB, MEI, and MLI strategies. As mentioned above, this is despite the significant level of noise, especially in the 50 ps runs. We also find that while longer MD simulations do not result in a reduction in the number of experiments required to achieve the design goal, it results in improved accuracy of the RF prediction. Tables 4.3, 4.4 and 4.5 show that the workflow with 50 ps runs significantly underestimates the temperature of the optimal alloy and the uncertainty estimates are overly optimistic.

To better understand how the models gain knowledge in the presence of noise we explore the evolution of the model predictions for $Fe_{50}Co_xNi_{50-x}$ alloys, the top performers, at various stages during the AL process. Figure 4.20 shows mean and uncertainty estimate predictions at four stages of an MLI workflow for the 50 ps case; results for 100 and 200 ps simulations show similar trends and are included in the SI as Figures S5, S6. Open circles denote MD simulations not already explored at the corresponding cycle and filled circles represent the same values for compositions that have been explored. We note that noise in the data affects both mean predictions and uncertainty estimates and can thus have non-trivial effects on AL decisions. Figure 4.20 shows that even with only the initial 39 datapoints (Iteration 0), the model predicts relatively high melting temperature for these alloys, comparable to the highest melting temperature materials in the initial set, see Figure 4.16. As the AL workflow with MLI explore compositions, the mean prediction for the selected family improves from 4.20 (a) to (b) as the mean moves towards the true MD temperatures of the alloys. At step 12, Fig. 4.20 (d), the model predicts high melting temperatures for the alloys with x=0, 10, 20 % and x=0 is selected for simulation. In subsequent steps, the MLI algorithm hones in the rest of the family and improves the model accuracy. When comparing the results for

50 ps runs with those of 100 and 200 ps (SI Figures S5 and S6) we observe similar paths to the optimal alloys with a decrease in uncertainties for longer simulation times. This is because the variability in the MD simulations is relatively small compared to the differences in melting temperatures across alloys.

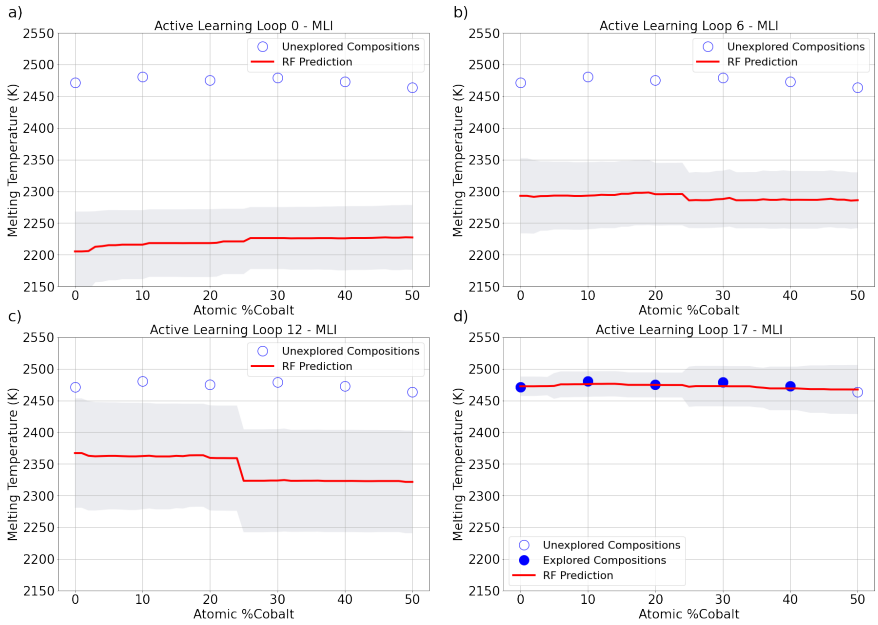

Figure 4.20. Red lines represent the predicted mean melting temperature and shaded region represents the uncertainty estimates for predictions on compositions as a function of Co content in $Fe_{50}Co_xNi_{50-x}$ at various stages of the AL workflow. Whereby x starts at 0% a goes to 50% with 1% step size. Results correspond to the MLI function with a 50 ps MD simulation time. Open circles represent MD-simulated values unknown to the model at the time and filled symbols represent the values included in the model.

4.3.4 Conclusions

The combination of machine learning and physics based simulations holds great promise to guide the experimental search and accelerate the development of novel materials. The use of surrogate-based models that can make use of computational simulations instead of experiments can save resources and time in design cycles. We find the molecular dynamics simulations, even under significant stochastic noise, can be a powerful tool to be used in active learning efforts.

We developed an AL workflow with MD simulations for the discovery of MPCAs materials for high-temperature applications. Starting from a limited initial set, we created a set of descriptors and a RF model to use as a surrogate for the optimization of our AL model. We determined that RF models are capable of handling multi-dimensional spaces with a limited initial dataset. We paired our RF model with a cloud-based simulation environment for automatic querying of melting temperatures from MD simulations with an interatomic potential for FCC alloys. We analyzed the effect of MD simulation time in the overall uncertainty quantification when running such calculations and showed that it has an important effect in capturing the sample-to-sample variability caused by the stochastic nature of MD. We compared the performance of four different acquisition functions in a closed, but high-dimensional, search space. We found that the best performing acquisition functions, MLI, MEI, and UCB, are combinations of exploitation and exploration.

The workflow we developed can be adapted toward the optimization of other material properties and to the use of other sources of information other than computational simulations. We provide the analysis and all other relevant information in a public SimTool on nanoHUB.org.

Continuing to explore deeper into our models we will use a separate database used in other work for feature explanation and model development. With the simplicity of random forests comes the ability to create unique explainers and tools tailored for both education and research purposes.

4.4 Feature selection and explanation for high entropy alloy strength

While considering a different subset of data, and applied labels for machine learning, the process for predicting the strength of an alloy is analogous to a model for melting point. The same principles for chemical based featurization will be applied, however with a more rigorous approach for feature selection and inspection. While not available in this work, the methods that follow have been applied to the models included in Sec. 4.2 and Sec. 4.3 for further investigation.

Complex concentrated alloys (RCCAs) and the sub-set of HEA are a relatively new class of materials attractive for a wide range applications [139], [253]–[255]. Some refractory CCAs exceed state of the art Ni-based superalloys in high-temperature strength and while the oxidation resistance remains to be improved these materials are of significant interest in applications such as turbines and aerospace re-entry materials. Significant efforts have focused on these class of materials and applications in particular with regard to engineering the strength [256]–[259]. The discovery of optimal CCAs for high-temperature applications is hindered by several factors. i) the high dimensionality of the design space (list composition, processing, testing, from the proposal). ii) data scarcity and the high cost and time involved in full-scale, high-temperature testing, iii) The lack of physics-based predictive models for the quantities of interest [260].

This section highlights RF models for the strength of CCAs that integrate both experimental results and theoretical information as input descriptors. We perform a systematic study of possible descriptors and find, not surprisingly, that hardness is the most important one. Also important, are components derived from Thermocalc (TC). Specifically, the phase change temperatures (liquidus and solidus), and the system enthalpy. To understand the strengths and weaknesses of the RF models as well as possible shortcomings in the existing data we use local explanation tools [261] understand how individual descriptors contribute to the total predicted strength in specific alloys. Understanding how predictions are made in cases when the models underestimate or overestimate strength, i.e. outlier cases, provides insight into pathfinding suggestions for feature selection and development for alloy exploration.

4.4.1 Initial Database

From a collected and openly published database of CCA mechanical properties [262], we extract hardness, strength, grain size, and processing conditions of CCAs. The database contains over 1600 entries; of these, 1100 contain strength at various temperatures. Using applied domain knowledge we will craft subsets of features manually, and use available data where alloys overlap in their experimental publishing. These descriptors can be generally categorized into three groups: i) surrogate experimental measurements, ii) physics-based

models, and iii) periodic table data that can provide information atomic properties of the constituents. Section 4.4.2 discusses features and their selection in detail, here we explore the available of surrogate experiments or descriptors associated with each of the materials of interest. One of the best surrogates for experimental strength is the experimental hardness. These methods are non-destructive to the alloy, and they scale linearly with the strength. Of the 380 entries with room temperature strength, only 160 have corresponding hardness values. Another set of features that could be used are the grain size, processing, and phase information of the material. Unfortunately, if we were to limit ourselves to a dense dataset with these features we would be limited to only 74 entries for training.

Settling for a database of 160 initial alloys containing both hardness and strength, we begin populating with additional physics based descriptors and periodic table assessments. To provide information about the base alloy will use single crystal strength model approximations used in literature [263]–[265]. The initial datasets for points containing grain size & strength, the correlation between single-crystal strength modeling and strength, and the hardness and strength are shown in Fig. 4.21.

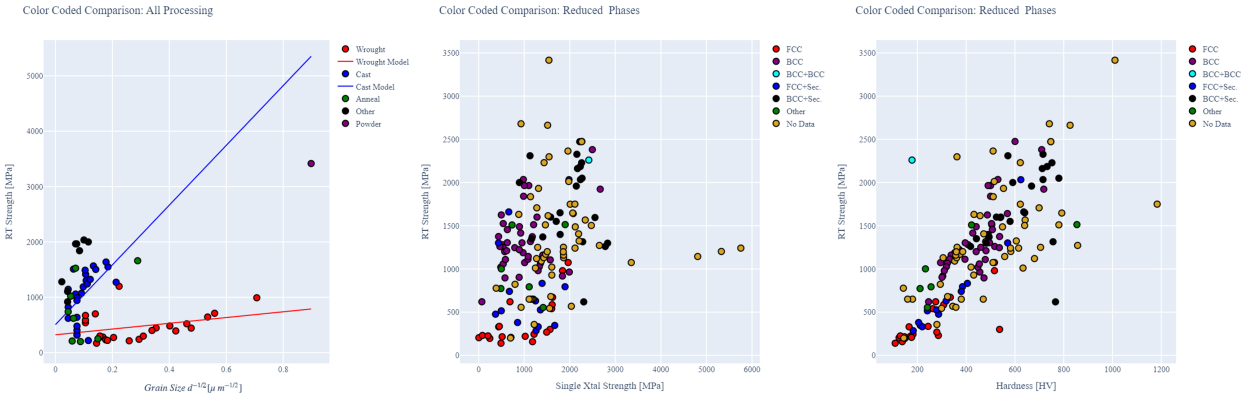

Figure 4.21. Initial database visualization of CCA strength. Properties shown are available data for grain size [left], single crystal strength modeling [middle], and experimental hardness [right].

4.4.2 Feature Generation

Significant work has been done on modeling of hardness in CCAs. These methods range from complex neural networks to more approximate linear methods [266]. In each of these a growing list of successful features have been teased out. These features for hardness include: elastic constants, atomic radii, entropy of mixing, valence electron concentration, and weighted parameters associated with each [267], [268]. These values have been taken as rules of mixtures, ranges, and static modes in previous work. Here we employ the same sets of features as in previous literature, but now to model the yield strength of CCAs. Experimental hardness is used as a feature in addition to the remaining set. Thermocalc predictions of phase stability, volume fraction of phases, and predicted solidus and liquidus temperatures were also used as features. For each phase present we employ a one-hot encoding scheme to differentiate between FCC, BCC, and multiphase alloys. This extends on previous efforts by Rickman et al. [266] where they only consider single or multiphase categories. For non-existent phase entries we use the phase information collected from Thermocalc. Phase composition profiles derived from Thermocalc were assessed at 0.3-0.9T_m and compared to existing data with phase information present. It was found that at 0.9T_m Thermocalc has good agreement of phase information with room temperature experiments. Therefore absent phase information was supplemented with 0.9T_m Thermocalc phase descriptions.

Pearson correlation analysis [269] was used as a first assessment of the list of descriptors selected. The full list of generated are shown below in Fig. 4.22 and the correlation between the entire set of features is included. Formalisms for the generation of each feature and their symbolic representation are provided in Table. 4.6.

4.4.3 Model Optimization

Given the small set of data we allow our random forest to extend well beyond the depth of the dataset to 300 trees. While this would generally saturate MAE reduction beyond 250 trees, the larger network helps calibrate the uncertainty measurement [188]. As is common practice we divide our data intro training and testing sets of 80 and 20% respectively. To avoid the variability associated with the specific choice of testing set, we quantify the discrepancy

Table 4.6. Features for RCCA strength modeling

Symbol	Formalism	Description
$\overline{\Delta S_{mix}}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}\Delta S_{mix,\mathrm{i}}$	Avg. Entropy of Mixing
$\overline{\rho}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}\rho_{\mathrm{i}}$	Avg. Density
$\overline{T_m}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}T_{m,\mathrm{i}}$	Avg. Melting Temp.
$\overline{r_{at}}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}r_{at,\mathrm{i}}$	Avg. Atomic Radii
$\overline{Y}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}Y_{\mathrm{i}}$	Avg. Young's Modulii
$\overline{G}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}G_{\mathrm{i}}$	Avg. Shear Modulii
$\overline{K}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}K_{\mathrm{i}}$	Avg. Bulk Modulii
$\overline{VEC}$	$\Sigma_{\mathrm{i}=1}^{N}\ c_{\mathrm{i}}VEC_{\mathrm{i}}$	Avg. Valence e^- Conc.
$\Delta\rho$	$\rho_{\mathrm{i},max} - \rho_{\mathrm{i},min}$	Range of Density
ΔT_m	$T_{m,\mathrm{i},max} - T_{m,\mathrm{i},min}$	Range of Melting Temp.
Δr_{at}	$r_{at,\mathrm{i},max} - r_{at,\mathrm{i},min}$	Range of Atomic Radii
ΔY	$Y_{\mathrm{i},max} - Y_{\mathrm{i},min}$	Range of Young's Modulii
ΔG	$G_{\mathrm{i},max} - G_{\mathrm{i},min}$	Range of Shear Modulii
ΔK	$K_{\mathrm{i},max} - K_{\mathrm{i},min}$	Range of Bulk Modulii
ΔVEC	$VEC_{\mathrm{i},max} - VEC_{\mathrm{i},min}$	Range of Valence e^- Conc.
δY	$\sqrt{\Sigma_{\mathrm{i}=1}^{N} c_{\mathrm{i}}(1-\frac{Y_{\mathrm{i}}}{\overline{Y}})^2}$	Asymmetry of Young's Modulii
δG	$\sqrt{\Sigma_{\mathrm{i}=1}^{N} c_{\mathrm{i}}(1-\frac{G_{\mathrm{i}}}{\overline{G}})^2}$	Asymmetry of Shear Modulii
δK	$\sqrt{\Sigma_{\mathrm{i}=1}^{N} c_{\mathrm{i}}(1-\frac{K_{\mathrm{i}}}{\overline{K}})^2}$	Asymmetry of Bulk Modulii
δr_{at}	$\sqrt{\Sigma_{\mathrm{i}=1}^{N} c_{\mathrm{i}}(1-\frac{r_{at,\mathrm{i}}}{\overline{r_{at}}})^2}$	Asymmetry of Atomic Radii
$T_{liquidus}$		Thermocalc liquidus temperature
$T_{solidus}$		Thermocalc solidus temperature
ρ_{TC}		Thermocalc alloy density
ΔH_{TC}		Thermocalc alloy enthalpy
ΔS_{TC}		Thermocalc alloy entropy
ΔG_{TC}		Thermocalc alloy Gibb's free energy
V_{misfit}	$\Sigma_{\mathrm{i}=1}^{N}\Delta V_{\mathrm{i}}^2\ \|\|\ \Delta V = \overline{V} - V_{\mathrm{i}}$	Atomic Volume Misfit
σ_{sx}	$\tau_y(T,\dot{\epsilon}) = \tau_{y,0}\mathrm{e}xp\left(-\frac{1}{0.51}\frac{kT}{\Delta E_b}ln\frac{\dot{\epsilon_0}}{\dot{\epsilon}}\right)$	Solid Solution Strength
[1 0 0 0 0]/[0 1 0 0 0] etc.		Reduced Phase One-Hot-Encoding (O.H.E.)
[1 0 0 0 0]/[0 1 0 0 0]		Processing O.H.E.
HV		Experimental Vicker's Hardness

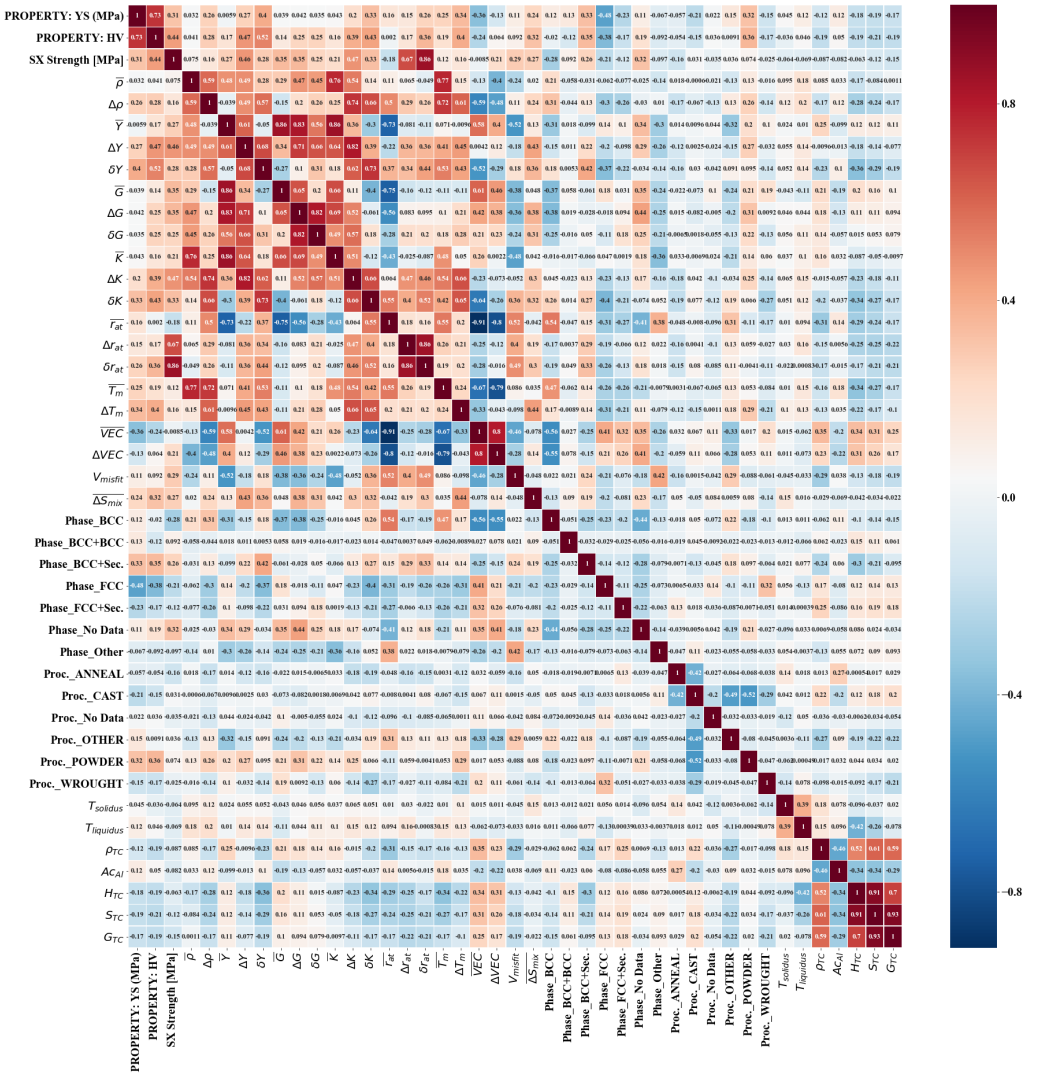

Figure 4.22. Pearson correlation plot for features used in strength modeling of CCAs

of our model using a method akin to k-fold cross validation. We use nested cross validation to report the effect of feature selection. For each set of features we select the data is shuffled and split into different testing and training sets 10 times. The MAE for each shuffling is aggregated together for a mean of means.

First, we will seek to find the optimal number of selected features to model strength in CCAs. To optimize the number of features one could naively perform a random selected of increasing number of features until a complexity was achieved in the random forest. owever, this method can be improved significantly by using the weights of the Pearson Correlation as adjustments to purely random selection. For any given number of features selected if a feature has a higher Pearson Correlation there is a higher likelihood that it is chosen as a feature to train. Fig. 4.23a shows the result of optimizing the list of selected features from before. In blue we show for increasing N features a reduction in aggregate MAE, but our overall error remains high. In green we show models that were trained on values randomly selected with weights associated with Pearson Correlations. In red we show any point where hardness was selected, random or weighted. We can see quite clearly that of these points models that contain hardness significantly improve the model, and create the entire boundary of the lower Pareto front.

4.4.4 Model Explanation

In this section, we seek to understand how the RF model makes decisions for specific materials. This not only can help gain confidence in the model but also understand the role of descriptors. We will do this using SHAP coefficients based on application of Game Theory postulated by Lloyd Shapely in the 1950s [270]. In principle, if we consider our model to predict some outcome based on features, we can allocate costs and rewards to removing or adding certain features from the strength model.

Assume that we have N features that are fed to the model and define a subset S of them. Let $\nu(S)$ be the value predicted by the model for material S. When a feature is added to the model we give them contribution $\nu(S \cup \{\mathrm{i}\}) - \nu(S)$. We introduce the features one at

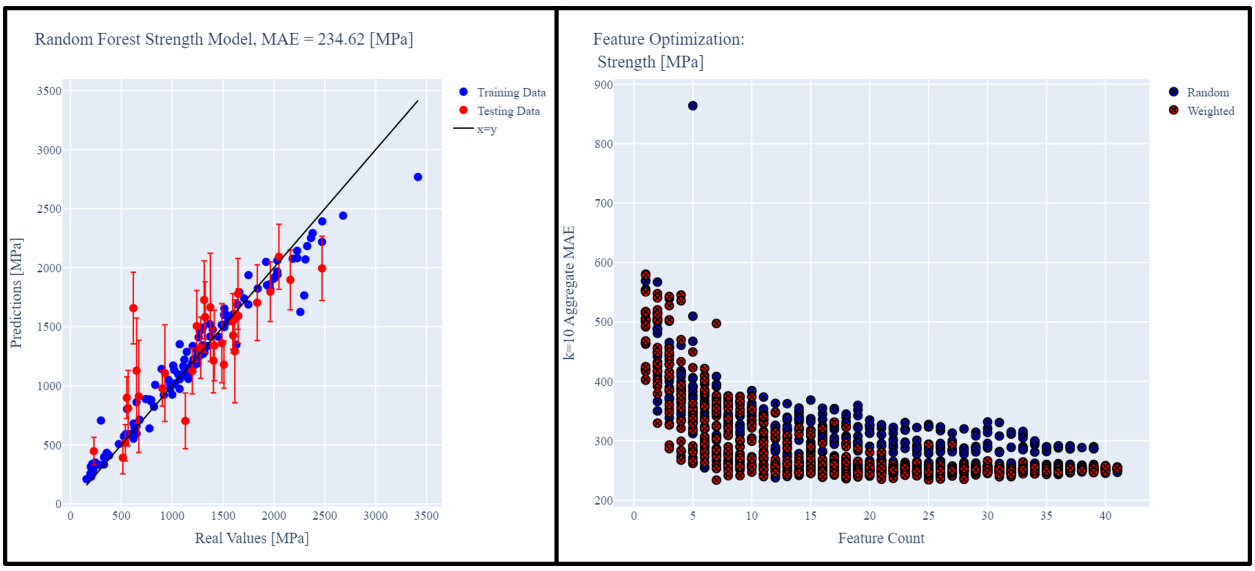

Figure 4.23. [Left] Individual sample seed parity figure with testing set uncertainties. [Right] Aggregate MAE scores with increase in feature count for random (blue) and weighted (red) sampling.

a time, and in each combination possible, while evaluating the gain or loss in model validity at each step. For a given feature, i, their contribution to the model can be measured by:

$$\phi_{\mathrm{i}}(\nu) = \Sigma \frac{|S|!(N - |S| - 1)!)}{N!}(\nu(S \cup \{\mathrm{i}\} - \nu(S)) \tag{4.11}$$

Here we also characterize the contribution to the model by four Axioms:

Axiom 1: Efficiency: The sum of the Shapely values of all agents equals the value of the total coalition.

Axiom 2: Symmetry: All features have a fair chance to join the game.

Axiom 3: Dummy. If feature i contributes nothing to any coalition S, then the contribution of feature i is zero, i.e. $\phi_{\mathrm{i}}(\nu) = 0$

Axiom 4: Additivity. For any pair of models v, w: $phi(v + w) = \phi(v) + \phi(w)$, where $(v + w)(S) = v(S) + w(S)$ for all S

In recent work, authors have built on the idea of game theory and Shapely coefficients to develop a robust method for feature and model interpretation [261]. Shapely additive explanations (SHAP) as a local explanation for decision paths in random forests, and reward allocation for individual features. Shapely values are computed based on a conditional expectation of the model output:

$$f_x(S) = E[f(X)|do(X_S = x_S)], \tag{4.12}$$

and for each perturbation to feature input we attribute the change. For each ordering of feature input the change is averaged to collect a global contribution. In Fig. 4.24 we show a selection of the highest contributing collected SHAP values and their correlation to the model output. Features are ranked according to *Feature Importance*, the metric of how much impact the feature had in magnitude for each addition to the model. *Horizontal importance* is characterized by the magnitude of the SHAP coefficient. Red coloring to the right of the axis indicates a high impact towards positive SHAP values, indicating a positive correlation towards model output. A red coloring to the left of the axis indicates a negative correlation between feature and model output. As expected from literature and intuition,

the parameters for hardness, VEC, density, and radii delta function are among the most contributing parameters.

For each feature we can analyze the impact of not just the grouped feature, but the individual SHAP coefficient for each local feature value. If we first consider hardness, we see that for an increase in hardness a linear response to the output strength is measured. This returns us to the logical conclusion that hardness is a linear approximator of experimental strength. This information can be leveraged for developing high temperature strength models based on elevated temperature hardness measurements. Unlike hardness, with increasing VEC we do not see an increase in SHAP value, or expected reward on the model expectation value. Rather, we see a region where low VEC contribute positively to the strength of a model, while high concentrations exhibit a stark dropoff in strength as well as hardness. This is consistent with literature on studies regarding VEC in CCAs [271]. For features such as the Thermocalc phase supplement we see a favorable model outcome when multi-phase systems are present versus single phase, and at high densities our alloy will exhibit higher strength. Delta functions to find asymmetry in atomic radii or elastic constants can be useful for initial screening metrics to maximize this function.

We also show that through individual feature explanation we can ascertain why certain alloys become outliers in the hardness vs. strength relationship. Here we isolate three regions of the data to explore the effect of features on total model output. Region I, in green, will be the high strength & high hardness family, Region II, in dark blue, the low strength & high hardness family, and Region III, in red, the high strength & low hardness family. Two alloys from each family have been selected with a different icon, and their corresponding waterfall plots for a random forest explanation are included.

Beginning with Region I, we track the experimental value of the database, and the random forest prediction. While the highest strength alloy in the database, CrMoNbTaVW, is slightly underestimated the contributions from all of the features are highly positive, with hardness pushing it furthest to the right from the model expectation value, or the database average.

Region II highlights the battle between the high hardness value for the alloy, and its negative impact on strength due to a low liquidus temperature, low atomic mismatch,

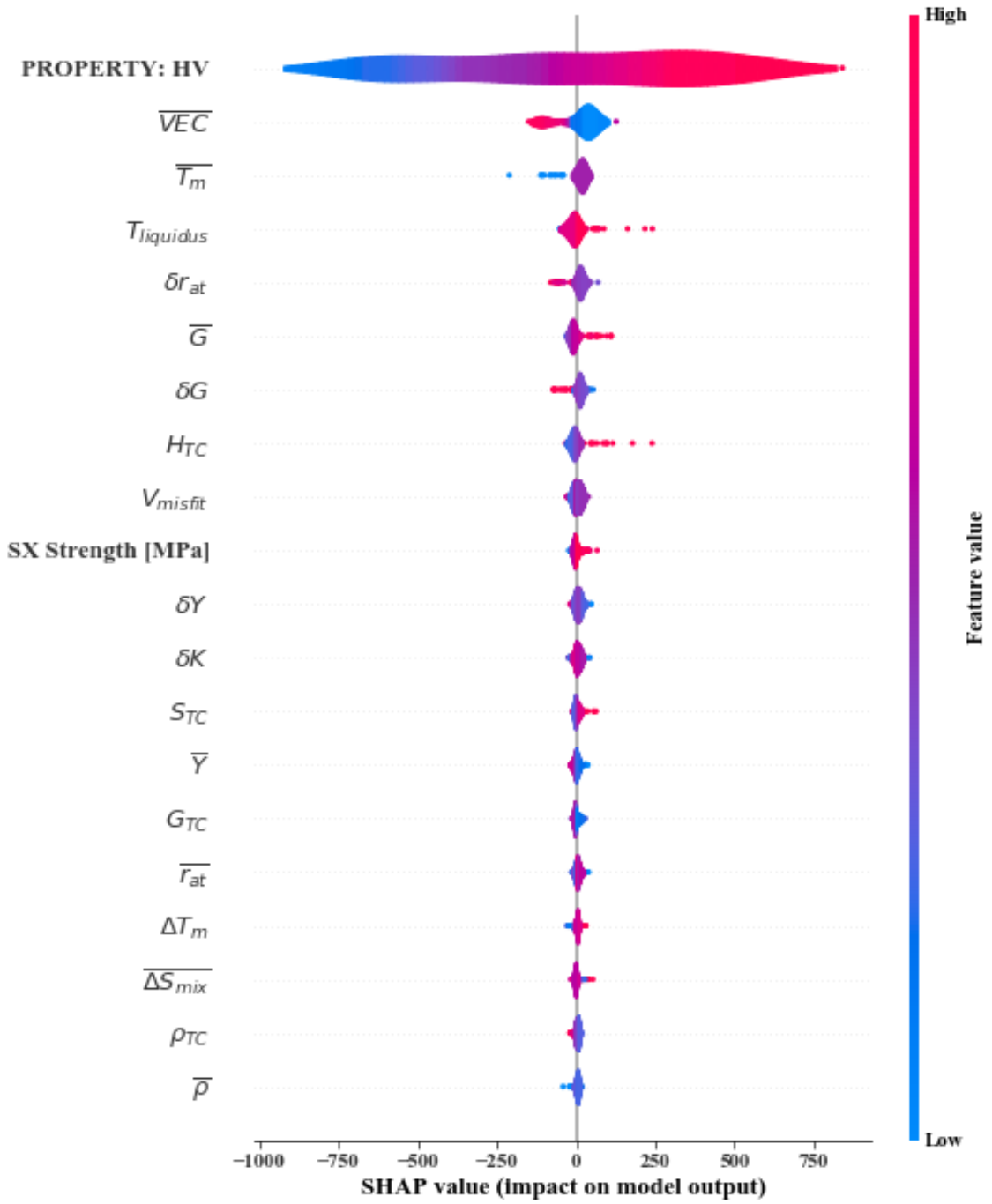

Figure 4.24. SHAP correlation values for strength model features. Red coloring indicates positive correlation, and a cluster to the right of the axis indications a positive impact on the model.

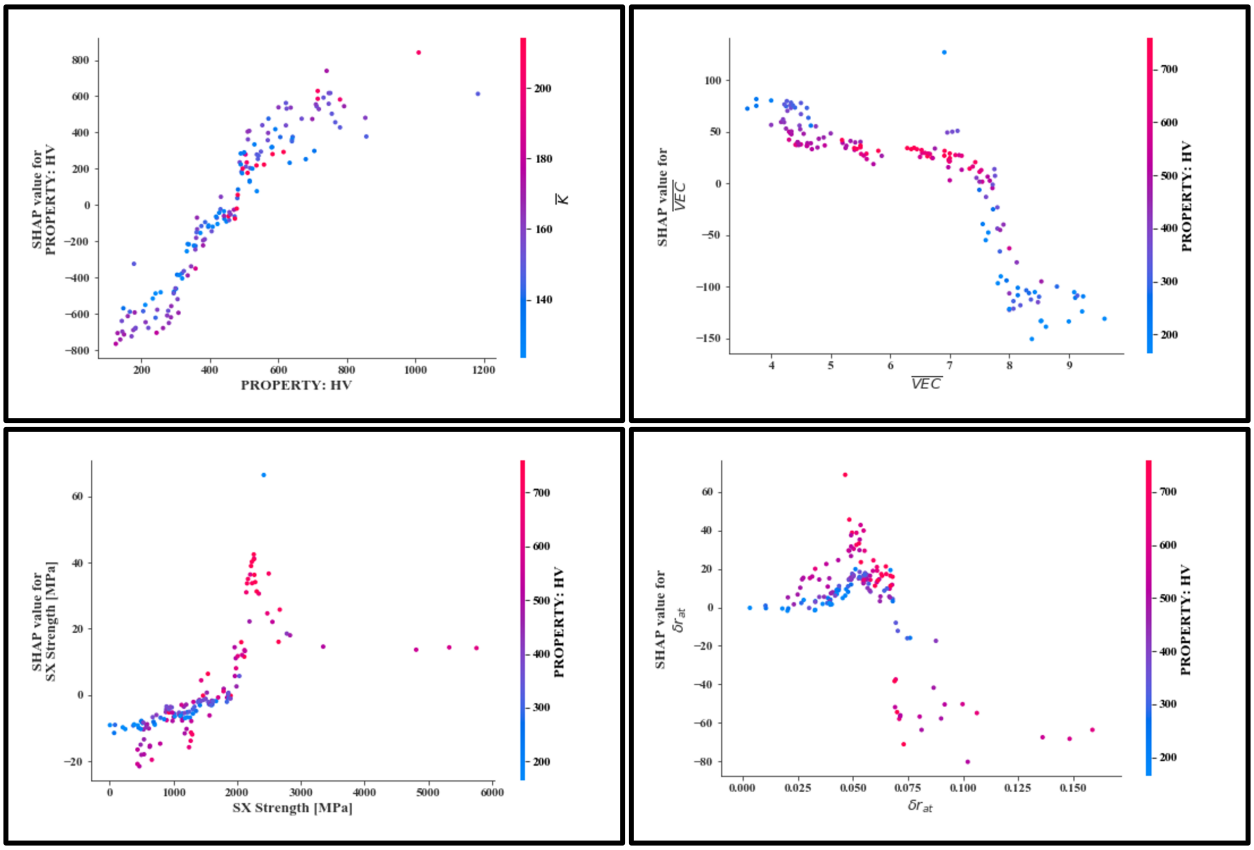

Figure 4.25. SHAP score for individual points for selected features. Correlation to the feature's strongest partner and its input values are shown in the color map. a) Curtin strengthening model b) Radii asymmetry function c) VEC d) Hardness

or a VEC approaching the steep dropoff effect. More interesting though is the effect of phase information in alloys that exhibit low hardness, but an inversely though high strength. Unfortunately for the $CrFeNiTi_{0.2}$ alloy no phase information was present at the time of this database modeling exercise. The model is unable to predict the high strength property as it struggles with the information of low hardness. However, in the case of $AlCoCrFeNiTi_{0.5}$ a two-phase equilibria of separate BCC phases were reported. This coupled with a low VEC and high enthalpic contribution from TC predictions allows for the model to come much closer to the value than if the phase information had not been present.

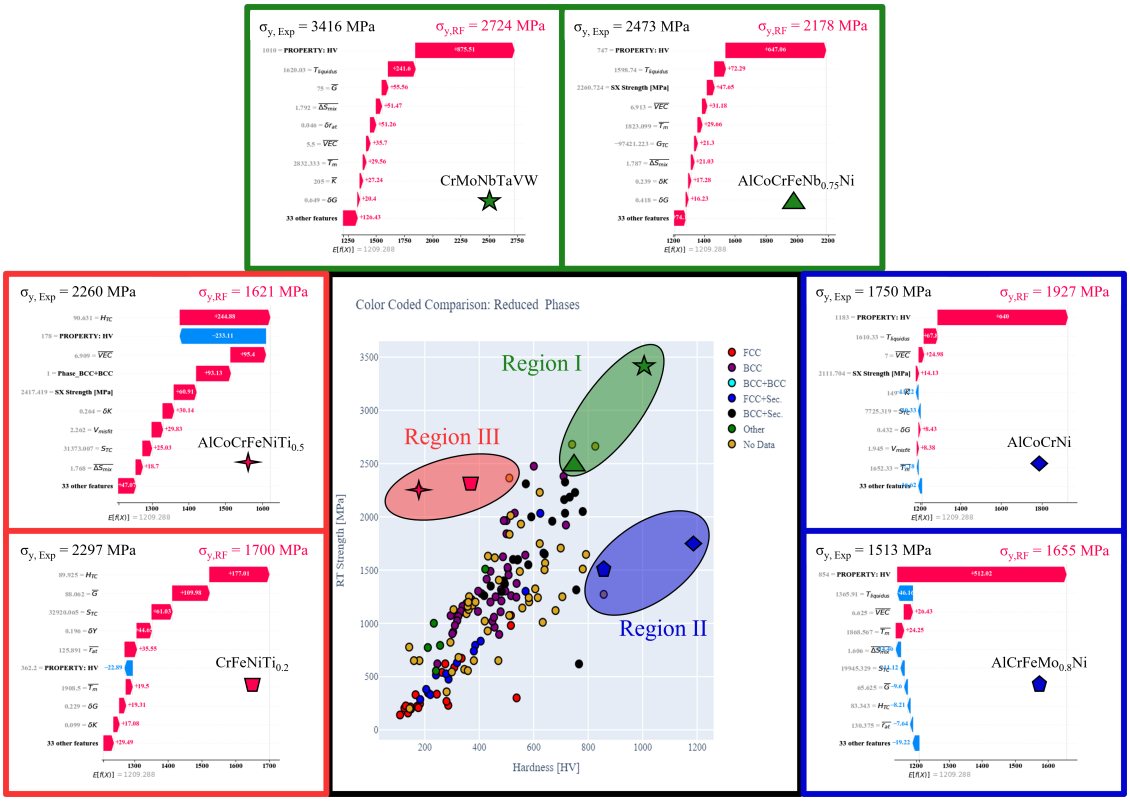

Figure 4.26. SHAP waterfall plots for CCA strength prediction. Region I [Green, Top panels]: High strength and high hardness. Region II [Blue, Right panels]: Low strength and high hardness. Region III [Red, Left panels]: High strength, low hardness.

4.4.5 Conclusions

The use of well engineered features for high entropy alloy strength coupled with optimization schemes and explainers were used to determine best set of features and their importance. In addition to global importance, localized explanations for model successes and failures were included for the process. Especially when working with small datasets, RF modeling has distinct advantages for explanation in single predictions, uncertainty quantification for active learning, and efficient algorithms for property exploration. The figures and methods used in this section are included in the online tool *ccaoptimizer* [272]. Interactive notebooks for feature collection, optimization, and explanation are provided.

4.5 Final Remarks

Using chemical based featurizers we use machine learning methods to expedite materials screening and explorations. Building on generally small datasets, random forest regressors proved to be highly useful both in their cost efficiency, and the explainability of models for strength, oxide melting point, and high entropy alloy melting point. The models provided within this chapter are accessible at their respective cited sources, and can be modified for other platform usage. Leveraging both physics-based descriptors and simulation derived quantities will be of interest for future material explorations.

5. DEEP LEARNING TOOLS FOR ATOMIC ENVIRONMENT MODELING

ADAPTED FROM: Mackinzie S. Farnell, **Zachary D. McClure**, Shivam Tripathi, and Alejandro Strachan. Modeling environment-dependent atomic-level properties in complex-concentrated alloys. The Journal of Chemical Physics, 156, March 2022. [273]

Zachary D. McClure, Robert Appleton, Nacim Bouia, Jean-Bernard Maillet, David Guzman, David P. Adams, and Alejandro Strachan. Practical database enrichment strategies with iterative learning: Neural Network Potentials for $Ge_2Sb_2Te_5$&$Ge_2Sb_2Te_5$+C. TBD.

5.1 Introduction: Atomic Environment Featurization

The properties and methods described in Ch. 4 are relevant for large scale properties assuming that all atomic species within the chemical composition repeat themselves periodically along a lattice, and are non-variable at their localized level. However, the very nature of materials such as amorphous glasses, and high entropy alloys is their disposition towards disorder and non-uniformity. To properly correlate atomic properties to larger scales careful descriptions of the atomic environments must be created. Unique fingerprints that represent the atomic environment with invariance to rotation and translational shifts are used as feature sets for property determination. In this chapter we will discuss two aspects of atomic-level feature engineering, and their implications for materials modeling. First, from simple molecular statics simulations powered by an interatomic potential we assess the variance in properties such as cohesive energy, vacancy formation energy, and atomic stress. Neural network methods are applied given the high dimensional properties of both our dataset, and our selected features. We find that even with toy model approximations the high temperature effects on atomic-variant properties can lead to drastic shifts from uniform property attributes. Finally, a combination of DFT, MD, and machine learning methods are applied to expedite simulation processes in phase change memory devices. Deep learning algorithms are applied to DFT datasets to correlate energies and forces necessary

for interatomic potential functional fitting. The trained neural net potentials are capable of exploring lengths and timescales unreachable with DFT methods while retaining high fidelity of the predicted configurations.

5.2 Modeling environment dependent atomic properties

5.2.1 Introduction

Complex concentrated alloys (CCAs) and multiple principal component alloys are crystalline materials consisting of four or more elements combined in similar fractions. They have attracted significant attention since their introduction in 2004 [145], [146] due to a range of desirable properties including high strength, even at high temperatures, thermal stability, and resistance to fatigue [274]. In addition, the vast space of potential alloy compositions makes them tailorable to specific applications [145], [274], [275]. The inherent variability in the local atomic configurations is the driving factor behind many of their unique properties, but also poses significant challenges to modeling and experimental characterization [276]. For example, the distribution of vacancy formation energies determines vacancy concentration which, in turn, dominates creep. Another key example is single crystal strength, which is dominated by local changes in the core energy along dislocation lines [263]–[265]. The local atomic environments govern the energy landscape under which dislocations move and their variability hinders their mobility, resulting in strengthening. For other examples of the relationship between local variability and properties see Refs. [248], [277]–[279]. These local properties can be assessed computationally via intensive atomistic simulations, but given the enormous number of local atomic configurations individual atoms can encounter in CCAs, computationally efficient models for local properties are highly desirable. For example, Chen et al. studied vacancy formation energy (VFE) in CrFeCoNi alloys using density functional theory (DFT) on special quasirandom structures (SQS) [280]. The authors found a wide distribution in VFE ranging from 1.5 to 2 eV with averages between 1.58 and 1.89 eV depending on the element. This work explored 24 of the most likely configurations given a 20 atom cell, a small subset of all the possibilities. For example, the number of local first nearest neighboring configurations in a five-element alloy is 5^Z, Z the coordination

number, divided by the multiplicity due to symmetry operations; clearly brute force *ab initio* calculations and even lower-fidelity interatomic force field calculations are out of the question.

Efforts to efficiently explore and characterize this enormous space have turned to machine learning methods for phase prediction, material screening, and through that best practices have begun to emerge [281]–[283]. The foundations for these screening processes built on early work for formation energy determination using cluster expansion (CE) methods [284]. Extensions of this model beyond binary components have shown great success in ternary semiconductors for predicting possible phase formations and separation [285], and multi-component CCA [286]. However, the method relies on unpacking 1st, 2nd, and higher order pairwise interactions in a symmetric, unrelaxed system. For systems that have been relaxed, and symmetry disrupted, the CE models begin to break down [287]. To overcome this limitation, rather than describing atomic interactions through the CE formalism, Shapeev used tensor descriptions to represent the energetics of multicomponent systems and showed better convergence rate with respect to training set size than CE for total energies [288], [289]. Each of these respective methods consider pair-wise interactions within a system, and sum their total contributions to determine total system energy. However, many of these methods focus on the macroscale properties and not on the local variability. To inform single crystal strength models, approximations to the local stresses have been developed from atomic radii and elastic constants [263], [265]. These model are easy to evaluate but involve several approximations and the associated uncertainties have not been quantified. In this paper we develop predictive models for various atomic-level properties of CCAs from molecular mechanics simulation data using invariant descriptors of local atomic environments and chemistry and neural networks. Recent work on high entropy diborides used atomistic simulations to develop models for VFE depending on the local environment. The authors showed the ability of pair approximation models with linear models and local structure up several neighboring shells to provide accurate descriptions [290].

In summary, the development of validated and computationally expedient models capable of predicting a variety of atomic-level properties of CCAs remains an active area of research and we are unaware of models capable of predicting a range of atomic-level properties needed to inform constitutive laws required for macroscopic predictions. To address this

gap, we combine molecular static calculations using a many-body interatomic potential with machine learning to create predictive models for local atomic properties of face centered cubic CCAs containing Co, Cr, Fe, and Ni. We model several properties (relaxed vacancy formation energies, atomic pressures and volumes, and cohesive energies) and assess the ability of the models to generalize and predict properties for new compositions and new chemistries. Importantly, the descriptors of local chemistry and geometry used as inputs to the models are generated from unrelaxed atomic configurations; thus, evaluating the models does not require computationally intensive structural relaxations.

Our work builds on the significant recent progress in the use of machine learning for atomistic simulations and a long history of modeling multicomponent systems [284]. Neural networks [291], Gaussian processes [218], and even linear regression [219] have been shown to be powerful models to relate local atomic environment and atomic energies, resulting in a new class of interatomic potentials. In these models, local atomic structures are described with descriptors that capture the symmetries of the underlying physics (e.g. translational and rotational invariance). Moment tensor potentials have also shown great promise to describe multicomponent systems [288], [292]. Approaches to descrisbe local atomic environments include smooth-overlap of atomic positions (SOAP) [293], two- and three-body symmetry functions [291], tensor formalisms [288], and bispectrum coefficients [219]. In this paper, we use bispectrum coefficients to relate the local, first nearest neighbor, environment of the *unrelaxed* structure to various *relaxed* local properties. Thus, our models need to learn not just the mapping between structure and property but also the relaxation of the local structure. In addition to the geometry, we use standard description of chemical properties of each environment. We explore the ability of the models to predict environments not seen during training including those originating from unseen compositions as well as the inclusion of new elements.

5.2.2 Methods

LAMMPS Simulations

The atomic properties of interest (relaxed vacancy formation energy, cohesive energy, stress, and volume) were obtained using the LAMMPS simulation package [40] with an embedded atom model interatomic potential developed by Farkas et al. [233]. Initial structures of the CCA alloys of interest, equiatomic Cr, Fe, Co, Ni, Cu, were obtained using an FCC lattice with lattice parameter a_0=3.56 Åwith atoms assigned following the SQS method. [294] All descriptors used as inputs for the neural network models are calculated from these initial structures, as described in sub-section 5.2.2.

After the descriptors are extracted, we relax the structure using molecular statics. We minimize the total energy with respect to both lattice parameters and atomic coordinates under ambient pressure with thresholds of 10^{-12} and 10^{-12} eV/Åfor scaled energy and force, respectively.

After relaxation, we compute the atomic energy (defined as the potential energy contribution of each atom), local atomic stress from the virial theorem [295], and local volume from a Voronoi tessellation [296]. Finally, the vacancy formation energy of each atomic site is computed by sequentially removing each atom and re-relaxing the structure (maintaining the simulation cell parameters constant). We define the relaxed vacancy formation energy (E_v^{i}) for site i from the energy difference between the perfect crystal E_0 and the system after the removal of corresponding atom E_{i}.

$$E_v^{\mathrm{i}} = (E_{\mathrm{i}} + \mu_{\mathrm{i}}) - E_0, \tag{5.1}$$

where μ_{i} is the chemical potential of atoms of element corresponding to atom i. This chemical potential is obtained as the cohesive energy of a pure element system.

The distributions of the resulting properties for each atom type obtained from a 5,000-atom SQS structure are shown in Fig. 5.1. These distributions compare well with prior ab initio calculations [280]. Our average relaxed vacancy formation energies for Cr, Fe, Co, and Ni are 1.52, 1.58, 1.44 and 1.63 eV, respectively. These points compare well with *ab*

initio results reporting average values of 1.61, 1.58, 1.70, and 1.89 eV for Cr, Fe, Co, and Ni obtained in 4-element CCAs.

We note that we use an interatomic potential since our goal is to establish the validity and accuracy of our proposed model of relaxed atomic-level properties. For more accurate models the interatomic potential would be replaced by DFT calculations that provide a good balance between accuracy and computational cost and can capture properties associated with the electronic structure of the systems, such as magnetism.

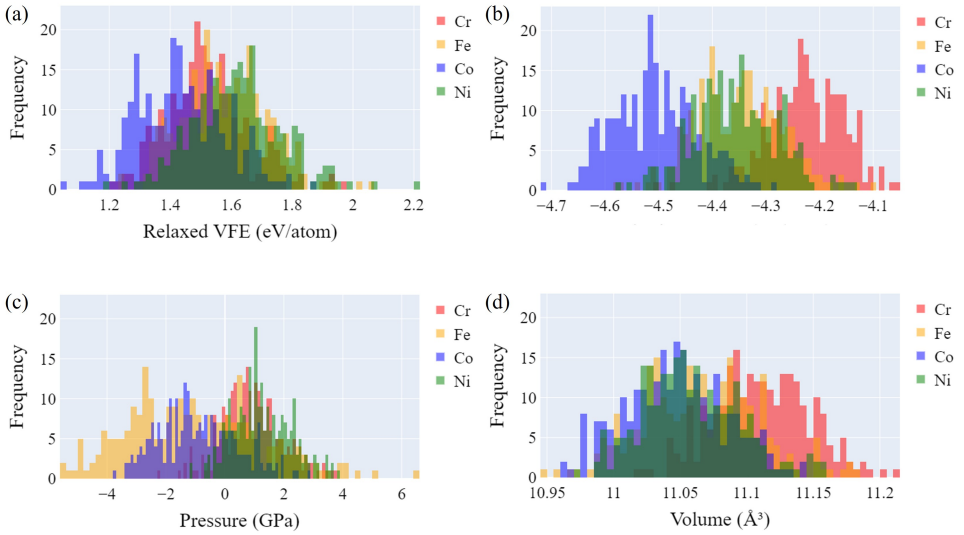

Figure 5.1. Distribution of values for Relaxed VFE (a), Cohesive Energy (b), Pressure (c), and Volume (d).

Model Features

We use a combination of chemical and geometrical descriptors to describe individual atoms. As described above, all descriptors are obtained from the initial, unrelaxed, structures. To describe the local geometrical environment we use bispectrum coefficients [219] that start from the local atomic density around an atom and create a list of translationally and rotationally invariant descriptors. To distinguish between atom types in the bispectrum calculation, we use atomic numbers as prefactors in the local density during the coefficient calculations. Bispectrum coefficients are obtained using a radial cutoff 10% beyond the

nearest neighbor distance (1.1 a_0 $\sqrt{2}/2$) and a band limit of eight for the resulting in a total of 55 coefficients. We note that the bispectrum coefficients capture up to four-body correlations and do not provide a complete description of atomic environments [297] and multiple local environments can lead to identical coefficients. This issue is less of a concern for multi-component systems and, from a practical point of view, near-DFT accuracy has been obtained for simple metals [298]. Thus, we believe the bispectrum coefficients provide an appropriate description for the problem at hand. In addition to the geometric descriptors, we use the atomic number of the central atom and the following chemical descriptors for the central atom queried from Pymatgen: [250] atomic radius, atomic mass, Poisson's ratio, electrical resistivity, thermal conductivity, and Brinell hardness. These properties were chosen to describe the size, bonding, and electronic structure of the central atom. We also studied the effects of using descriptors capturing the central atom and the 12 nearest neighboring atoms using a rule of mixtures, but found that these did not improve model performance; these results are discussed in the open data repository *Machine learning for high entropy atomic properties* in the section "Train Neural Network on Equiatomic CrFeCoNi". These descriptors were added as additional physics informed descriptors, and have good overlap with previously investigated descriptors used in material classifications [299].

Neural network architecture

Machine learning models were implemented in the Jupyter notebook environment [300] on nanoHUB [12] using Tensorflow [301] and Keras [302] libraries. The models use shallow neural networks with a first hidden layer containing 512 neurons connected to the 63 input features. This hidden layer used exponential linear unit (elu) activation functions and was followed by a dropout layer with dropout ratio of 0.2. During training, the loss function was mean squared error and the *Adagrad* optimizer was used [49]. Also, the learning rate was 0.002 and the models were trained for 5000 epochs. The model architecture and hyperparameters were chosen after testing several models, as detailed in the data repository material workflow.

To train the model, the data was split into testing and training sets, with 80% of data used for training and 20% used for testing. The inputs and outputs were normalized using

the standard approach of subtracting the mean and dividing by the standard deviation of the training data. During training, 10% of the training data was used for validation. The validation data differs from the testing data in that it is used during the training of the model to assess convergence, while the testing data is hidden during training and only used after training to evaluate the model. Initially, an early stopping criterion based on validation data was used to determine number of epochs for training. However, models had similar errors when trained with early stopping and with 5000 epochs, so 5000 epochs were used to train all models. Independent models were developed for each property of interest to describe all elements in the system. The initial model architecture was developed using equiatomic CrFeCoNi structures with a data set containing 5000 atoms. However, we found that training with 2000 atoms was sufficient. Thus, models were then trained and tested on equiatomic four-element alloys CrFeNiCu, FeCoNiCu, CrCoNiCu, and CrFeCoCu with data sets containing 2000 entries (atoms) each. The predictive ability of these models was tested on the five element alloy CrFeCoNiCu and on non-equiatomic alloys.

5.2.3 Models for atomistic properties of CCAs

As described above, we trained neural network models to predict relaxed vacancy formation energy, atomic cohesive energy, atomic pressure, and local Voronoi volume. Figure 5.2 shows parity plots of the four properties for CrFeCoNi alloy. Only testing data points are shown, these have not been used in training. The results highlight the large atomic variability of all the properties studied, the range for each element is larger than the difference in mean values between elements. The dash lines bound errors corresponding to 10% of the range of each property. In absolute terms, the the mean absolute errors are 0.042 eV for cohesive energy, 0.059 eV for VFE, 0.809 GPa for pressure, and 0.020 Å^3 for atomic volume. Figure 5.3 compares the accuracy of the models for the five four-element alloys used for training. We show the mean absolute error of all predictions normalized by the range over the testing data points. Our models have comparable performance across the different chemistries. Importantly, models can predict properties with an accuracy of approximately

10% of the range for each of the properties studied. This level of accuracy is comparable to that achieved in high-entropy borides using first nearest descriptors [290].

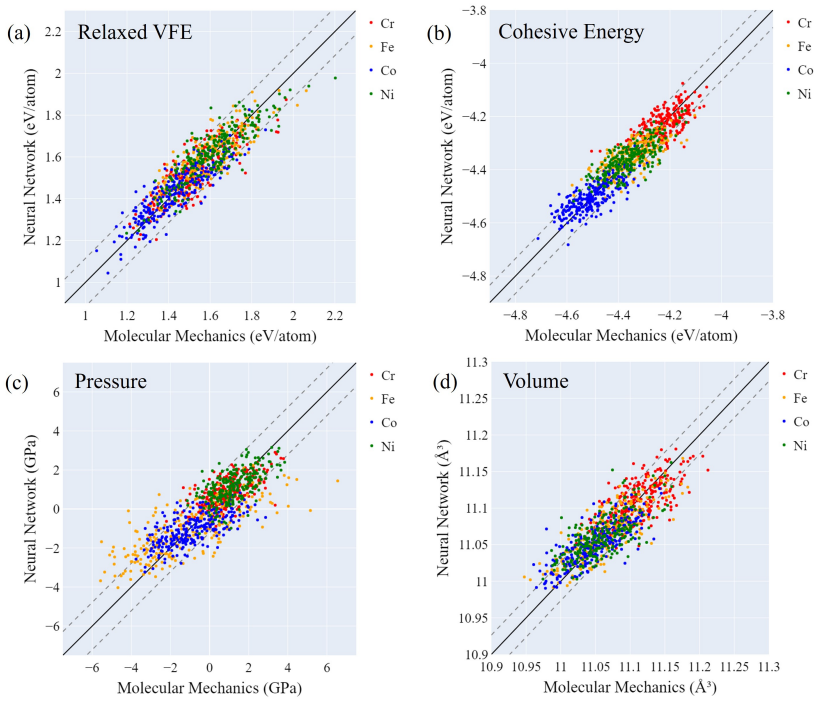

Figure 5.2. Machine learning model predictions compared to molecular statics results for relaxed VFE (a), cohesive energy (b), atomic pressure (c), and atomic volume (d) for equiatomic CoCrFeNi configurations belonging to the testing set. The grey, dashed lines indicate errors of ± 10 % of the range for each property, in absolute terms these represent ± 0.115 eV for relaxed VFE, ± 0.065 eV for cohesive energy, ± 1.213 GPa for atomic pressure, and ± 0.026 Å^3 for atomic volume.

Predicting properties for new compositions

The model trained on equiatomic CrFeCoNi was used to predict properties of alloys with different compositions with the same four elements. Neural network predictions are compared to molecular statics predictions in Figures 5.4 and 5.5. Figure 5.4 assesses the model accuracy for $Cr_{20}Fe_{40}Co_{20}Ni_{20}$. We find the model to be able to make accurate predictions across all properties. The normalized MAE values are slightly larger than those for the composition used for training, with models predicting with an accuracy of roughly 20% of

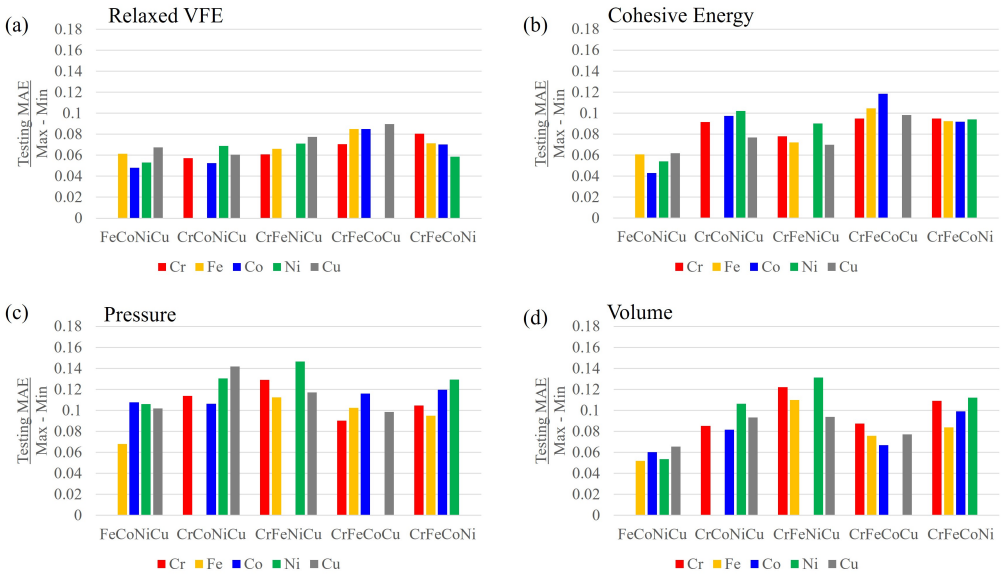

Figure 5.3. MAE normalized by range for the testing data for each of the four-atom systems for Relaxed VFE (a), Cohesive Energy (b), Pressure (c), and Volume (d).

the range of each property. The slight underestimation of the Voronoi volumes is due to the larger overall volume of this Fe-rich alloy. Figure 5.5 assesses the ability of the model trained on equiatomic CoCrFeNi to predict on $Cr_{15}Fe_{55}Co_{15}Ni_{15}$. For this composition, with environments more rich in Fe that deviate further from the training data, the model accuracy degrades further. The model is still able to capture overall trends in properties but the trend observed above of underestimating atomic volumes accentuates with increasing Fe. Going from the equiatomic systems to the $Cr_{15}Fe_{55}Co_{15}Ni_{15}$, the average volume computed using molecular mechanics increases from 11.070 Å^3 to 11.146 Å^3. In contrast, the model average volume predictions are essentially unchanged. This indicates that the model cannot capture the overall expansion observed with increasing Fe content, this is not surprising as this information was not provided to the model during training.

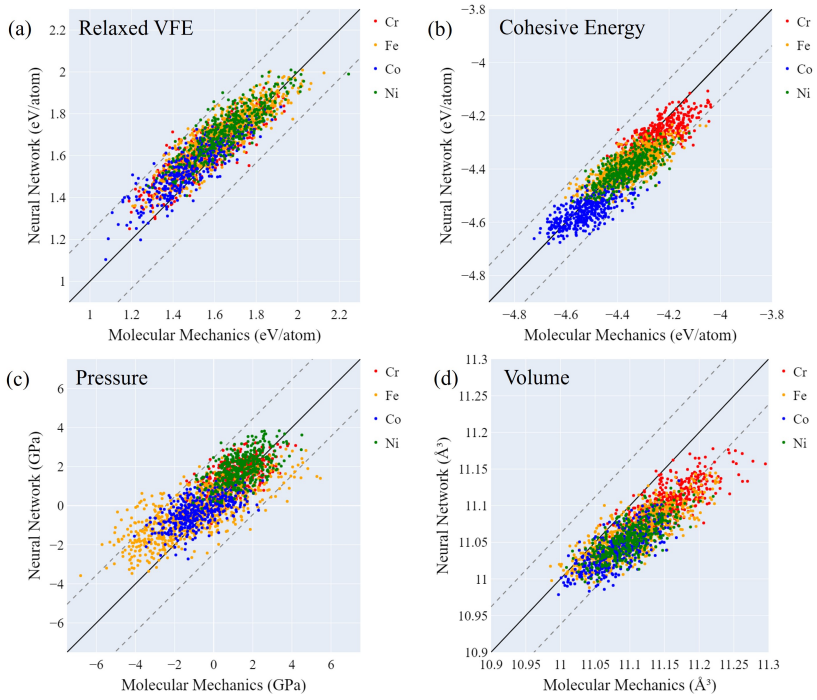

Figure 5.4. Parity plots for $Cr_{20}Fe_{40}Co_{20}Ni_{20}$ for Relaxed VFE (a), Cohesive Energy (b), Pressure (c), and Volume (d). Predictions were made using model trained on equiatomic CrFeCoNi. The grey, dashed lines bound ± 20% of the range for each property.

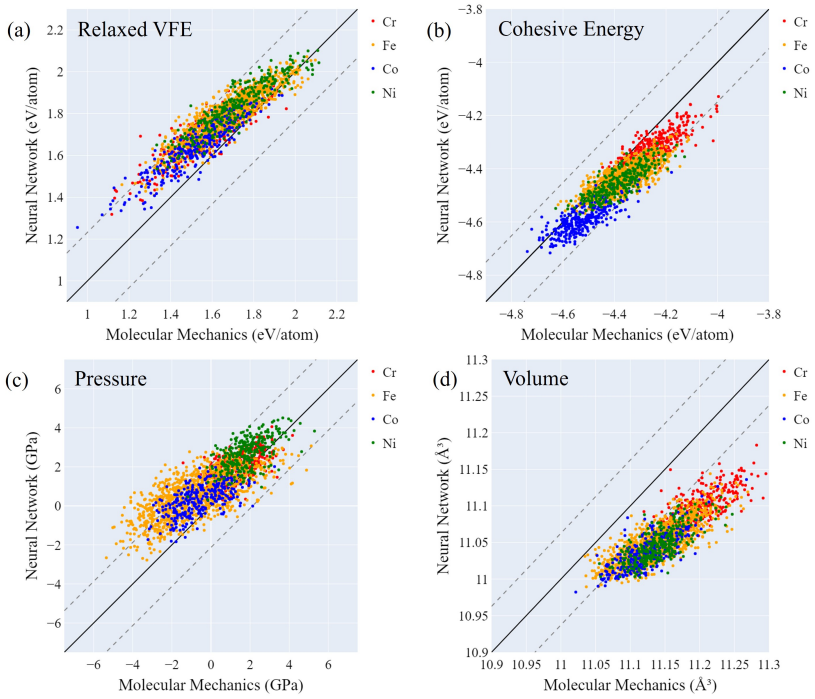

Figure 5.5. Parity plots for $Cr_{15}Fe_{55}Co_{15}Ni_{15}$ for Relaxed VFE (a), Cohesive Energy (b), Pressure (c), and Volume (d). Predictions were made using model trained on equiatomic CrFeCoNi. The grey, dashed lines bound ± 20% of the range for each property.

The model trained on equiatomic CrFeCoNi was also used to make predictions on several other alloys with different compositions. The error in these predictions, for the four properties of interest, is shown in Figure 5.6. The first composition in each panel of Figure 5.6 represents the one used for training. These results indicate that the model has some predictive power on unseen compositions, giving better predictions on compositions closer to training set. For compositions with 40% of a particular atom and 20% of each of the other atoms, the model accuracy is roughly 20% of the property range. For compositions with 55% of a specific atom and 15% of each of the other atoms, the model accuracy is roughly 30% of the property range for relaxed vacancy formation energy and cohesive energy and 50% of the range for atomic volume.

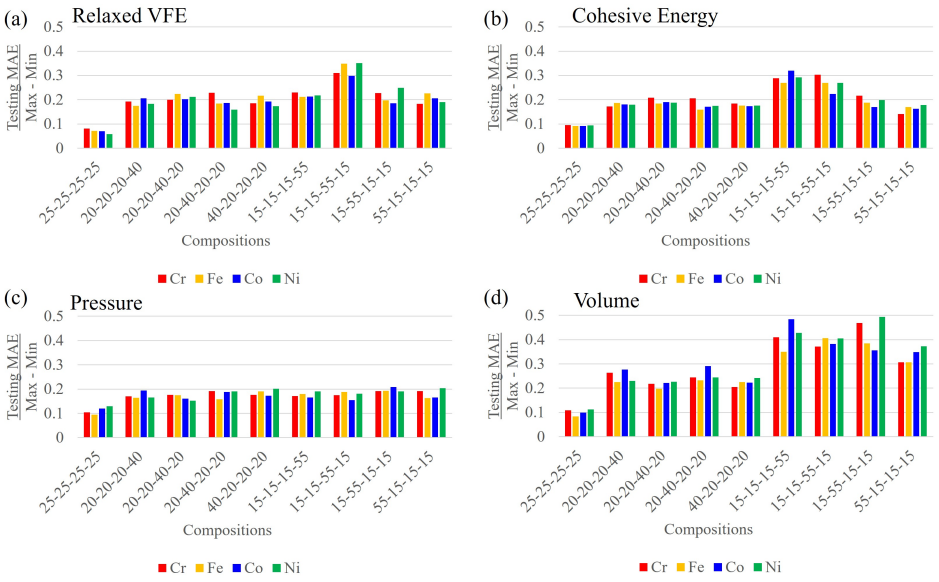

Figure 5.6. MAE for predictions on untrained compositions for: (a) Relaxed vfe, (b) Cohesive Energy, (c) Pressure, and (d) Volume. The model was trained on equiatomic CrFeCoNi.

Predicting properties for new chemistries: CrFeCoNiCu

Finally, we tested the model's ability to predict properties of systems with unseen elements. We used five models trained on single four-element alloys (CrFeCoNi, CrFeNiCu, FeCoNiCu, CrCoNiCu, and CrFeCoCu) to make predictions on CrFeCoNiCu. Results for

vacancy formation energies are shown in Figure 5.7, with the other properties included in the data repository information workflow. Figure 5.7 indicates that the relaxed vacancy formation predictions of all elements on the CoCrCuFeNi are accurately described by the models trained on CrFeCoCu (missing Ni), CrFeNiCu (missing Co), and CrCoNiCu (missing Fe) but rather poorly by the models trained on FeCoNiCu (missing Cr) and CrFeCoNi (missing Cu); note that Cr and Cu are the end elements within our group in terms of atomic number.

To understand the underlying reason for these differences, we compared the inputs between the various alloys, specifically the unrelaxed bispectrum coefficients for CrFeCoNiCu with those for the four-element alloys. Figure 5.8 shows the distributions of the first coefficient. We find that the systems trained without Fe, Co, and Ni have relatively similar local descriptors (bispectrum coefficients) to the CrFeCoNiCu system. However, the descriptors for the alloys lacking Cu or Cr show significantly different distributions of descriptors as compared to the 5-element CCA. For FeCoNiCu (without Cr), the differences in the local environments are more pronounced than for CrFeCoNi (without Cu), explaining why the model shows very poor performance. We observe the same trends for the other bispectrum coefficients. This is due to the use of atomic number as prefactors in the construction of bispectrum coefficients. Ni, Fe, and Co lie between the elements trained on while Cr has the lowest atomic number of the group and Cu has the highest atomic number.

5.2.4 Discussion and conclusions

We combined molecular statics, atomic level featurization, and data science to develop models for atomic properties in high entropy alloys from local atomic environment and elemental information. Our approach relates descriptors that are easy to obtain from unrelaxed atomic structures to properties that require atomic relaxations and, thus, are computationally more intensive to obtain. Evaluation of the models requires simply generating an atomic structure, performing a local structure calculation, computing atomic-based descriptors, and evaluating a neural network. For testing data, the model predictions were within 10% of the range for each of the properties studied. This level of accuracy is comparable with that of the pair approximation models of Daigle et al. when only the first neighboring cell is used.

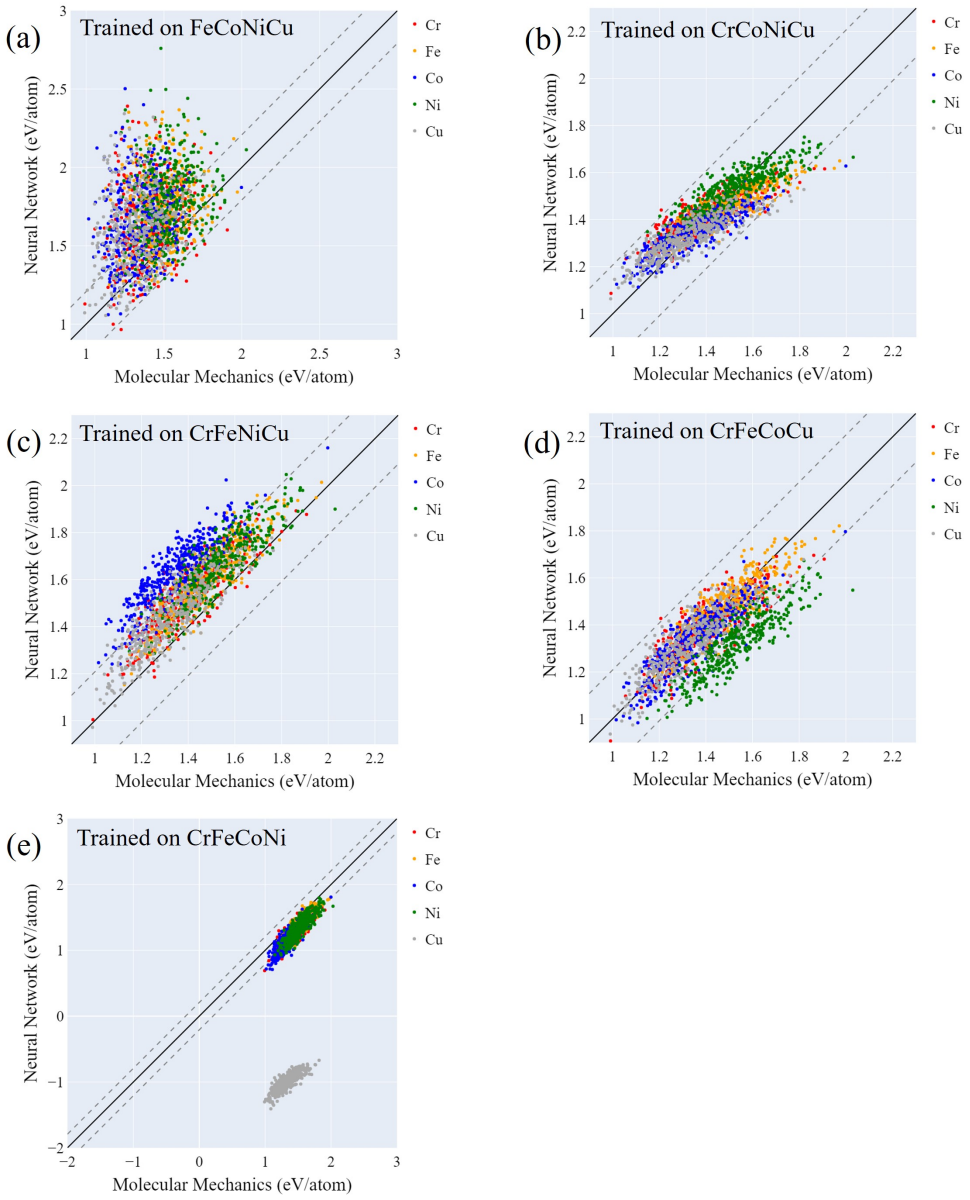

Figure 5.7. Parity plots for relaxed vacancy formation energy for predictions on CrFeCoNiCu. Model was trained on FeCoNiCu (a), CrCoNiCu (b), CrFeNiCu (c), CrFeCoCu (d), and CrFeCoNi (e). The grey, dashed lines bound ± 20% of the range for each property.

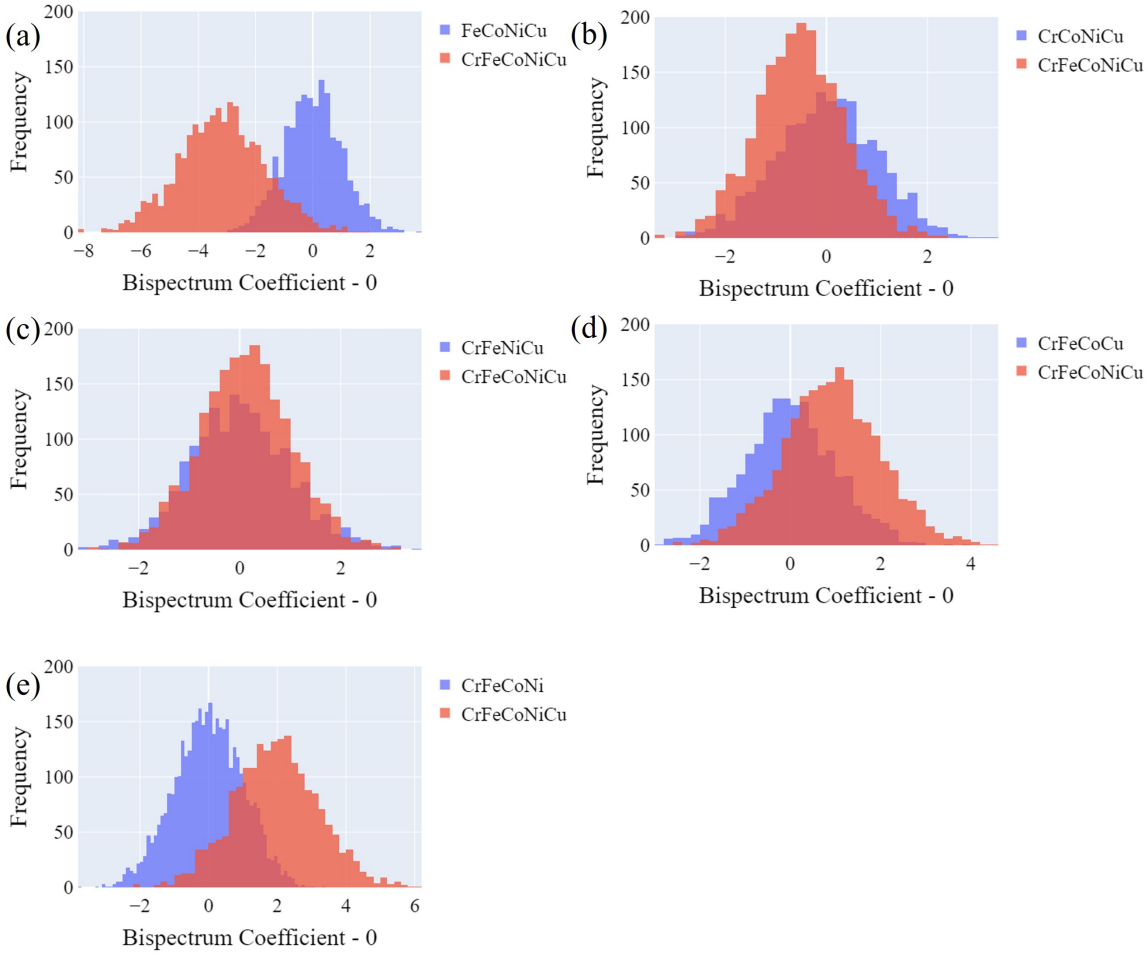

Figure 5.8. Distribution for zeroth bispectrum coefficient for FeCoNiCu (a), CrCoNiCu (b), CrFeNiCu (c), CrFeCoCu (d), and CrFeCoNi (e) compared with CrFeCoNiCu. Bispectrum coefficient was normalized using the mean and standard deviation for FeCoNiCu (a), CrCoNiCu (b), CrFeNiCu (c), CrFeCoCu (d), and CrFeCoNi (e).

[290] The authors demonstrate improvements in accuracy as additional shells are included. We assessed the ability of our models to predict concentrations and chemistries not used during training, and we find that the model has can predict properties for several unseen concentrations and chemistries.

The local atomic properties modeled are important in determining several macroscopic properties of CCAs. As mentioned above, models for local volumes and stresses can inform single crystal strength models [263]. In addition, the distribution of VFEs affect vacancy concentrations. To exemplify the importance of capturing distributions, Figure 5.9 compares the equilibrium vacancy concentrations vs. inverse temperature for each element in a CrFeCoNi alloy considering the distribution of VFEs (solid circles) with the values assuming a constant value (set to the mean VFE for each element). The vacancy fraction calculated from neural network predictions of VFE compares well with the vacancy fraction calculated from molecular mechanics predictions of VFE. As also observed in shown borides, [290] a distribution of VFEs results in non-Arrhenius behavior as the relative contribution of different values is temperature dependent. All calculation details are included as Jupyter notebooks in our open data repository.

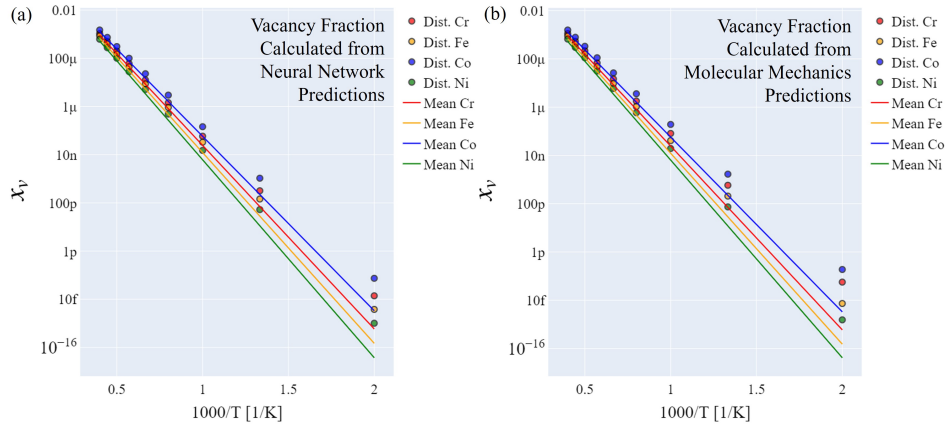

Figure 5.9. Vacancy Fraction of HEA elements in an alloy given the mean VFE (solid lines), and calculating a population of vacancies based on the full distribution (circles) using neural network predictions (a) and molecular mechanics predictions (b)

In summary, atomic-level fluctuations in CCAs and other multi-principal component materials result in unique and often desirable properties. Our results indicate that atomic level simulations, appropriate descriptors, and machine learning tools can be used to capture such variability. In this work we used properties computed from a many body force field for computational expediency, but the overall approach can be used with more accurate *ab initio* results.

Continuing with the work of atomic level descriptors, we use the knowledge gained from high-fidelity DFT simulations for interatomic potential training and functionality. Explorations with high-dimensional neural networks are shown, but with a different set of atomic fingerprints for featurization. Rather than computing a range of atomic properties we will focus on the energies and forces needed to propagate dynamics in a system. These trained potentials are capable of producing DFT level accuracy for hundreds of thousands of atoms at nanosecond timescales, unlocking dynamics unobtainable by *ab initio* methods.

5.3 Development of neural network potentials for phase change memory devices

5.3.1 Introduction

$Ge_2Sb_2Te_5$ (GST) alloys for phase change memory (PCM) are a critical component in electronic devices for rewritable memory storage, neuromorphic computing, and resistive random access memory (RRAM) components [61], [303], [304]. Two phases of the material, crystalline and amorphous, can be rapidly switched between using well controlled energy pulses through electrical or optical discharge [305]. The kinetics of phase change and phase stability can be altered by composition of the base alloy [306], or with dopants such as N, Al, Cu, C, Sn, and O [307]–[311]. The basic read-write process (liquid to crystal or amorphous, and amorphous to crystal) of a device takes mere nanoseconds to complete. While bulk phases can be determined through resistance measurements, the atomstic configuration regarding bonding and phase change nucleation is difficult to capture. However, to great success, the physics behind the GST amorphous to crystalline phase transition has been well characterized using single point density functional theory (DFT) calculations [312] and *ab initio* molecular dynamics (MD) methods [313]–[316]. However, the kinetics of phase transformation, and

device molecular scale often exceed the current state of the art super-computing resources with *ab initio* MD methods and the need for an interatomic potential approximation is needed.

For the binary system of GeTe a modified Tersoff-type potential [317], and a neural-network trained potential [318] have been used, and to date only one machine learning (ML) potential exists for GST using the Gaussian approximation potential (GAP) framework [319]. These potentials have helped characterize medium and long-range order effects in GeTe and GST recrystallization beyond *ab initio* scales, but quaternary effects with the addition of dopants have not been explored.

To bridge the multi-scale gap between *ab initio* computing, and nano-to-mesoscale operation, efforts to train machine learning models on DFT databases for larger length and time scale extrapolation have flourished. These ML potentials primarily include, but are not limited to, atomic cluster expansion [320], neural network potentials (NNP) [220], [321], graph networks [322], Gaussian approximation potentials (GAP) [218], kernel ridge regression [323], spectral neighbor analysis potentials (SNAP) [219], moment tensor potentials (MTP) [288], gradient-domain machine learning cite [324], and support vector machines (SVM) [325].

A recent benchmark on a standard database was published evaluating current inter-atomic potentials based on classical mechanics such as EAM [326], MEAM [35], and were compared to the state-of-the-art ML methods [298]. Of the listed potentials, MTPs create the pareto front for current limits of computational cost and accuracy, with GAP being one of the most expensive. Trading accuracy for cost, ableit less than SVMs, NNPs show great promise for materials exploration. However, the training database often is the limiting factor in physics exploration [298]. Augmentations to these databases is an active area of exploration for materials research through both iterative learning [210], and active learning methods [327]–[330]. The ability to scale a machine learning interatomic potential with on-the-fly acquisitions of new training structures is of essential benefit for taking already established potentials and adding additional constituents. For GST in particular, extending ML methods for development of easily acquirable interatomic potentials is of great interest.

Acknowledging that HDNNPs are not the highest accuracy potential with regard to MTPs, and a highly parameterized GAP approximation, their flexibility towards material

systems while still performing at near DFT accuracy is of great interest. In addition, the low cost for training these potentials offers advantageous study of the hyperparameter optimization of the NNP, rapid acquisition of new trajectories and DFT sampling for iterative data augmentation [210], and interpretable feature generation for an N-dimensional system.

In this work we will begin with curation details of our GST DFT database. Using equation of state single point calculations and *ab initio* MD trajectories a representative space of 1500 total configurations of Ge, Sb, Te, GeTe, and GST were sampled. Following the training database description will be an overview of our iterative workflow, including selection of weighted atomic-centered symmetry functions (wACSF) via grid approximation and CUR decomposition. In part to their high-dimensional nature, the likelihood of the network to settle in local minima is a concerning component. The effects of re-initializing the network force weights and biases explored in detail, and their impacts on iterative training explored.

Comparing separate iterative training strategies we propose guidelines and a workflow for rapid generation of training trajectories for NNPs, extrapolation to structures and dynamics unavailable with *ab initio* MD methods. Finally, we showcase the capability of our workflow to incorporate a highly sough after dopant: C. We develop and compare fully trained NNPs for GST & GST+C and evaluate the differences in atomic structure.

5.3.2 Database Generation & Augmentation with an Iterative Loop

We begin with an overview our of iterative learning workflow, and the structure of how we build our database. The initial database of Ge, Sb, Te, GeTe, and GST configurations with atomic forces and system energies were collected from DFT+MD and single point calculations. Using atomic centered symmetry functions, a set descriptors were generated for each respective element. A subsequent initial NNP was trained on the DFT configurations. Following the convergence of the network we assemble sets of configurations for MD simulations with the NNP. The trained NNP will be used to drive an NVT MD simulation as the interatomic potential expression. Trajectories of the hexagonal, cubic rocksalt, amorphous, and liquid phases are collected at a range of temperatures, and a range of densities for the amorphous phase. Initial setpoint densities were used from the final trajectory of the respective DFT

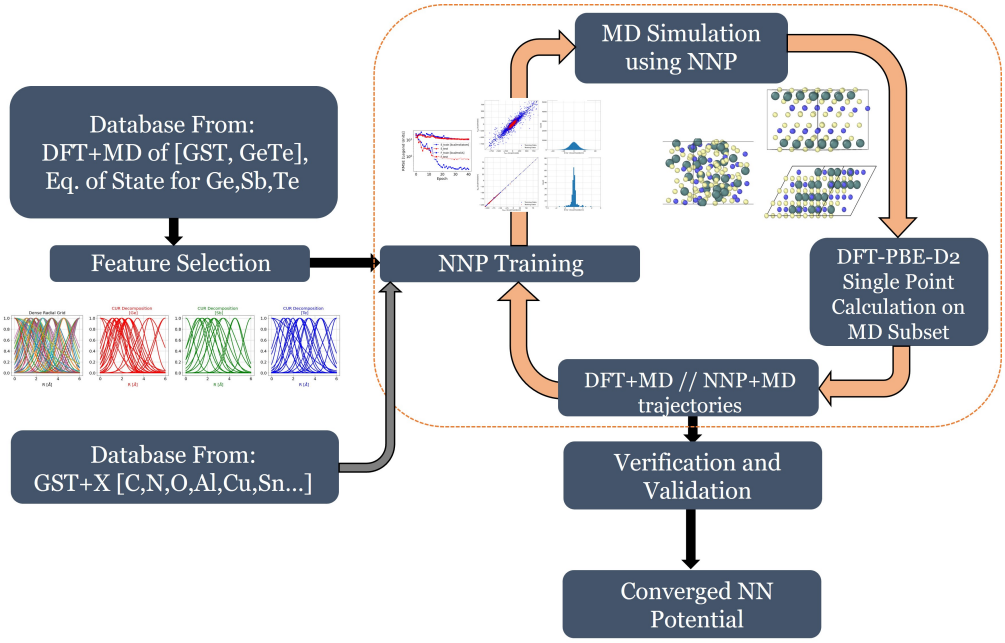

Figure 5.10. GST Workflow

simulation. Following the NNP-MD simulation, 10 evenly spaced trajectories are collected and a single-point DFT calculation is performed. It is often the case that the first generation of a ML potential yields erroneous trajectories, but these outlier explorations can be useful in the next generation of the augmented database. The cycle of training, MD+NNP extrapolation, DFT assessment, and database augmentation continues until sufficient evidence of high accuracy with respect towards validation and testing sets are achieved.

Initial Density Functional Theory Database

Electronic state calculations were completed within the generalized gradient approximation (GGA) with the Perdew-Burke-Ernezerhof (PBE) exchange-correlation functional [126] via projector-augmented wave pseudopotentials [331]. Extended range London dispersion effects were added in the form of the empirical D2 correction [332]. *Ab initio* MD simultions were performed with a 1 fs timestep, 300 eV KE cutoff, and 1x1x1 k-point grid for gamma point sampling. All DFT simulations were completed using the Vienna *ab initio* simulation package (VASP) [333], [334].

Volume expansion and contractions of the single element Ge, Sb, and Te unit cells were performed to obtain single element reference energies, as well as equation of state volumes. Initial structures for the hexagonal phase were generated from a 4x4x1 replica of the ground state unit cell [335]. The structure coordinate files were queried from the Materials Project Ref-ID mp-1224375 [14], [336]. The 144 atom structure was shifted to match experimental lattice parameters determined via XRD at 300K [337], and equilibrated under NVT conditions. The cubic-rocksalt phase was generated with 144 atoms in an SQS supercell where Te occupies one sublattice and the other is occupied by random distributions Ge, Sb (40% each) and vacancies (20%). This was done to model previous computational works [316], [338]–[340] and with initial lattice parameters of 6.00 Å, in good agreement with past experiments [341]–[343]. GeTe trajectories were obtained from previous in house trajectories of *ab initio* MD simulations.

Amorphous structures were generated via a melt-quench protocol of the supercell under NVT conditions at 5.6 g/cm^3. The system was heated to 2000K for 20ps to allow

for full atomic diffusion. The final liquid trajectory was copied, and quenched to 1100K, and compressed to an experimental density of 5.88 g/cm^3 [344]. The final trajectory of the 1100K simulation was quenched to temperatures ranging from 300-800K, with densities corresponding to experimental, minimized residual stress, and best practices DFT estimates [313], [344]. Liquid trajectories at 1100K after quenching from 2000K were collected as well.

For each structure [hexagonal, cubic, amorphous], we sample MD trajectories over 20ps with a frequency of 1ps (20 frames/20ps simulations), for temperatures of 300-1000K. Stresses in the X, Y, and Z directions were monitored during dynamics, and if residual stresses remained the cell was expanded or contracted accordingly to match hydrostatic pressure. Systems were equilibrated to within 0.1 GPa residual stress. However, some amorphous trajectories at compressed densities (5.88 and 6.11 g/cm^3) were included with residual stress built in.

Table 5.1. Initial DFT database composition for GST iterative interatomic potential training. Database is a small subset of full DFT+MD trajectories available for validation and sourcing.

Data Type	Temperature (K)	Structures
Ge/Sb/Te	0	150
GeTe	300-1000	828
Hexagonal $Ge_2Sb_2Te_5$	300-800	100
Cubic $Ge_2Sb_2Te_5$	300-800	100
Amorphous $Ge_2Sb_2Te_5$	300-800	114
Liquid $Ge_2Sb_2Te_5$	1100 and 2000	40

5.3.3 High Density Neural Net Potentials [HDNNPs]

To fully describe the total energy of the atomic system, a set of descriptors are required for each atom to be characterized in its local environment. Critical to the generation of a descriptor is its highly unique identity to the atomic configuration, invariance to rotational or translational shifts, and ease of generation. Common sets of atomic descriptors used in modern literature include the Coulomb matrix [323], bag of bonds [345], moment tensor potentials [288], bispectrum coefficients [218], [219], and the atomic-centered symmetry functions both unweighted and weighted (ACSFs/wACSFs) [220], [346]. In this work we will implement the

latter most feature set given their ability to scale well to higher dimensional elements, but our final dataset could be used for training with a different feature set for benchmarking in future work

In 2007, Behler and Parinello [220] applied Gaussian symmetry functions centered around an atomic position to model local energies, E_{short} up to a cutoff radius R_c. Rather than considering every atomic environment and configuration in a system at once, the method allows for rapid approximation of a local environment, and summing across all configurations for atomic species. For a given chemical specie in a Cartesian coordinate position, R_n^x, the system is converted into a series of symmetry functions, G_n^x. However, to fingerprint each local environment the symmetry function is dependent on all other atoms within the Cartesian coordinate cutoff sphere. Symmetry functions are used as the representation of features in a neural network unique to each chemical specie. Unique contributions to the total system energy are approximated for each atomic local environment.

$$E_{\text{short}} = \sum_{N_{\text{i}=1}}^{N_{\text{elem}}} \sum_{\text{j}=1}^{N_{\text{atom}}^{\text{i}}} E_{\text{j}}^{\text{i}} \tag{5.2}$$

Two types of symmetry functions are used to describe the neighboring atoms from the perspective of a central atom. The radial and angular functions represent the many-body interactions of local system, but primarily account for two-body and three-body interactions respectively. The radial symmetry function takes the form:

$$G_{\text{i}}^{rad} = \sum_{\text{j}\neq\text{i}}^{N_{\text{atom}}\in R_c} \text{e}^{-\eta(R_{\text{ij}}-R_s)^2} f_c(R_{\text{ij}}) \tag{5.3}$$

with R_{ij} the interatomic position between two atoms, the cutoff radius of interaction R_c, a radial shift R_s to shift the Gaussian function peaks to describe non-centered environments,

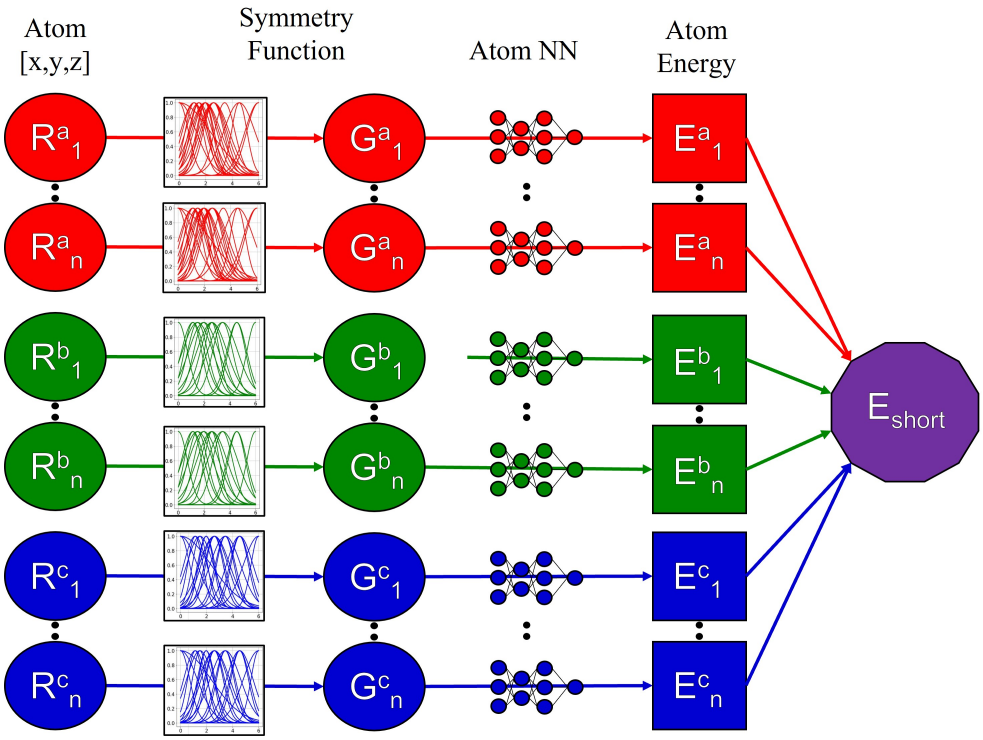

Figure 5.11. NNP Design

and η to control the width of the function description. Capturing 3-body interactions requires the addition of angular components to the radial description of an environment:

$$\begin{aligned} G_{\mathrm{i}}^{ang} =& 2^{1-\zeta} \sum_{\substack{\mathrm{j},k\neq\mathrm{i} \\ \mathrm{j}<k}} (1+\lambda\ cos\ \theta_{\mathrm{ij}k})^{\zeta} \mathrm{e}^{-\eta[(R_{\mathrm{ij}}-R_s)^2+(R_{\mathrm{i}k}-R_s)^2+(R_{\mathrm{j}k}-R_s^2)]} \\ & \cdot f_c(R_{\mathrm{ij}}) f_c(R_{\mathrm{i}k}) f_c(R_{\mathrm{j}k}) \end{aligned} \tag{5.4}$$

with $\theta_{\mathrm{ij}k}$ the angle formed with central atom i, and neighbors j and k, λ as ± 1 to center the maxima of the cosine function at 0^o and 180^o, and the ζ exponent to explore angular resolution.

ACSFs are one of the most generalizable set of features for atomic structures, and have seen wide success in atomistic modeling including: elemental bulk, multi-element bulk, water and aqueous solutions, surfaces, and molecular clusters [347]. However, since the description of atomic environments are defined by pairs and triples of element combinations the number of required features scales poorly with atomic specie.

To overcome the limitation of ACSFs, but still have the generalizability to a large distribution of configurations the wACSFs were developed [346]. Rather than using separate feature functions to describe each combination, the composition of the local chemical environment within R_c is computed with the atomic number Z, and a weighting function to modify the contribution of radial and angular components. As we extend our potential development to 3+ components in doped GST the use of wACSFs will be critical for linear scaling.

The wACSF representation of the radial and angular symmetry functions takes a similar form to the ACSF, but with the following modifications:

$$G_{\mathrm{i}}^{w,rad} = \sum_{\mathrm{j}\neq\mathrm{i}}^{N_{\mathrm{atom}}\in R_c} g(Z_{\mathrm{j}}) \mathrm{e}^{-\eta(R_{\mathrm{ij}}-R_s)^2} f_c(R_{\mathrm{ij}}) \tag{5.5}$$

and,

$$\begin{aligned} G_{\mathrm{i}}^{w,ang} =& 2^{1-\zeta} \sum_{\substack{\mathrm{j},k\neq\mathrm{i} \\ \mathrm{j}<k}} h(Z_{\mathrm{j}}, Z_k)(1+\lambda\ cos\ \theta_{\mathrm{ij}k})^{\zeta} \mathrm{e}^{-\eta[(R_{\mathrm{ij}}-R_s)^2+(R_{\mathrm{i}k}-R_s)^2+(R_{\mathrm{j}k}-R_s^2)]} \\ & \cdot f_c(R_{\mathrm{ij}}) f_c(R_{\mathrm{i}k}) f_c(R_{\mathrm{j}k}) \end{aligned} \tag{5.6}$$

where now instead of considering every combination of two-body, and three-body pairs we simplify the local chemical composition around central atom, i, by weighting functions $g(Z_{\mathrm{j}})$, and $h(Z_{\mathrm{j}}, Z_k)$.

Feature Selection

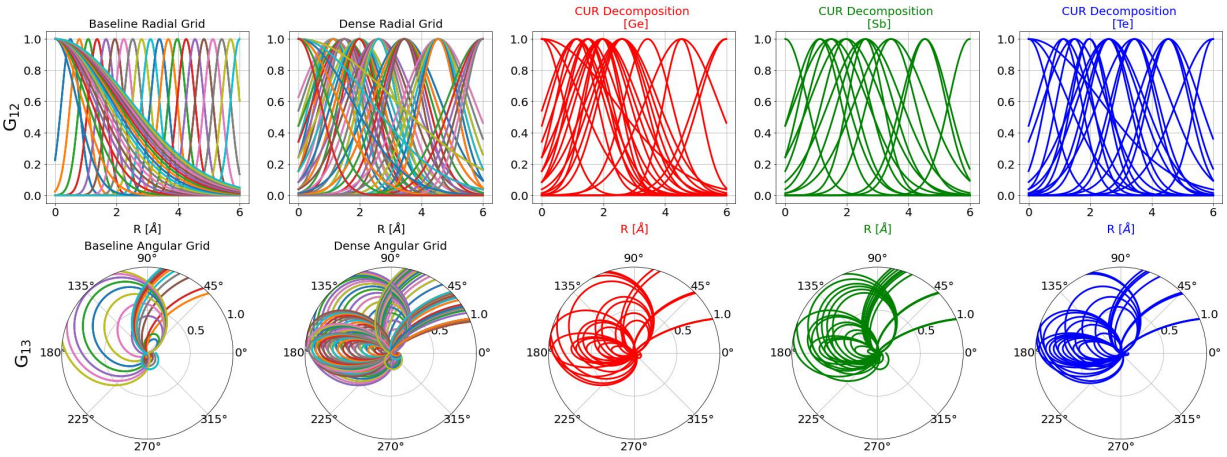

Figure 5.12. Radial [G12], and Angular [G13] Symmetry Function Visualization for: a) An agnostic grid selection of element features, b) An overly dense grid of features for CUR selection c) Individual element symmetry functions selected by CUR decomposition.

The process for feature generation classically involves generating a sufficiently dense grid of symmetry functions to describe the potential atomic configurations, while selecting features that do not duplicate information, force redundancies, or are unused entirely. Efforts to automate this process with principal component analysis (PCA), Pearson correlations, and CUR decomposition of the selected features have yielded promising results in literature [348].

We will showcase two approaches to the generation of features, and highlight how our choices have long-term propagation effects in our training process: 1) A naive baseline grid of 2:1 radial and angular symmetry functions: $N_{func,ele} = 49$, $R_c = 6$Å, λ=$\pm$ 1, $\zeta = 1$. 2) Features selected from CUR decomposition within a dense grid of 240 symmetry functions per element pa:$N_{func,ele} = 50$, $R_c = 6$Å, λ=$\pm$ 1, ζ = [1,4,8,16]

After creating a baseline grid of features, and isolating our initial database, this set of features and initial data will be sequestered into Family 1. Using our initial database, we

assess the expressivity of the dense grid of features and determine which symmetry functions are acted on for each atomic configuration the most, while not duplicating information in the set. For a matrix of features, each element and its symmetry function is expressed as a row. To select the highest rank symmetry functions from a matrix M, we apply CUR decomposition [349], [350], to select the highest valued symmetry functions, without changing the original shape or representation of the data. Other methods such as PCA, or the Pearson Correlation, while useful, do not retain the information from the original feature itself. For a matrix, M, a low-rank approximation of the matrix is provided as:

$$\tilde{M} \approx CUR, \tag{5.7}$$

where C and R are rows and columns respectively from $\tilde{M}$, and U a scoring matrix given by:

$$\pi_c = \Sigma_{\mathrm{j}=1}^{k} (v_c^{(\mathrm{j})})^2 \tag{5.8}$$

The respective scores from rows and columns allow us to slowly remove data, while still approximating the original matrix. The remaining features left can be fixed for any given N. We choose N=50 for the lead-up to the iterative workflow. Included in the Supplemental Information is a table for each set of symmetry functions, and their hyper-parameters.

Neural Network Training

All tools used for scaling, normalizing, and training our neural networks were done using the *n2p2* neural network package [351]. Recent extensions of this package for training optimization include the introduction of Standard Kalman filter training [352]. We use the recommended parameters of $\epsilon = 10^{-2}$, $q_{min} = 10^{-6}$, η=0.01. From our initial database 10% of the data is isolated for testing, while using 90% for training.

Like many problems in machine learning, the confidence in our model to produce realistic results is critical. To test errors we often aggregate results from many shuffled initializers, splits in the training database, or cross-fold validation. However, the impact of using a poorly trained NNP on an interative workflow could have cascading effects. Since we

rely on the NNP expression for energies and forces, any poor prediction ultimately leads to unstable molecular dynamics.

Network Variability Classification

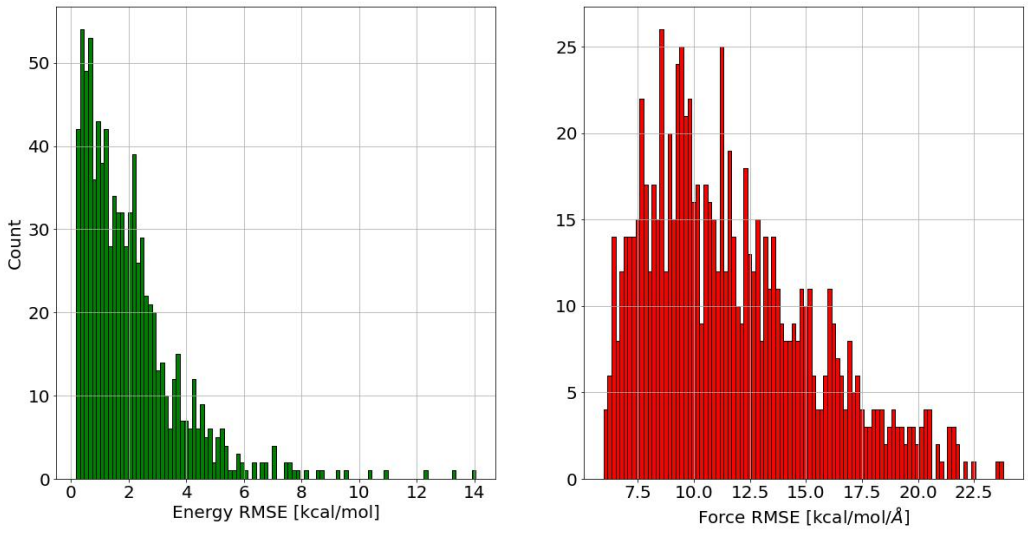

Figure 5.13. N=1000 Error distributions

Given the high dimensionality of deep neural networks, their solutions can often be perturbed by simple modifications to the initial force weights or training data. With poor initialization a network will find a local minima, but lack the proper momentum to find the global solution. Depending on the algorithm, the method and selection of weights can have drastic effects on model performance [353].

To help characterize the variability, and uncertainty of our predictions we train multiple networks on our initial database and classify the distributions of error. With reduced network architecture of 15 nodes in 2 hidden layers ([15,15]) we trained 1000 networks on the same subset of data, but with a different random seed for initializing force weights.

Fig. 5.13 shows the distribution of minimum energy and force errors for the 1000 trained networks. Interestingly, the energy and force distributions seem to follow distinctly shaped curves. We classify the error histograms by comparing the empirically found data

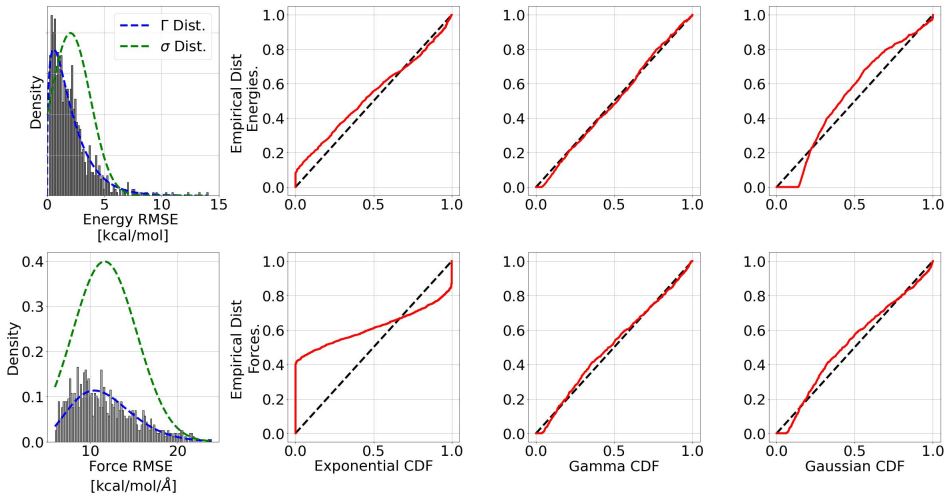

Figure 5.14. Error distributions and correlations for NNP variability testing. N = 1000. Exponential, gamma and Gaussian distributions considered, with gamma being the most likely distribution.

with common distribution representations. We include pairwise comparisons for exponential, gamma, Gaussian, and Poisson distributions for the energy and force distributions. With respect to the force and energy distributions, each set of data correlate highly to a gamma like distribution.

Our results clearly illustrate the impact of force weight initialization on model performance, and elucidate the need for a third Family to compare in our iterative training approach. Rather than accepting the results of a trained model, and blindly running a new MD simulation to collect trajectories, five different networks will be trained and evaluated. Since the trajectories of the MD simulation are highly dependent on the forces, we will select the network that performs the best (lowest root-mean-squared-error [RMSE]) for the forces. While the total energy of the system is a highly important metric, the volume of our force data outweighs our energy 3:1. If a model scores high on error it could be penalized for one configuration in our dataset, rather than the general trend of misaligned forces. Assuming that all networks within the iterative loop follow a similar distribution to our baseline sample, the probability of selecting a network with an RMSE below the mean of the distribution for five networks is within safe bounds of operation.

Beginning with the same DFT database for Ge/Sb/Te, GeTe, and GST, we showcase three families of training procedures for generating a stable NNP:

1. Family 1: A generated grid of 2:1 radial and angular features. Single model per generation.

2. Family 2: A CUR decomposition selected grid from dense set. Single model per generation.

3. Family 3: A CUR decomposition selected grid from dense set. Five trained models per generation, with minimum force RMSE selected.

5.3.4 Molecular Dynamics with Neural Network Potentials

A generational trained NNP was used in our simulations for dynamic extrapolation of our database. All MD simulations were performed using the LAMMPS software package from

Sandia National Laboratories [39], [40] with the HDNNP implementation [354]. Visualization of atomic structures and structure identification was performed with OVITO [238].

A timestep of 1 fs was used throughout, with damping constants of 0.1 ps and 1 ps for the Nose-Hoover thermostat and barostat respectively [41]–[43]. All atomic structures consisted of 144 atoms for the hexagonal, cubic, amorphous, and liquid phases.

5.3.5 Evaluation of Iterative Training

Database Enrichment

Using our initial DFT+MD database, we train NNPs according to the method prescribed by their Family, we annotate this as Gen 1.1. Since our NNPs have shown high variability, and inability to extrapolate well, we restrict our NNP+MD simulations to constant volume ensembles (NVT) in the early stages. We perform NNP+MD simulations of the hexagonal, cubic, amorphous, and liquid phases of GST at varied temperatures, and amorphous densities. Each MD simulation is held for 1 ns at isothermal temperatures of 300, 400, 600, 800 K for the solid phases, and 1100 K for the liquid phase. From these trajectories 10 snapshots of the configuration are sampled with even distribution across the runtime.

After five generations of constant volume ensemble simulations, we begin to add simulations with constant pressure ensembles (NPT). Howver, we continue to simulate and sample from NVT ensembles to ensure decent ratio balances of our training database. For the remaining generations [1.5 - 1.15] we use both NVT and NPT ensembles of all the phases to fully explore the trajectories between each. Fig. 5.15

Network training results

Beginning with Gen 1.1, we compare the three families ability to train, and accurately reproduce the results of a testing subset. We evaluate parity plots for the model's ability to reproduce training and testing data, as well as evaluate the distribution of the error. Evaluations of Gen1.1, Gen1.5, Gen1.10, and Gen1.15 for each family are shown below in Fig. 5.16. Highlighted in red, are the results of the iterative loop for Family 1. The coarse grid of features used here show decent representation of the system energies for training and testing,

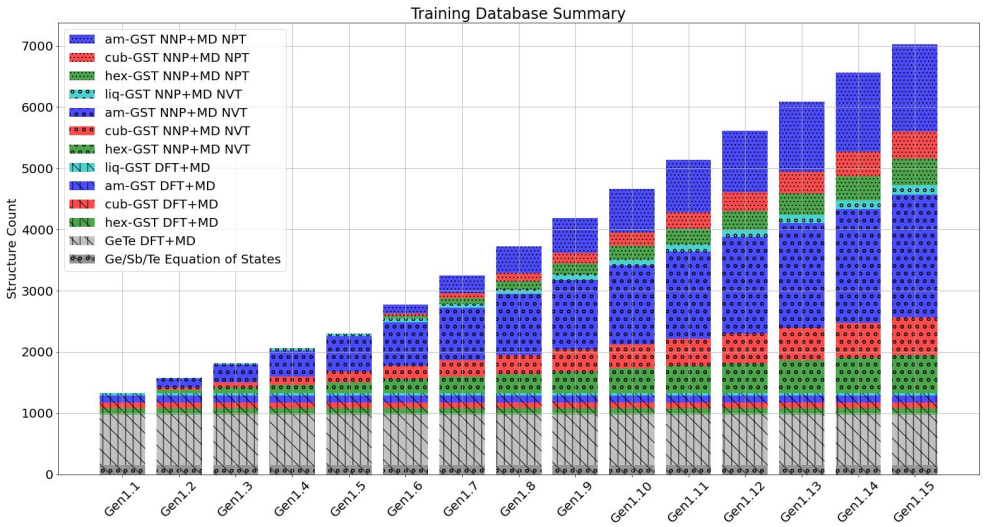

Figure 5.15. Configuration database for GST. Vertical and diagonal slashes correspond to our initial DFT+MD database of GST/GeTe/Ge/Sb/Te. NVT NNP+MD [large circles] and NPT NNP+MD [small circles] contruct the primary majority of the Gen 1.15 database

but the forces begin to deviate to large degrees with more configurations. As a consequence of using a poorly trained network for extrapolative purposes it is likely that many of the configurations obtained by Gen 1.15 in Family 1 are more harmful than helpful.

In green and blue are Family 2 and 3 respectively. Using the CUR selected features we score lower in RMSE at earlier generations compared to Family 1, and with Family 3 we yield a much better convergence of the system forces. The network struggles to reproduce energies of some of the configurations as the system becomes more complex, but the well controlled forces are more essential for dynamics. Comparing the three families in Fig. 5.17 we show the variability error scores across generations. Using a coarse grid in Family 1 with only one network per training generation yields often chaotic results as shown in Fig. 5.16. Both the energy and force error scores can sometimes outperform the CUR selected features if they are lucky, but often this is not the case. When observing the CUR selected features (Family 2 and 3), we see consistently better error scores across the generations. Family 2 shows a better generalization of the system energies, while Family 3 shows a better convergence and control of the system forces, with a few spikes in total energy error score. However, Family 3 strongly outperforms Family 2 by the time we approach Gen 1.10 - Gen 1.15.

To help elucidate some of the variability associated with the networks, in addition to what we described in Fig. 5.13, we show the errors for energy and force of each network trained for Family 3. Fig. 5.18 shows the collected errors for all 75 networks in the iterative loop. Here we show average network scores in vertical dashed lines, the network error in colored bars, and the selected network with a star. As we can see by the attached lines the convergence of energies is reached by Gen 1.15 with some noise in between, and the forces of the selected network are always well below the average.

5.3.6 NNP Validation: Static & Dynamic

As a test of validation, we compare our NNP performance with a long time-scale DFT+MD simulation. Based on a recrystallization study of amorphous GST, we repeat classical work done by Hegedus et al. [313]. An amorphous cell of GST is fixed at 6.11 g/cm^3 density, and held at 600K under NVT conditions. During the DFT run significant

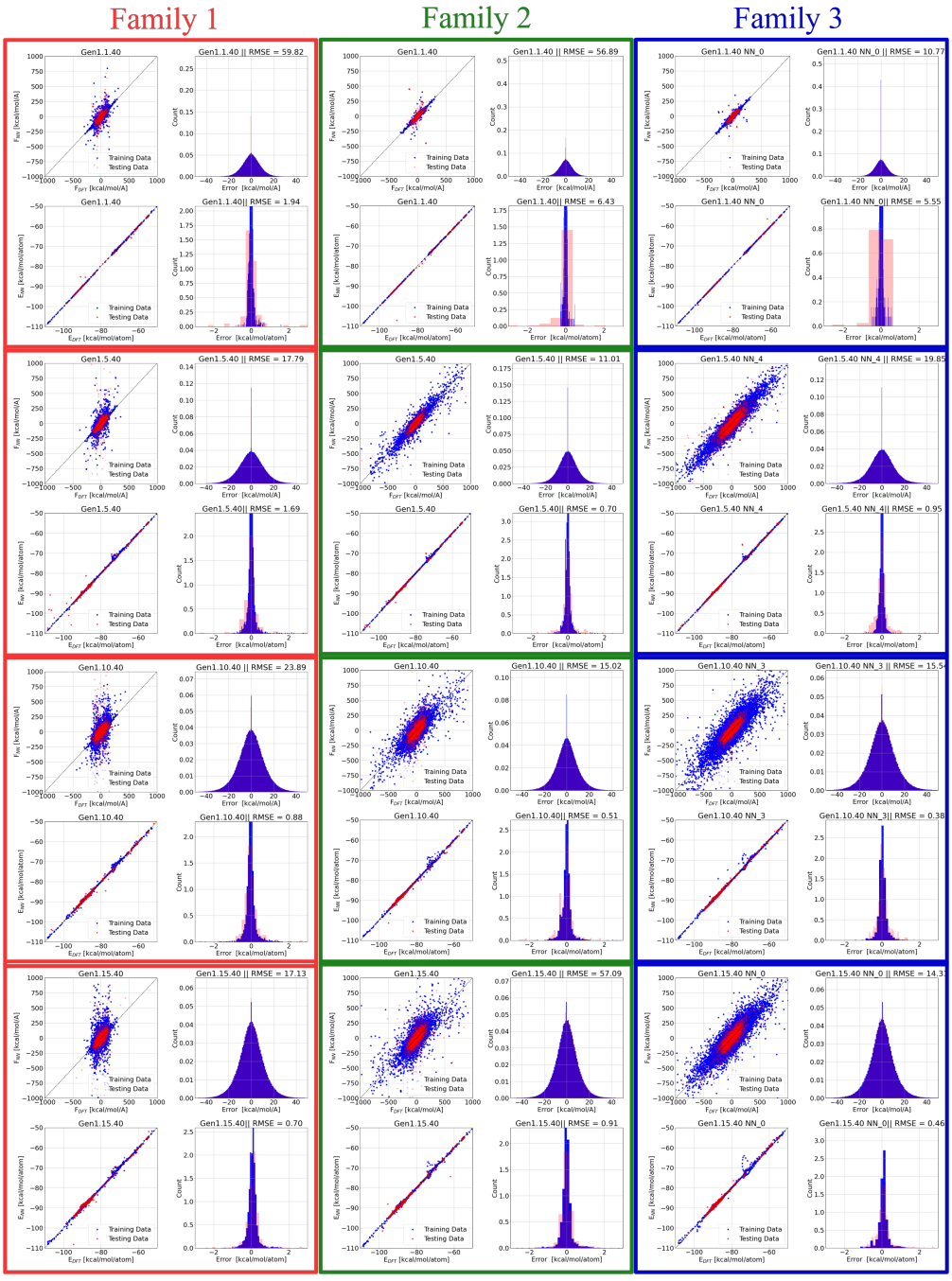

Figure 5.16. Parity and distribution plots for Families 1, 2, 3, and Generations 1, 5, 10, and 15.

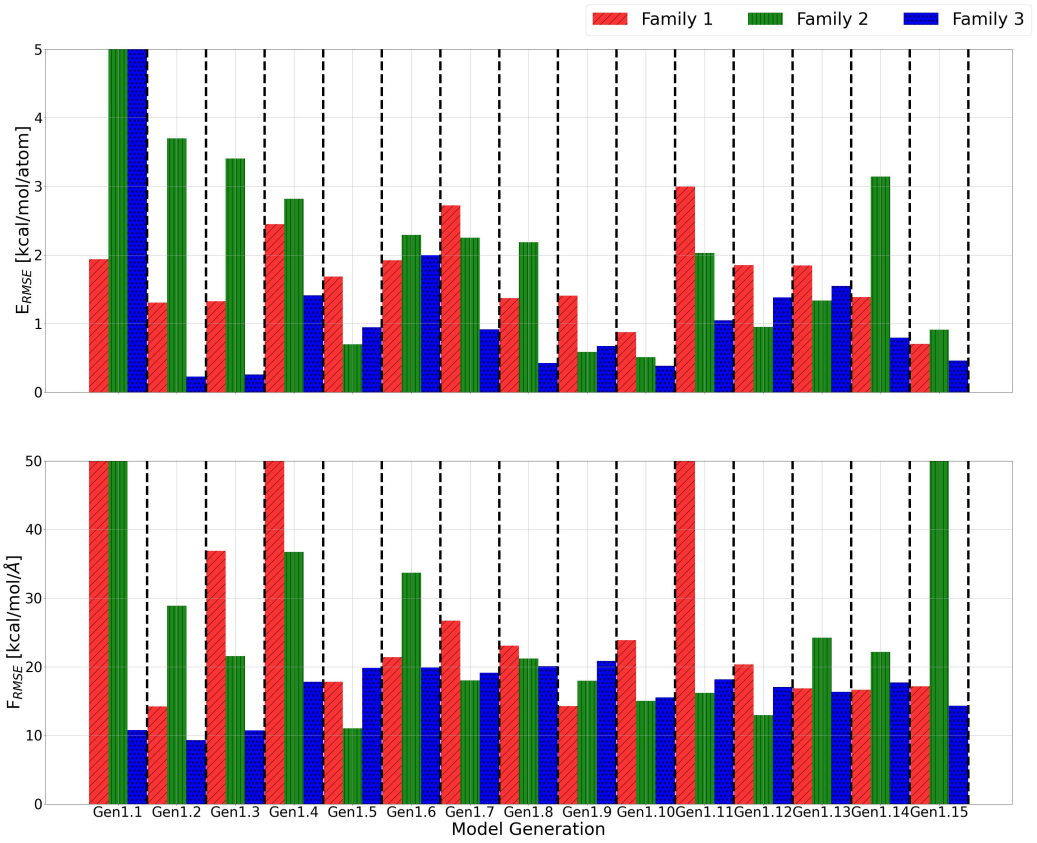

Figure 5.17. Energy [Top] and force [Bottom] RMSE scores for each Family and Generation.

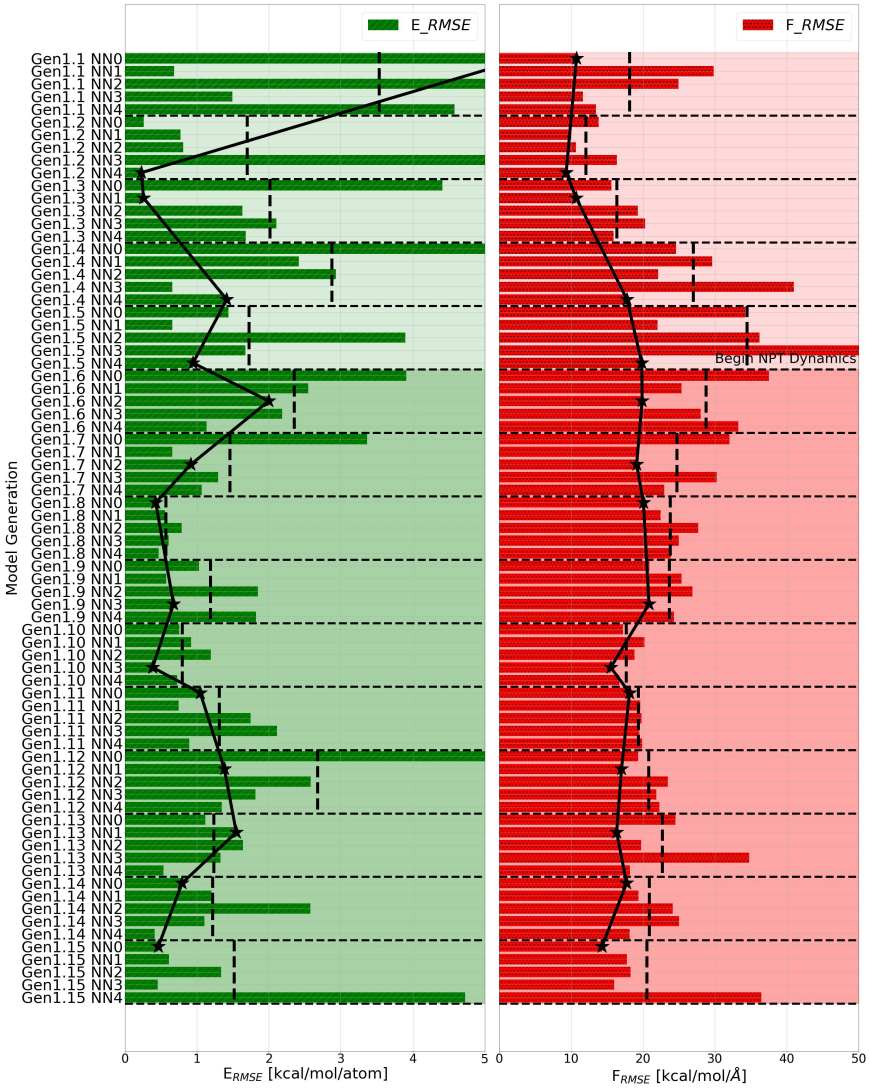

Figure 5.18. Scores of the five networks at each Generation in Family 3. Darker shade indicates when NNP+MD dynamics included NPT ensembles. Stars represent selected network for continued iterative loop study. Errors shown in the star selection correspond to errors in Fig. 5.17.

recrystallization can be seen, and we track the evolution of the structure as it orders. We choose to evaluate our force field performance on two types of validation: static and dynamic. We define static validation as the common practice of sequestering high-fidelity data from our training set, and using our force field to reproduce the energies of a set of configurations. This highlights our model's ability to reproduce data from an existing dataset, but it does not account for generalizability of the model, or its ability to extrapolate to new trajectories with reasonable energetic states. Here we evaluate the dynamic process involved from using a force field to run full MD simulations. We attempt to reproduce the same amorphous recrystallization process from our DFT+MD validation set, but this time starting with an initial configuration and allowing the force field to explore the process on its own.

Static Validation

Fig. 5.19 shows a comparison of the three families and their ability to reproduce energies of the DFT+MD simulation run. In Gen 1.1, all three families show high accuracy, but as additional data is added to our database the model has a more difficult time reproducing simpler DFT configurations. However, as we will see in the next section, the ability to reproduce a configuration is a secondary goal to the ability to run well characterized MD simulations. This energy shift in the validation set is a penalty our model must pay to ensure decent generalization to other datasets. Comparing the three families we see that Family 3's force fields consistently out perform the other two with respect to DFT+MD validation of the energy. By Gen1.15 we also capture a plateau in Family 3's RMSE error, indicating that we have reached a saturation point with respect to our validation set.

Dynamic Validation

Beginning with an amorphous sample at 6.11 g/cm^3, we use our force field to run an MD simulation at 600K to reproduce our DFT+MD validation set. As discussed before, but highlighted in Fig. 5.19, we see the poor ability of our model to accurately captue dynamic effects of new configurations. In the early generations it is common for certain configurations to enter highly energetically unfavorable states since the model itself does not understand

anything beyond its initial DFT database. However, these high energy configurations help guide the model in later generations since it understand which states to avoid, and which to drive itself towards to minimize total system energy. As shown by Gen1.15, we see decent agreement with the potential energy curve of the DFT+MD run up to 250 ps. While our model does not capture the initial nucleation event, and subsequent recrystallization on the same initial time scale as our DFT+MD run, we do see consistent recrystallization at 750 ps - 1 ns of anneal time.

In addition to tracking the potential energy, we can use the trajectory data generated by the NNP+MD run to evaluate the vibrational density of states in the system. We collect velocity components of individual atoms in a system at incremental timesteps and perform a Fourier transform on the time-dependent atomic velocities. We follow the methods of Berens et al. [355] for calculating the power spectrum of a system:

$$P(\omega) = \frac{\beta\tau}{N} \sum_{\mathrm{j}=1}^{3N} m\mathrm{j} \Big| \sum_{n=0}^{N-1} v\mathrm{j}(n\Delta t) \mathrm{e}^{-\mathrm{i}2\pi\omega n\Delta t} \Big|^2 \tag{5.9}$$

where v(t) are the atomic velocities at time t, β is defined as $(k_B T)^{-1}$ with T as absolute temperature, τ is total sampling period, m_j is the atomic mass of atom j, Δt is the sampling rate, and N is the number of snapshots analyzed. We use a fast Fourier transform to solve Eq. 5.9. The vibrational density of states is obtained by dividing Eq. 5.9 by $1/2k_B T$ under thermal equilibrium. At this condition every vibrational mode in a classical system will contribute $1/2k_B T$ of kinetic energy.

Structural information including radial distribution functions (RDF), angular distributions functions (ADF), and vibrational density of states (VibDOS) have been included for both NNP results and validation sets for DFT. In Fig. 5.20 we show the comparisons of each structural value, with the evolution of structure between Gen 1.1 and Gen 1.15. In red is Gen 1.1, with the RDF and ADF for each respective phase. Here we see poor agreement with the structural information, and if we follow the solid lines from blue to red we see worse agreement with the black dashed lines corresponding to DFT. However, with iterations in the dynamic loop the NNP grows to understand which trajectories are favorable for phases, and recreates each phase with excellent precision. Shown in the green highlight is Gen 1.15. The RDF

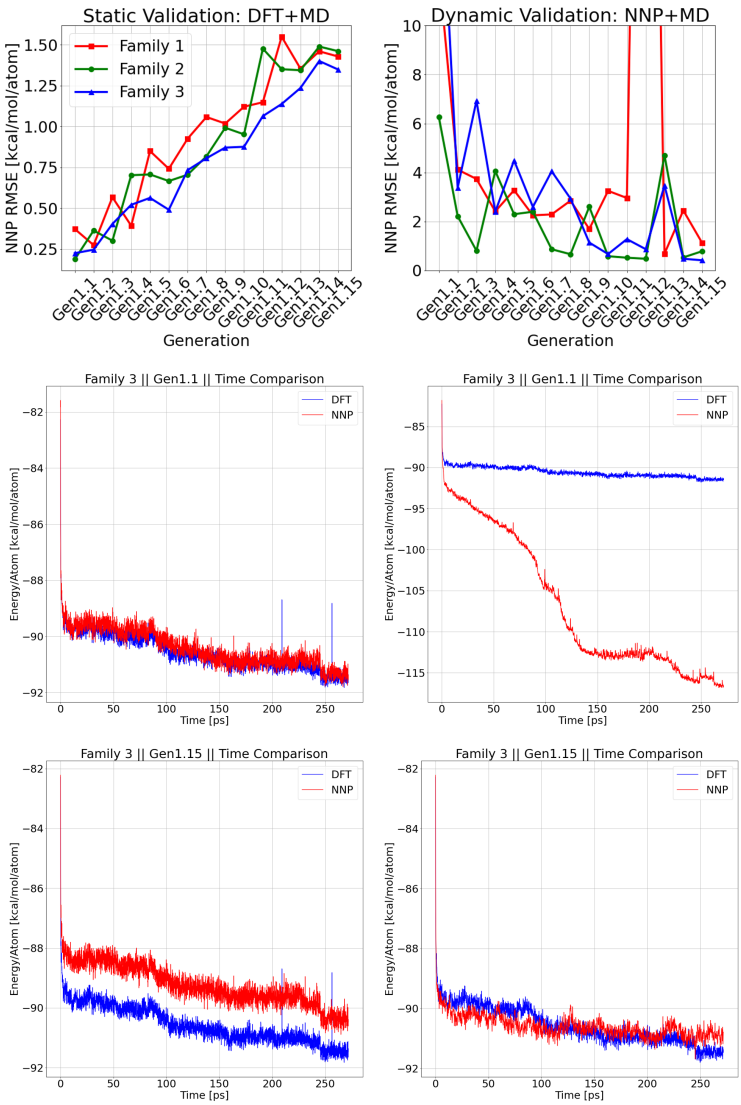

Figure 5.19. Static [left] and dynamic [right] validation. Static validation set defined as sequestered recyrstallization simulation of amorphous GST at 6.11 g/cm^3 [313], and the ability for a trained NNP force field to reproduce the energy of a pre-defined configuration. Dynamic validation set defined as an extrapolated dataset, using the NNP force field to drive molecular dynamics of an amorphous GST cell at 6.11 g/cm^3. The difference in potential energy landscape is shown for static and dynamic validation of Gen1.1 and Gen1.15. Y-axis scale bars are different for each plot, and are kept so intentionally.

and ADF for hexagonal and cubic agree well, with the amorphous showing significant signs of recrystallization. Individual atomic contributions to vibrational modes are in reasonable agreement with DFT by Gen 1.15.

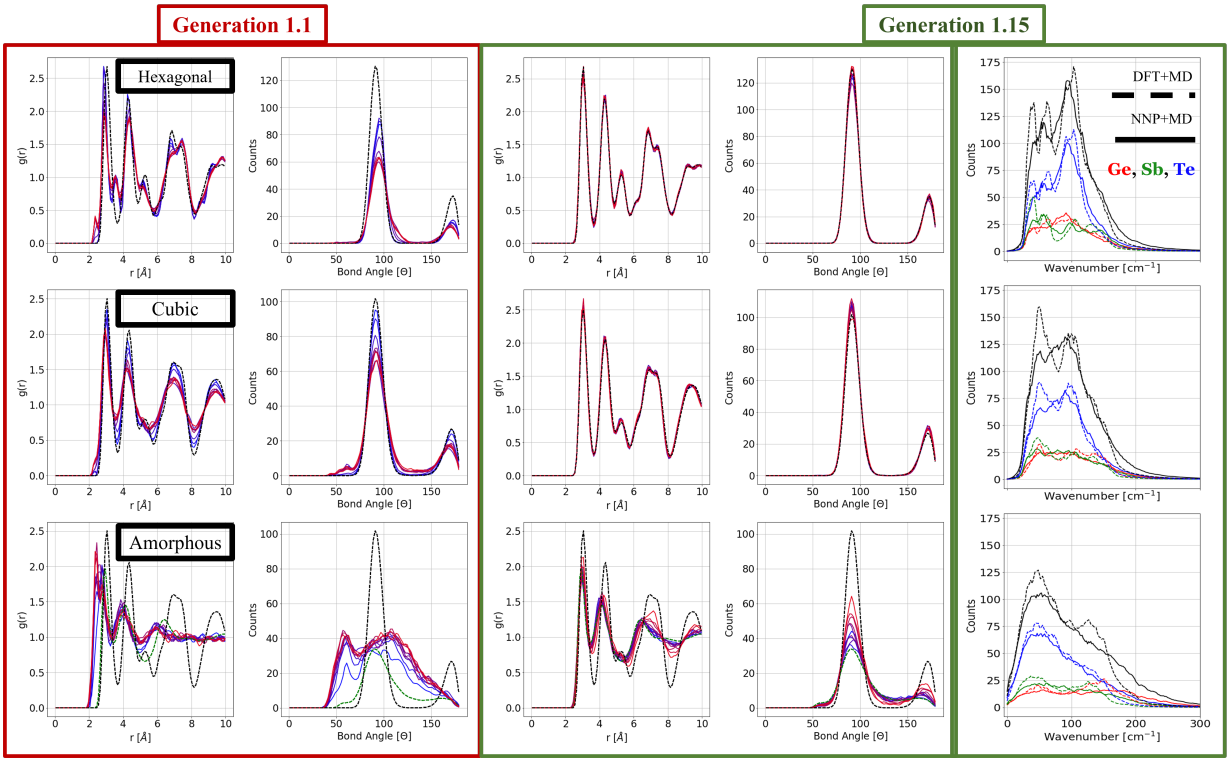

Figure 5.20. Gen 1.1 (red highlight) RDF [left] and ADF [right] for Hexagonal [Top], Cubic [Middle], and Amorphous [Bottom]. Solid lines represent NNP predicted values at 100 ps intervals and 10 ps of time averaging. Temporal evolution from 100 ps to 1 ns is highlighted from blue to red. Dashed black lines represent the DFT structure analysis for a given phase with the exception of the amorphous. The cubic phase is plotted in black as a representative guide to show recrystallization, with the amorphous structure shown in dashed green. Gen 1.15 (green highlight) shows excellent agreement with RDF, ADF, vibrational density of states, for all phases, and shows significant recrystallization of the amorphous phase to the cubic.

In addition to capturing the physical modes of vibration, we are interested in how well our potential can reproduce physical quantities such as the minimum stress density calculated by DFT. In Gen 1.1 through Gen1.4, we do not run any NPT ensemble dynamics due to the erratic behavior of the NNP force field in early stages. Gen1.5 and on we begin sampling non-constant variables such as the density and collecting the stochastic average.

For each temperature and initial phase that we simulate we collect the thermodynamic states and assess their performance with DFT predictions. Fig. 5.21 shows the difference between the average NNP+MD density and the minimum stress DFT+MD density. Here we see poor agreement between the two methods of simulation in early generations, but well converged densities across all temperatures at later generations. Examples of highly irregular configurations sampled in our database can be found at higher temperatures of the middle generations.

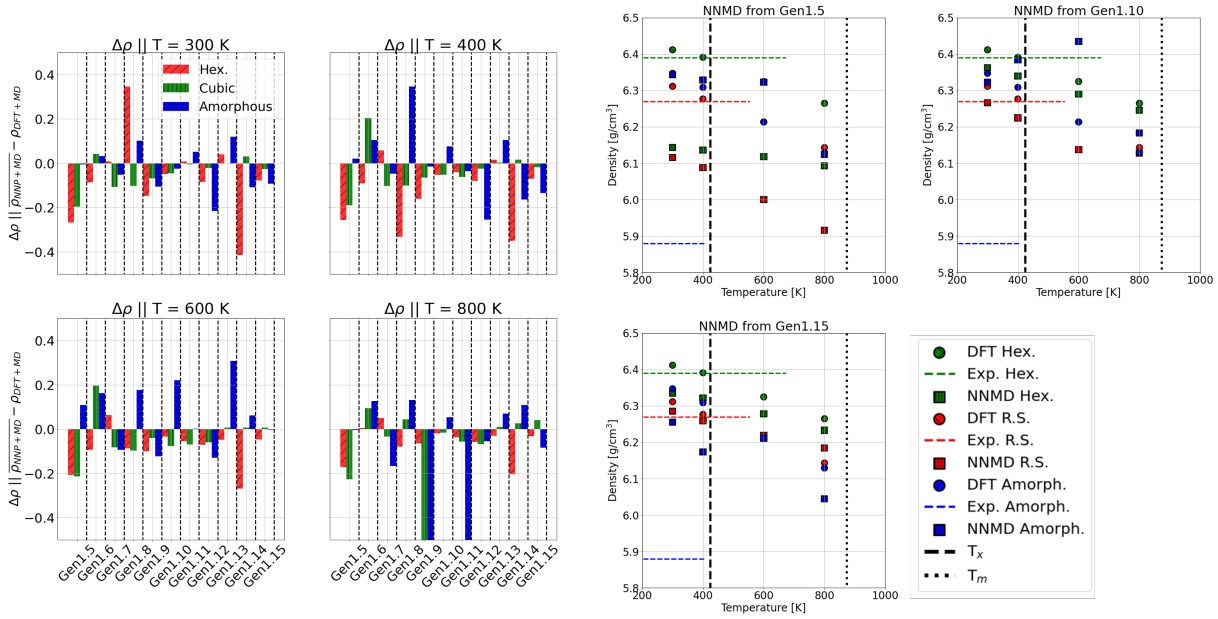

Figure 5.21. Density analysis of DFT+MD [circle] and NNP+MD [square] simulations. Horizontal dashed lines correspond to the experimental density [344], with vertical dashed lines the phase transition temperature. DFT densities show good agreement with experimental results. NNP densities from NPT simulations are shown as the average density of the simulation.

5.3.7 Discussion

One of the highlights to come from this approach is a potential that was able to learn the kinetics of recrystallization on its own, but with physical limitations that resemble realistic nucleation kinetics. Of the initial DFT trajectories that were used for Gen1.1 training, only 20 ps of MD+DFT configurations were acquired - much before any onset of recrystallization

in the amorphous phase. It is important to highlight that initial trajectory information in our database included only stable or metastable phase dynamics, and no transient information. The value of this gap is highlighted in our approach since the path between states was extrapolated, and solved via the iterative approach without the need for manual insertion of those states.

When creating additional trajectories for DFT sampling, we run varied densities of the amorphous phase at temperatures above and below the recrystallization temperature. Two sets of runs are generated under NVT and NPT conditions. In the early stages of the iterative loop, running 144 atom simulations, we do not see any evidence of ordering, recrystallization, or long range structure formation of the amorphous phase. However, by Gen1.11 in Family 2 and 3 we begin to see evidence of rapid nucleation, and recrystallization of the system albeit at a longer timescale of 700 ps. Systems that are able to relax in the NPT ensemble tend to recrystallize more than the NVT ensemble. These results are encouraging, and consistent with DFT and GAP+MD literature [313], [319]. However, given the small system size if there is any amount of nucleation there is a high probability of the system rapidly recrystallizing due to small boundary conditions, and lack of secondary amorphous phase to pin the process.

A natural next step when considering recrystallization of a system is the critical nucleus needed to propagate a spontaneous phase change. Especially in the transition from amorphous to crystalline, the energy barrier needed for a group of atoms to nucleate without being annihilated by the surrounding disordered phase can be large. While extensive work has been done with *ab initio* to study the recrystallization kinetics of GST, due to their limited size effects they tend to over-characterize the speed of phase transition by orders of magnitude [303]. This can be remedied by larger atomic supercells, but with *ab initio* methods the compuational cost is quite high. The trained GAP for GST has been used to study systems of up to 24,300 atoms, and with it they investigated the effect of quench rates and system size on the potential energy of the system, but there has yet to be an analysis on the critical nucleus and recrystallization rates. To harness the power of the NNP, we scale a rocksalt system to ∼7000 atoms for preliminary study of the critical nucleus needed for GST recrystallization, as well as assess first order approximations of the recrystallization rate.

Beginning with a replicated system of cubic rocksalt with 3 nm box edge lengths, an NPT thermostat is applied at 600 K to equilibrate the structure to a relaxed density for 10 ps. After thermalization we select a region of atoms that we will denote as the *seed atoms*. This region selection is performed by isolating a group of atoms within a sphere of a selected radius. The *seed atoms* are given a separate thermostat from the rest of the system while a melt-quench protocol is performed. NVT conditions are given to the *seed atoms* and the melt atoms. The *melt atoms* temperature is increased to 1500 K in 10 ps, above the melting point, to melt the system, while the seed atoms are held at 600K. The *melt atoms* are held at 1500 K for 10 ps to ensure well diffused liquid, and rapidly quenched to 600 K in 10 ps ($\sim$100 K ps^{-1}. Once both sets of atoms are at 600 K the NVT thermostats are removed, and a global NPT thermostat is applied to all atoms at 600 K to anneal for 5 ns. During the melt-quench protocol some of the *seed atoms* are melted, but continue to be individually thermostatted at 600 K. While erroneous, the volume comparison of atoms in *seed atoms* that may be locally grouped in the *melt atoms* is statistically insignificant to have an effect on simulation effects.

For crystalline characterization we combined many of the tools available in Ovito[238] for both structure and cluster analysis. The steps included are the pipeline processes involved for the Ovito script to map cluster properties of crystalline phases:

1. Using the Polyhedtral Template Matching (PTM) algorithm [86], select simple cubic (SC) crystal structure and RMSD cutoff of 0.15.

2. Select Type Modifier: Select atoms denoted as *Other*. Here *Other* is anything non detected as *SC*. The PTM algorithm unfortunately does not recognize vacancy neighborhoods that are characteristic to the cubic phase of GST, but rather characterizes them as defects, or *Other*. The selection of type *Other* overselects the number of atoms not in the crystalline phase, but it still allows for initial screening of the structure.

3. Delete Selected Atoms of type *Other*, leaving only *SC*.

4. Cluster analysis on remaining *SC* atoms with cutoff radius slightly above lattice bonding (3.5 Å). Sort clusters by size, and report radius of gyration.

Fig. 5.22 shows the comparative results of different initial radii selection for *seed atoms*, and the visual representation over the 5 ns anneal at 600 K. Initial visualization starts at 20 ps, where the atomic rendering is a slice in the middle of the simulation cell to capture the *seed atoms*. For very small selected radii, the *seed atoms* do not survive the initial ramp to 1500 K, and are consumed by the melt. Systems with an initial nucleus smaller than 6 Å were unlikely to survive the melt-quench prior to the system anneal. However, this gives these systems an opportunity to demonstrate their ability to homogeneously nucleate. While these systems do not give us enough statistical sampling on the time to nucleation since we have only one primary nuclei, we do collect critical sampling on the necessary radius of gyration to generate a feasible seed for recrystallization. For the system of initial seed size 4 and 6 Å the atoms within the cell fluctuate between amorphous and crystalline, with minor regions connecting, but ultimately unable to form a stable crystallite. In time ranges from 1-2 ns, we begin to see that once a threshold of approximately 7 Å is crossed the seed can continue to grow, and ultimately recrystallize a majority of the simulation cell, albeit with significant voids. The same critical size barrier can be seen for systems of *seed atoms* initial radii greater than 8 Å, where both continued growth is tracked from the initial surviving seed, and other nuclei that spontaneously form from the amorphous phase.

To overcome the challenge of poor statistical sampling for number of nuclei in a simulation, we scale our rocksalt system to the largest GST simulation in current literate. Using a cubic cell with 18 nm box lengths, a system of 194,400 atoms is explored with an initial group of *seed atoms* as a nuclei, and a majority simulation cell in the amorphous. The initial radii of the crystalline seed was grouped at 30 Å, one third the size of a sphere extending to the edge of the box. This was to ensure that a large cell could be studied for induced nucleation growth, but small enough that a large number of homogenous crystallization events can occur, and potentially interact with the larger seed. The same melt-quench protocol was performed on the system using the methods described above, with the exception of the quench-rate being reduced to 10 K ps $^{-1}$.

Confirming initial velocity characterizations of the smaller simulation cells, the large 30 Å seed grows with a velocity of roughly 1 m s^{-1}, in agreement with experimental observations in TEM [356]. As the seed grows, there is little evidence of spontaneous recrystallization prior

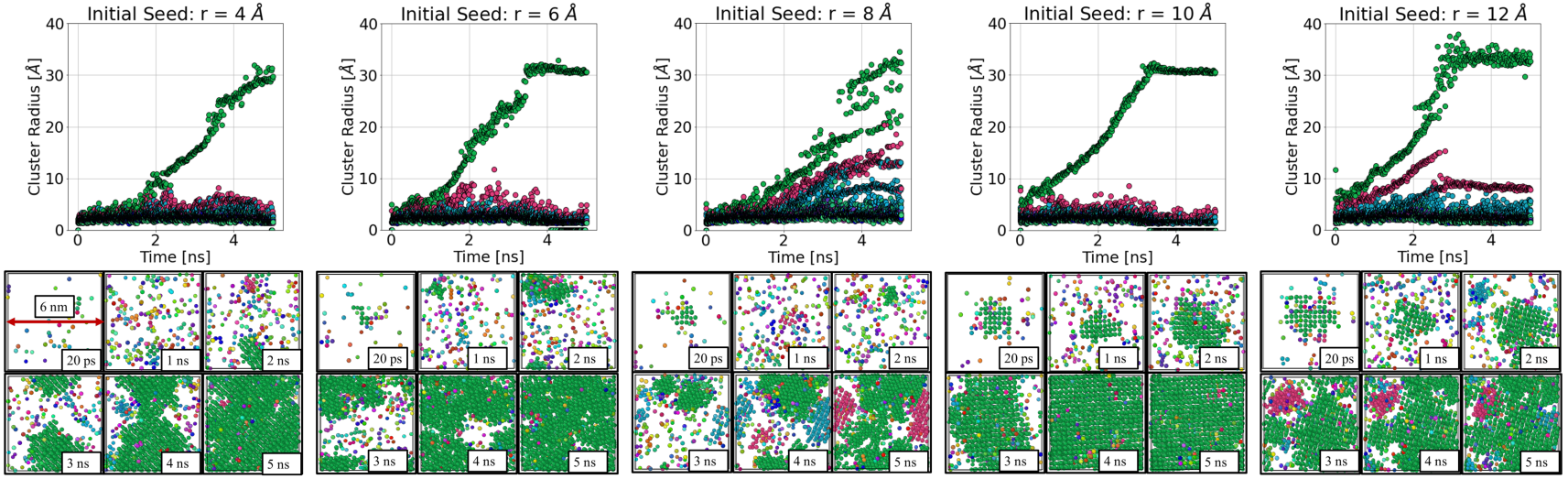

Figure 5.22. Recrystallization snapshots of 7000 atom GST during 600 K anneal. Initial 'seed' of spherical region modified to assess stability during melt-quench protocol. Colors in scatter plot and atomic configurations are matched to aid visualization.

to 1 ns. After an incubation period, a number of clusters begin to appear after they reach the critical threshold of 7 Å. The re-occurence of critical clusters with a larger simulation cell compared to the system of 7000 atoms is re-assuring for the validity of the potential across different length-scales outside of the small periodic systems trained on for DFT.

Characterizing recrystallization as a function of quench rate, system size, and anneal temperature are all factors that will be the basis of future explored work. In this brief study we established a solution to a previously challenging problem of the critical nucleus for GST recrystallization, and have shown that in both homogeneous and induced homogeneous recrystallization we match well with physically expected quantities.

5.3.8 Conclusions

In this work, we explore three different methods for generating an enriched database. Using an iterative loop, we leverage trained NNPs to extrapolate to configurations our model has yet to see in an original database. Sampling from these trajectories allows us to supplement our database with well converged low energy configurations, and non-physical high energy configurations. The duality of acquiring these sets of data help the model to learn which routes are favorable for dynamics, versus poor extrapolations of a model that is unable to generalize outside of a rigid training set.

Most interestingly, even though our initial database only contained trajectories from the initial phases (hexagonal, cubic, amorphous, liquid, GeTe, Ge/Sb/Te), and no transitory information, our NNPs for Family 2 and Family 3 are able to successfully capture the phase transition from amorphous to the metastable cubic phase. Traditional DFT+MD simulations of recrystallization are highly costly to include given the minimum simulation time of 100 ps for an initial onset of recrystallization. While including these trajectories may have allowed our model to find these dynamic recystallization configurations sooner, we assert that the lower barrier of required trajectories is highly advantageous for extended applications of the NNP iterative loop method.

A secondary outcome of this study is the classification of the NNP network initialization, and the effect of this variability on model performance. We often find that if a single network

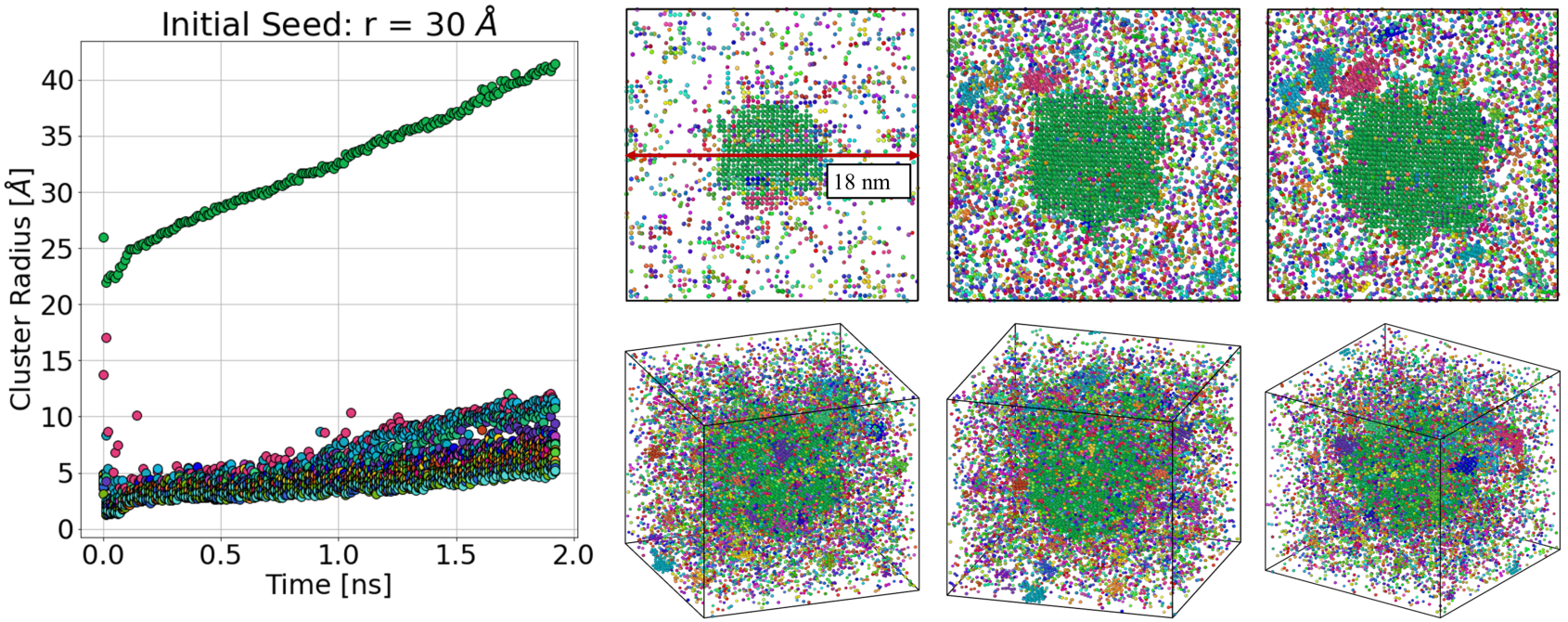

Figure 5.23. Recrystallization snapshots 194,400 atom GST anneal at 600K. Initial 'seed' of 30 Å. Slice snapshots at 20 ps, 1 ns, and 2 ns of the seed are included in the [Top] row, with different aspect views at 2 ns of recrystallization shown in the [Bottom]. Colors in scatter plot and atomic configurations are matched to aid visualization.

is poorly converged, the cascading effects of sampling erroneous trajectories can overwhelm our database with strange data. Using a minimum force RMSE selection method we create a pseudo-control scheme that provides a higher probability of success in finding a reasonable NNP for dynamics.

Finally, we show that starting with a small database of 1000 configurations is sufficient for initial training, and acts well for describing our chemical space. However, as highlighted in our work, the use of these NNPs to extrapolate outside of a rigid database in early generations must be used with caution. The implementation of a semi-autonomous workflow for iterative learning allows us to quickly sample space outside of our initial DFT+MD database with little to no human intervention.

5.4 Final Remarks

As we refine our scope of material problem to the atomistic level, the unique fingerprint and configuration of an atom is highly relevant to accurate material prediction. Assuming that all atoms exemplify the same uniform property no longer applies for multi-component systems in both the crystalline and liquid phases. To capture the localized effects of energy, stress, and forces we train neural networks with geometric based featurizers to capture atomic level fluctuations in behavior, and the drive dynamic processes with the obtained energies and forces. The combination of chemical based featurizers described in Ch. 4 and geometric descriptors here in Ch. 5 can be used in tandem for higher fidelity feature sets, and for enhanced material screening applications.

6. EDUCATIONAL SUPPORT FOR THE NEXT GENERATION WORKFORCE

6.1 Introduction

In this chapter we will briefly discuss efforts made through this thesis to align with the third mission statement of the MGI for the development of the next generation workforce, but also the enhancement and supplement we can provide to the current field. Primarily through the use of nanoHUB, we have pushed to have our work be as open as possible for publication review, and for reduced barrier of entry to novice researchers. The products of this work help shift fields of academic, educational, and social growth through easy to use tools and accompanied instructional guides. First we will discuss the shift in our academic discussion of supplementary material. Included with our more recent works are functional Jupyter Notebooks [300] with live code and markdown cells for well documented operations. These notebooks are available on nanoHUB, and allow users to run the code from their web-browser without the need for any code installation or supplementary libraries. The code is open source, and files including the code and data can be downloaded as a zip file. The low entry barrier for code access, operation, and download is ideal for novice users, or those with limited resources for code reproduction. For more advanced users the code is available, and can be transferred to a preferred platform, or Python based environment. Next, our contributions during a rather strange period of our history will be included for serendipitous reasons. These are included as a showcase of the magnitude of impact well applied cyberinfrastructure tools can have, and for setting the stage of future workforce development. Finally, we will conclude with a brief collaboration between academic consortium, nanoHUB, and educational practitioners.

6.2 Enhancement of Supplemental Materials

In traditional work, figures or tables that do not help with the main text of a document are placed in supplemental information. Even one step further removed is the data itself that was used to create the supplemental figures. In best practice with F.A.I.R. data guidelines

[11], we enhance Git practices with fully operatable code and data powered by cloud-based cyberinfrastructure. The code can be modified directly from the web browser to re-train models, modify plots, and perform additional queries from databases. Using these interactive notebooks as supplemental information helps to lower the entry barrier for new researchers in the field, but also as an interactive journal for peer-review and internal discussion.

6.2.1 High-Temperature Property Explorer: *htoxideprop*

Included in the tool *htoxideprop* are reference notebooks for data acquisition and model training for high-temperature oxides property prediction [196] including supplemental information for work in Ch. 4. Composition featurizers are used to train models for stiffness, vacancy formation energy, and melting temperature. Physics derived descriptors such as the Lindemann law for melting are incorporated into model calculations to show the viability of low dataset predictions with transfer learning and feature engineering. Property models are trained for material screening processes, and uncertainties within are quantified for model validation. As an example of the querying workflow, Fig. 6.1 shows the combination of live code for accessing the Materials Project database, filtering and parsing of data, with visualization included in a live code cell. The notebook includes documented code for both review and educational purposes.

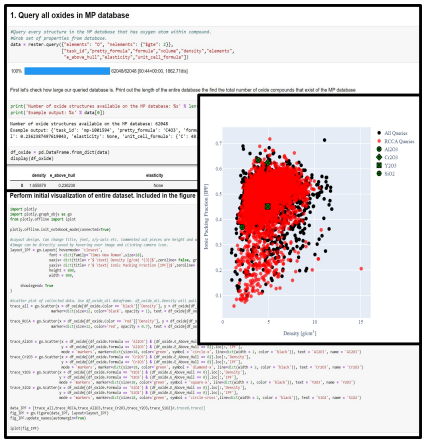

Figure 6.1. Example of live code and plot for tool 'nanohub.org/tools/htoxideprop'

6.2.2 Melting point for high entropy alloys: *meltheas* and *activemeltheas*

This Sim2L [13] calculates the melting point of high entropy alloys through a phase coexistence method [357]. We compute the melting point by heating one half of the system above the melting point, and one half of the system below the melting point, and then perform a constant enthalpy simulation. When coexistence is established the temperatures of the two regions will converge to the defined melting point. The tool outputs a final snapshot of the simulated device, and can display time evolution of potential energy, volume, and system temperature.

To explore the large space of high entropy alloys users can select from five different elements and choose the composition and reproduce results given an initial seed condition for the random configuration. What makes this tool unique is that each input and output of a user simulation is cacheable, and can be used for future viewing as necessary. If an input is selected that exists within the global cache the output will be queried rather than a simulation being repeated.

In *activemeltheas*, we couple active learning with molecular dynamics simulations to identify multiple principal component alloys (MPCAs) with high melting temperatures [358]. We present a fully autonomous workflow for the efficient exploration of the high dimensional compositional space of MPCAs. Visualization of the cached Sim2L results can be viewed using the Jupyter dashboard for Sim2L visualization of *meltheas* [359]. Centralizing these repositories in alignment with other entities such as Materials Project [14], and OQMD [153] allows for more accessible materials data practices.

6.2.3 Machine learning for atomic properties: *mlatomprop*

This tool can be used to train neural networks to predict properties of high entropy alloys [360]. High entropy alloys are metal alloys with 4 or more metals present in significant percentages that are of interest due to their mechanical properties and tailorability to specific applications. However, the range of local atomic environments in HEAs makes it difficult to predict properties of atoms in the alloys. Therefore, we present neural networks to predict properties of the CoCrCuFeNi-family of alloys, including the properties of relaxed vacancy

formation energy, cohesive energy, pressure, and volume. Inputs to the model are bispectrum coefficients and central atom descriptors. We test the models ability to predict on the system it was trained on and see how well it can predict properties of systems with different compositions than the training system.

6.3 nanoHUB Tools and Workshops

6.3.1 Tools for Hands-On Learning

During the initial stages of the Covid-19 pandemic, our group realized that we could fill a unique niche in the scientific community with the development of hands on workshops that incorporated entry level coding exercises, data science, and machine learning in the context of materials engineering [361]. Beginning in the spring of 2020, we created combined Jupyter Notebook tools and live instructional videos for hands-on workshops provided free of charge to the community. Users from around the world were able to join, learn new coding skills on demand, and participate in Q&A sessions afterwards. Live lectures and supplemental staged recordings were included as part of our mission to aid accessibility for new resources and skills. Live lecture audiences ranged from 50-200 individual viewers. These modules were co-organized with a group of peers that contributed each aspect of their knowledge to create a suite of courses for data science expertise from basic Python operation, to automated active learning protocols for materials discovery.

6.3.2 Tools for Classroom Development

From this personal work, the recorded lecture and nanoHUB tool *matdatarepo* [362], [363] have since gained over 1,000 users across the world since its original use in the nanoHUB workshops, and its inclusion in materials engineering courses. The use of these querying mechanisms have been extended to other tools for feature engineering and selection [364].

Leaning on the success of our educational workshops, and the web-based computing platform of nanoHUB, a collaborative effort between the School of Engineering Education, and a Purdue research development initiative for 'Scalable Asymmetric Lifecycle Engagement' (SCALE), we helped to create educational tools for first-year engineering students to learn

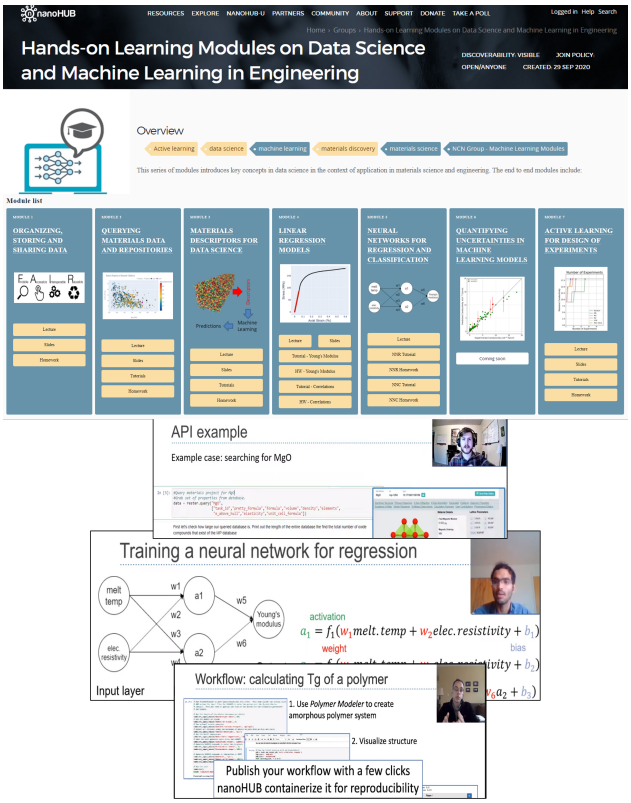

Figure 6.2. Machine learning modules for materials engineering: a multi-part hands-on workshop sponsored by nanoHUB.

Figure 6.3. 'matdatarepo' tool cumulative users. Sourced from https://-nanohub.org/resources/matdatarepo/usage

MATLAB coding within the context of materials engineering for radiation hardening in electronics. Through this effort tools were deployed in introductory engineering courses with fundamental MATLAB coding principles integrated into unique engineering problems related to radiation hardening and electronics. These tools were published through a large collaborative effort and can be found under tool names 'matlabrad' [365] and 'matlabdata' [366].

6.4 Final Remarks

Using online cyberinfrastructure platforms, we disseminate our work for academic and educational purposes to a worldwide audience. For both research and classroom utility, the tools included in this chapter have served as peer-review supplementary resources, educational modules, and live-classroom exercises. The processes included here are ubiquitous across fields, and can be applied for further transparency in work, or for real-world teaching applications. One of the largest challenges in any technical field is the ability to communicate your results to a wide range of audiences effectively, and in a way that does not lose them before they reach your main content. Through modern day visualization techniques, GUI management, and online repositories our goal is to make data and simulations for materials engineering more accessible to new users, and more reliable for returning contributors.

7. CONCLUSIONS

This thesis is a combinatorial toolbox of materials engineering assets for computational efforts and design. By leveraging well-defined physics based models we derive thermodynamic and macroscale quantities of materials through molecular dynamic and density functional theory simulations. Our simulations have allowed for new explorations of material mechanisms, as well as provide necessary databases, but there remain gaps in how to bridge these models together, and how to use them as extrapolatory systems. Enhancing our databases and systems with cyberinfrastructure, we are able to expedite materials screening and molecular dynamics methods through the use of high-performance machine learning models. Using well constructed models with uncertainties, explainable features, and dense datasets we are able to tie together each piece of design through the Pareto front of computational cost. These efforts have been published in online, open repositories for enhancement of educational and academic initiatives.

7.1 Outlook on Experimental and Computational Collaboration for Materials Initiatives

As we showed in Ch. 3, the combination of experimental and computational initiatives is essential for materials design. However, the alignment between these two entities is often non-synchronous. This misalignment comes in many forms including approximation bounds, acceptable noise, and above all data handling and infrastructure. To remedy this, efforts throughout the community have shown that well curated and catered data repositories are viable options for both materials exploration and data storage. The metadata that has slowly begun to accumulate from different aspects of DFT, MD, and experimental repositories will be highly useful for future deep learning modeling and active learning protocols. In particular, the use of cacheable data infrastructures like Sim2Ls [13] for a wide suite of simulations will help expedite materials research with the addition of successful and 'failed' data points. A large hurdle to overcome in the future years will be the representative skew of data in the field, and what we deem publishable. By creating centralized repositories of data with all runs, both successful and failed, we can develop models that learn from non-idealized data

and can construct real implicit dynamics for us. These dynamics can come in the form of molecular modeling, but also scale to the dynamics of experimental design.

7.2 Outlook on Machine Learning for Materials Discovery and Dynamics

A majority of the work in the field of materials design and discovery is still deemed to be within the 'Inverse Problem' of machine learning. While the field has made critical decisions for experimental ventures, and models for computational extrapolation, there remains a lifetime of work to be done. With the acceleration of active learning protocols for both experimental and computational methods the exploratory nature of the field is soon to be tied together in new ways. In particular, leveraging of deep learning algorithms for acceleration of materials search space, and simulation dynamics will aid in the scientific workflow for design.

Bridging the gap between electronic structure theory and molecular simulations allows for new investigative fronts for whatever set of data we can create an initial set for. Shown in Ch. 4, the development of well-converged interatomic potentials driven by neural networks can rely on very small initial databases of DFT, and with enrichment of structures be used to simulate near device scale material. This showcases just one example of deep learning acceleration for bridging density functional theory and interatomic potential design, but other efforts for using autoencoders for dimensionality reduction of chemical pathways [367], scalable algorithms for next generation computation [368], and transfer learning with deep learning for phase field free energy information [369] are exciting examples of what the field is capable of.

In this work, we have helped to shift the boundary of the computational Pareto front for phase change memory exploration, but we have also demonstrated the combination of tools for combining pieces of information from across time and length scales. In particular, the work with GST NNPs shows promise with large scale simulations of device operation, and determination of nucleation kinetics both heterogeneous and homogeneous. Continued work to unravel the mechanisms behind nucleation, growth, and effects of environment on device performance will be interesting avenues to explore with our new NNP tool. A series of additional iterative workflows to follow up on the GST database for doped compounds of

GST+C will also be of interest for future work. These tools are ubiquitous across the field, and can be applied with flexibility as long as the creation of features and datasets are well understood. With continued collaboration across fields the available databases for training and feature engineering will hopefully expand and create new avenues for materials design.

7.3 Final Remarks

Materials engineering is entering into an exciting paradigm that will define an age of materials classes, and with it new ways of interactions between societies. There are going to be many uncertain corners that the research community enters through, but with the tools presented in this thesis, and the modernization of the materials data communities there are high hopes that materials discovery and industrial production processes will become synchronous and reduced in time. As the tools used in our field become more advanced the learning curve will become steeper for each new entry to the field. Part of our responsibility as researchers outside of science is the communication of our results for funding, but also creating accessible pathways for new scholars and minds. Many of us are going to find times during this new age of materials where the unknown is scary, daunting, and intimidating. Tools and methods are going to evolve faster than we can apply them, but hopefully with an integrated community we find ways to use each other's work to its best ability. There are going to be many moments within the growth of our field that growing pains are going to be experienced. As new tools and infrastructures are built the culture around our research will change. More integrated designs of computational and experimental autonomous systems will force scientists and engineers to learn more of each others skills to create the best work they can. Exemplified by the last 10 years of the MGI, the development of computational and experimental infrastructure for materials design has shown high yields for the innovation and curation of unique materials and individuals.

REFERENCES

[1] F. Walsh, P. Boyens, S. Sinclair, and P. Jackson, *The lord of the rings: The two towers*, 2002.

[2] S. P. McPherron, Z. Alemseged, C. W. Marean, *et al.*, "Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at dikika, ethiopia," *Nature*, vol. 466, no. 7308, pp. 857–860, 2010.

[3] Z. Khurshid, M. Zafar, S. Qasim, S. Shahab, M. Naseem, and A. AbuReqaiba, "Advances in nanotechnology for restorative dentistry," *Materials*, vol. 8, no. 2, pp. 717–731, 2015.

[4] F. Ye, Y. Zhao, R. El-Sayed, M. Muhammed, and M. Hassan, "Advances in nanotechnology for cancer biomarkers," *Nano Today*, vol. 18, pp. 103–123, 2018.

[5] S. K. Saxena, R. Nyodu, S. Kumar, and V. K. Maurya, "Current advances in nanotechnology and medicine," in *NanoBioMedicine*, Springer, 2020, pp. 3–16.

[6] V. G. Childe, *The bronze age*. CUP Archive, 1963, vol. 3.

[7] G. Minois, *History of old age: From antiquity to the renaissance*. University of Chicago Press, 1989.

[8] M. Might, *The Illustrated Guide to a Ph.D.* [Online]. Available: https://matt.might.net/articles/phd-school-in-pictures.

[9] N. Science and T. C. (US), *Materials genome initiative for global competitiveness*. Executive Office of the President, National Science and Technology Council, 2011.

[10] N. Science and T. C. (US), *Materials Genome Initiative Strategic Plan*. Executive Office of the President, National Science and Technology Council, 2021.

[11] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[12] A. Strachan, G. Klimeck, and M. Lundstrom, "Cyber-enabled simulations in nanoscale science and engineering," *Computing in Science & Engineering*, vol. 12, no. 2, pp. 12–17, 2010.

[13] M. Hunt, S. Clark, D. Mejia, S. Desai, and A. Strachan, "Sim2ls: Fair simulation workflows and data," *Plos one*, vol. 17, no. 3, e0264492, 2022.

[14] A. Jain, S. P. Ong, G. Hautier, *et al.*, "The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011 002, 2013, ISSN: 2166532X. DOI: 10.1063/1.4812323. [Online]. Available: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1%5C&Agg=doi.

[15] S. P. Ong, S. Cholia, A. Jain, *et al.*, "The materials application programming interface (API): A simple, flexible and efficient API for materials data based on REpresentational state transfer (REST) principles," *Computational Materials Science*, vol. 97, pp. 209–215, 2015. DOI: 10.1016/j.commatsci.2014.10.037. [Online]. Available: http://dx.doi.org/10.1016/j.commatsci.2014.10.037.

[16] J. O'Mara, B. Meredig, and K. Michel, "Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access," *Jom*, vol. 68, no. 8, pp. 2031–2034, 2016.

[17] E. B. Tadmor, R. S. Elliott, S. R. Phillpot, and S. B. Sinnott, "Nsf cyberinfrastructures: A new paradigm for advancing materials simulation," *Current Opinion in Solid State and Materials Science*, vol. 17, no. 6, pp. 298–304, 2013.

[18] M. Planck, "On the theory of the energy distribution law of the normal spectrum," *Verh. Deut. Phys. Ges*, vol. 2, pp. 237–245, 1900.

[19] J. Franck and G. Hertz, "Über zusammenstöße zwischen elektronen und den molekülen des quecksilberdampfes und die ionisierungsspannung desselben," *Physikalische Blätter*, vol. 23, no. 7, pp. 294–301, 1967.

[20] H. Liu, G. H. Low, D. S. Steiger, T. Häner, M. Reiher, and M. Troyer, "Prospects of quantum computing for molecular sciences," *Materials Theory*, vol. 6, no. 1, pp. 1–17, 2022.

[21] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln," *Annalen der physik*, vol. 389, no. 20, pp. 457–484, 1927.

[22] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," *Physical review*, vol. 136, no. 3B, B864, 1964.

[23] A. E. Mattsson, P. A. Schultz, M. P. Desjarlais, T. R. Mattsson, and K. Leung, "Designing meaningful density functional theory calculations in materials science—a primer," *Modelling and Simulation in Materials Science and Engineering*, vol. 13, no. 1, R1, 2004.

[24] R. O. Jones and O. Gunnarsson, "The density functional formalism, its applications and prospects," *Reviews of Modern Physics*, vol. 61, no. 3, p. 689, 1989.

[25] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical review B*, vol. 54, no. 16, p. 11 169, 1996.

[26] J. M. Soler, E. Artacho, J. D. Gale, *et al.*, "The SIESTA method for ab initio order-N materials simulation," *Journal of Physics Condensed Matter*, vol. 14, no. 11, pp. 2745–2779, 2002. DOI: 10.1088/0953-8984/14/11/302. [Online]. Available: http://stacks.iop.org/0953-8984/14/i=11/a=302?key=crossref.8ed2406c09184bcd143191af26e9f492.

[27] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*. Elsevier, 2001, vol. 1.

[28] C. Kittel, "Introduction to solid state physics eighth edition," 2021.

[29] J. E. Lennard-Jones, "Cohesion," *Proceedings of the Physical Society (1926-1948)*, vol. 43, no. 5, p. 461, 1931.

[30] P. M. Morse and E. C. G. Stueckelberg, "Diatomic molecules according to the wave mechanics i: Electronic levels of the hydrogen molecular ion," *Physical Review*, vol. 33, no. 6, p. 932, 1929.

[31] J. Tersoff, "New empirical model for the structural properties of silicon," *Physical review letters*, vol. 56, no. 6, p. 632, 1986.

[32] F. H. Stillinger and T. A. Weber, "Computer simulation of local order in condensed phases of silicon," *Physical review B*, vol. 31, no. 8, p. 5262, 1985.

[33] S. Foiles, M. Baskes, and M. S. Daw, "Embedded-atom-method functions for the fcc metals cu, ag, au, ni, pd, pt, and their alloys," *Physical review B*, vol. 33, no. 12, p. 7983, 1986.

[34] M. S. Daw, S. M. Foiles, and M. I. Baskes, "The embedded-atom method: A review of theory and applications," *Materials Science Reports*, vol. 9, no. 7-8, pp. 251–310, 1993.

[35] M. I. Baskes, "Modified embedded-atom potentials for cubic materials and impurities," *Physical review B*, vol. 46, no. 5, p. 2727, 1992.

[36] A. D. MacKerell Jr, D. Bashford, M. Bellott, *et al.*, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *The journal of physical chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[37] S. L. Mayo, B. D. Olafson, and W. A. Goddard, "Dreiding: A generic force field for molecular simulations," *Journal of Physical chemistry*, vol. 94, no. 26, pp. 8897–8909, 1990.

[38] A. C. Van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, "Reaxff: A reactive force field for hydrocarbons," *The Journal of Physical Chemistry A*, vol. 105, no. 41, pp. 9396–9409, 2001.

[39] LAMMPS, *Http://lammps.sandia.gov/*, 2018. [Online]. Available: http://lammps.sandia.gov/.

[40] S. Plimpton, "Fast Parallel Algorithms for Short – Range Molecular Dynamics," *Journal of Computational Physics*, vol. 117, no. June 1994, pp. 1–19, 1995. DOI: 10.1006/jcph.1995.1039. [Online]. Available: http://lammps.sandia.gov.

[41] S. Nosé, "A molecular dynamics method for simulations in the canonical ensemble," *Molecular physics*, vol. 52, no. 2, pp. 255–268, 1984.

[42] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of chemical physics*, vol. 81, no. 1, pp. 511–519, 1984.

[43] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Physical review A*, vol. 31, no. 3, p. 1695, 1985.

[44] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[45] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. Psychology Press, 2005.

[46] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, 1943.

[47] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[48] T. Dozat, "Incorporating nesterov momentum into adam," 2016.

[49] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization.," *Journal of machine learning research*, vol. 12, no. 7, 2011.

[50] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," *Advances in neural information processing systems*, vol. 27, 2014.

[51] J. Dean, G. Corrado, R. Monga, *et al.*, "Large scale distributed deep networks," *Advances in neural information processing systems*, vol. 25, 2012.

[52] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[53] User:Dhp1080, *Neuron Synapse Image*. "Anatomy, Physiology" by the US National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. [Online]. Available: https://upload.wikimedia.org/wikipedia/commons/b/b5/Neuron.svg.

[54] A. Lenail, *Nn-svg*. [Online]. Available: https://alexlenail.me/NN-SVG/index.html.

[55] Z. D. McClure, S. T. Reeve, and A. Strachan, "Role of electronic thermal transport in amorphous metal recrystallization: A molecular dynamics study," *The Journal of Chemical Physics*, vol. 149, no. 6, p. 064 502, 2018.

[56] C. B. Saltonstall, Z. D. McClure, M. J. Abere, *et al.*, "Complexion dictated thermal resistance with interface density in reactive metal multilayers," *Physical Review B*, vol. 101, no. 24, p. 245 422, 2020.

[57] W. Johnson, C. Kim, and A. Peker, "Golf club made of a bulk-solidifying amorphous metal," no. 7357731, 2008. [Online]. Available: https://www.freepatentsonline.com/7357731.html.

[58] S. Jayalakshmi and M. Gupta, *Metallic amorphous alloy reinforcements in light metal matrices*. Springer, 2015.

[59] C. A. Schuh and A. C. Lund, "Atomistic basis for the plastic yield criterion of metallic glass," *Nature materials*, vol. 2, no. 7, pp. 449–452, 2003.

[60] Y. Wang, X. Xie, H. Li, *et al.*, "Biodegradable camgzn bulk metallic glass for potential skeletal application," *Acta biomaterialia*, vol. 7, no. 8, pp. 3196–3208, 2011.

[61] M. Wuttig and N. Yamada, "Phase-change materials for rewriteable data storage," *Nature materials*, vol. 6, no. 11, pp. 824–832, 2007.

[62] G. Herzer, "Modern soft magnets: Amorphous and nanocrystalline materials," *Acta Materialia*, vol. 61, no. 3, pp. 718–734, 2013.

[63] K. G. Vishnu, M. J. Cherukara, H. Kim, and A. Strachan, "Amorphous ni/al nanoscale laminates as high-energy intermolecular reactive composites," *Physical Review B*, vol. 85, no. 18, p. 184 206, 2012.

[64] K. V. Manukyan, C. E. Shuck, M. J. Cherukara, *et al.*, "Exothermic self-sustained waves with amorphous nickel," *The Journal of Physical Chemistry C*, vol. 120, no. 10, pp. 5827–5838, 2016.

[65] L. Wang, L. Tu, and J. Wen, "Application of phase-change materials in memory taxonomy," *Science and Technology of advanced MaTerialS*, vol. 18, no. 1, pp. 406–429, 2017.

[66] D. Loke, T. Lee, W. Wang, *et al.*, "Breaking the speed limits of phase-change memory," *Science*, vol. 336, no. 6088, pp. 1566–1569, 2012.

[67] J. Giles, "News features-green explosives: Collateral damage," *Nature*, vol. 427, no. 6975, pp. 580–581, 2004.

[68] K. T. Higa, "Energetic nanocomposite lead-free electric primers," *Journal of propulsion and power*, vol. 23, no. 4, pp. 722–727, 2007.

[69] A. Swiston Jr, E. Besnoin, A. Duckham, O. Knio, T. Weihs, and T. Hufnagel, "Thermal and microstructural effects of welding metallic glasses by self-propagating reactions in multilayer foils," *Acta materialia*, vol. 53, no. 13, pp. 3713–3719, 2005.

[70] J. Wang, E. Besnoin, A. Duckham, *et al.*, "Room-temperature soldering with nanostructured foils," *Applied physics letters*, vol. 83, no. 19, pp. 3987–3989, 2003.

[71] J. Schroers, "Processing of bulk metallic glass," *Advanced materials*, vol. 22, no. 14, pp. 1566–1597, 2010.

[72] A. Peker and W. L. Johnson, "A highly processable metallic glass: Zr41. 2ti13. 8cu12. 5ni10. 0be22. 5," *Applied Physics Letters*, vol. 63, no. 17, pp. 2342–2344, 1993.

[73] Y. Pei, G. Zhou, N. Luan, B. Zong, M. Qiao, and F. F. Tao, "Synthesis and catalysis of chemically reduced metal–metalloid amorphous alloys," *Chemical Society Reviews*, vol. 41, no. 24, pp. 8140–8162, 2012.

[74] T. Umegaki, J.-M. Yan, X.-B. Zhang, H. Shioyama, N. Kuriyama, and Q. Xu, "Co–sio2 nanosphere-catalyzed hydrolytic dehydrogenation of ammonia borane for chemical hydrogen storage," *Journal of Power Sources*, vol. 195, no. 24, pp. 8209–8214, 2010.

[75] K. Lu, "Nanocrystalline metals crystallized from amorphous solids: Nanocrystallization, structure, and properties," *Materials Science and Engineering: R: Reports*, vol. 16, no. 4, pp. 161–221, 1996.

[76] H.-D. Geiler, E. Glaser, G. Götz, and M. Wagner, "Explosive crystallization in silicon," *Journal of applied physics*, vol. 59, no. 9, pp. 3091–3099, 1986.

[77] M. O. Thompson, G. Galvin, J. Mayer, *et al.*, "Melting temperature and explosive crystallization of amorphous silicon during pulsed laser irradiation," *Physical review letters*, vol. 52, no. 26, p. 2360, 1984.

[78] O. Bostanjoglo and R. Liedtke, "Tracing fast phase transitions by electron microscopy," *physica status solidi (a)*, vol. 60, no. 2, pp. 451–455, 1980.

[79] C. Krzeminski, Q. Brulin, V. Cuny, E. Lecat, E. Lampin, and F. Cleri, "Molecular dynamics simulation of the recrystallization of amorphous si layers: Comprehensive study of the dependence of the recrystallization velocity on the interatomic potential," *Journal of applied physics*, vol. 101, no. 12, p. 123 506, 2007.

[80] L. A. Marqués, M.-J. Caturla, T. Díaz de la Rubia, and G. H. Gilmer, "Ion beam induced recrystallization of amorphous silicon: A molecular dynamics study," *Journal of applied physics*, vol. 80, no. 11, pp. 6160–6169, 1996.

[81] B. C. Gundrum, D. G. Cahill, and R. S. Averback, "Thermal conductance of metal-metal interfaces," *Physical Review B*, vol. 72, no. 24, p. 245 426, 2005.

[82] D. Duffy and A. Rutherford, "Including the effects of electronic stopping and electron–ion interactions in radiation damage simulations," *Journal of Physics: Condensed Matter*, vol. 19, no. 1, p. 016 207, 2006.

[83] A. Rutherford and D. Duffy, "The effect of electron–ion interactions on radiation damage simulations," *Journal of Physics: Condensed Matter*, vol. 19, no. 49, p. 496 201, 2007.

[84] K.-H. Lin, B. L. Holian, T. C. Germann, and A. Strachan, "Mesodynamics with implicit degrees of freedom," *The Journal of Chemical Physics*, vol. 141, no. 6, p. 064 107, 2014.

[85] A. Stukowski, "Visualization and analysis of atomistic simulation data with OVITO-the Open Visualization Tool," *Modelling and Simulation in Materials Science and Engineering*, vol. 18, no. 1, p. 015 012, 2010, ISSN: 09650393. DOI: 10.1088/0965-0393/18/1/015012.

[86] P. M. Larsen, S. Schmidt, and J. Schiøtz, "Robust structural identification via polyhedral template matching," *Modelling and Simulation in Materials Science and Engineering*, vol. 24, no. 5, p. 055 007, 2016.

[87] Y. Mishin, "Atomistic modeling of the γ and γ -phases of the ni–al system," *Acta Materialia*, vol. 52, no. 6, pp. 1451–1467, 2004.

[88] J. R. Morris, C. Wang, K. Ho, and C. Chan, "Melting line of aluminum from simulations of coexisting phases," *Physical Review B*, vol. 49, no. 5, p. 3109, 1994.

[89] P. Desai, "Thermodynamic properties of nickel," *International journal of thermophysics*, vol. 8, no. 6, pp. 763–780, 1987.

[90] A. Caro and M. Victoria, "Ion-electron interaction in molecular-dynamics cascades," *Physical Review A*, vol. 40, no. 5, p. 2287, 1989.

[91] I. Koponen, "Energy transfer between electrons and ions in dense displacement cascades," *Physical Review B*, vol. 47, no. 21, p. 14 011, 1993.

[92] G. Norman, S. Starikov, V. Stegailov, I. Saitov, and P. Zhilyaev, "Atomistic modeling of warm dense matter in the two-temperature state," *Contributions to Plasma Physics*, vol. 53, no. 2, pp. 129–139, 2013.

[93] V. Pisarev and S. Starikov, "Atomistic simulation of ion track formation in uo2," *Journal of Physics: Condensed Matter*, vol. 26, no. 47, p. 475 401, 2014.

[94] Z. Lin, L. V. Zhigilei, and V. Celli, "Electron-phonon coupling and electron heat capacity of metals under conditions of strong electron-phonon nonequilibrium," *Physical Review B*, vol. 77, no. 7, p. 075 133, 2008.

[95] W. Ma, H. Wang, X. Zhang, and W. Wang, "Study of the electron–phonon relaxation in thin metal films using transient thermoreflectance technique," *International Journal of Thermophysics*, vol. 34, no. 12, pp. 2400–2415, 2013.

[96] M. Sandhofer, I. Y. Sklyadneva, R. Heid, *et al.*, "Electron-phonon coupling in quantum-well states of the pb/si (111) system," 2014.

[97] Z. Lin and L. V. Zhigilei, "Temperature dependences of the electron–phonon coupling, electron heat capacity and thermal conductivity in ni under femtosecond laser irradiation," *Applied Surface Science*, vol. 253, no. 15, pp. 6295–6300, 2007.

[98] Z. Wang, C. Dufour, E. Paumier, and M. Toulemonde, "The se sensitivity of metals under swift-heavy-ion irradiation: A transient thermal process," *Journal of Physics: Condensed Matter*, vol. 6, no. 34, p. 6733, 1994.

[99] R. Powell, C. Y. Ho, and P. E. Liley, *Thermal conductivity of selected materials*. US Department of Commerce, National Bureau of Standards Washington, DC, 1966, vol. 8.

[100] S. Yuan and P. Jiang, "Thermal conductivity of small nickel particles," *International journal of thermophysics*, vol. 27, no. 2, pp. 581–595, 2006.

[101] J. Narayan, G. Rozgonyi, D. Bensahel, G. Auvert, V. Nguyen, and A. Rai, "Explosive recrystallization of ion implantation amorphous silicon layers," *MRS Online Proceedings Library*, vol. 13, no. 1, pp. 177–184, 1982.

[102] C. Chiritescu, D. G. Cahill, N. Nguyen, *et al.*, "Ultralow thermal conductivity in disordered, layered wse2 crystals," *Science*, vol. 315, no. 5810, pp. 351–353, 2007.

[103] R. Costescu, D. Cahill, F. Fabreguette, Z. Sechrist, and S. George, "Ultra-low thermal conductivity in w/al2o3 nanolaminates," *Science*, vol. 303, no. 5660, pp. 989–990, 2004.

[104] B. Poudel, Q. Hao, Y. Ma, *et al.*, "High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys," *Science*, vol. 320, no. 5876, pp. 634–638, 2008.

[105] T. Beechem, S. Graham, P. Hopkins, and P. Norris, "Role of interface disorder on thermal boundary conductance using a virtual crystal approach," *Applied Physics Letters*, vol. 90, no. 5, p. 054 104, 2007.

[106] P. E. Hopkins, P. M. Norris, R. J. Stevens, T. E. Beechem, and S. Graham, "Influence of interfacial mixing on thermal boundary conductance across a chromium/silicon interface," *Journal of Heat Transfer*, vol. 130, no. 6, 2008.

[107] T. Beechem and P. E. Hopkins, "Predictions of thermal boundary conductance for systems of disordered solids and interfaces," *Journal of Applied Physics*, vol. 106, no. 12, p. 124 301, 2009.

[108] P. E. Hopkins, L. M. Phinney, J. R. Serrano, and T. E. Beechem, "Effects of surface roughness and oxide layer on the thermal boundary conductance at aluminum/silicon interfaces," in *International Heat Transfer Conference*, vol. 49415, 2010, pp. 313–319.

[109] J. C. Duda, T. S. English, E. S. Piekos, T. E. Beechem, T. W. Kenny, and P. E. Hopkins, "Bidirectionally tuning kapitza conductance through the inclusion of substitutional impurities," *Journal of Applied Physics*, vol. 112, no. 7, p. 073 519, 2012.

[110] C. S. Gorham, K. Hattar, R. Cheaito, *et al.*, "Ion irradiation of the native oxide/silicon surface increases the thermal boundary conductance across aluminum/silicon interfaces," *Physical Review B*, vol. 90, no. 2, p. 024 301, 2014.

[111] M. D. Losego, M. E. Grady, N. R. Sottos, D. G. Cahill, and P. V. Braun, "Effects of chemical bonding on heat transport across interfaces," *Nature materials*, vol. 11, no. 6, pp. 502–506, 2012.

[112] P. E. Hopkins, "Thermal transport across solid interfaces with nanoscale imperfections: Effects of roughness, disorder, dislocations, and bonding on thermal boundary conductance," *International Scholarly Research Notices*, vol. 2013, 2013.

[113] R. J. Stevens, L. V. Zhigilei, and P. M. Norris, "Effects of temperature and disorder on thermal boundary conductance at solid–solid interfaces: Nonequilibrium molecular dynamics simulations," *International Journal of Heat and Mass Transfer*, vol. 50, no. 19-20, pp. 3977–3989, 2007.

[114] J. D. Schuler and T. J. Rupert, "Materials selection rules for amorphous complexion formation in binary metallic alloys," *Acta Materialia*, vol. 140, pp. 196–205, 2017.

[115] S. J. Dillon, M. Tang, W. C. Carter, and M. P. Harmer, "Complexion: A new concept for kinetic engineering in materials science," *Acta Materialia*, vol. 55, no. 18, pp. 6208–6218, 2007.

[116] A. R. Krause, P. R. Cantwell, C. J. Marvel, C. Compson, J. M. Rickman, and M. P. Harmer, "Review of grain boundary complexion engineering: Know your boundaries," *Journal of the American Ceramic Society*, vol. 102, no. 2, pp. 778–800, 2019.

[117] D. P. Adams, R. V. Reeves, M. Abere, *et al.*, "Ignition and self-propagating reactions in al/pt multilayers of varied design," *Journal of Applied Physics*, vol. 124, no. 9, p. 095 105, 2018.

[118] R. D. Deslattes, E. G. Kessler Jr, P. Indelicato, L. De Billy, E. Lindroth, and J. Anton, "X-ray transition energies: New approach to a comprehensive evaluation," *Reviews of Modern Physics*, vol. 75, no. 1, p. 35, 2003.

[119] G. Cliff and G. W. Lorimer, "The quantitative analysis of thin specimens," *Journal of Microscopy*, vol. 103, no. 2, pp. 203–207, 1975.

[120] S. Kabekkodu, "Icdd 2016 powder diffraction file inorganic and organic data book," *International Centre for Diffraction Data, PA, USA*, 2016.

[121] C. D. Landon, R. H. Wilke, M. T. Brumbach, *et al.*, "Thermal transport in tantalum oxide films for memristive applications," *Applied Physics Letters*, vol. 107, no. 2, p. 023 108, 2015.

[122] T. E. Beechem, A. E. McDonald, E. J. Fuller, *et al.*, "Size dictated thermal conductivity of gan," *Journal of Applied Physics*, vol. 120, no. 9, p. 095 104, 2016.

[123] R. Cheaito, K. Hattar, J. T. Gaskins, *et al.*, "Thermal flux limited electron Kapitza conductance in copper-niobium multilayers," *Applied Physics Letters*, vol. 106, no. 9, pp. 3–8, 2015. DOI: 10.1063/1.4913420.

[124] A. J. Schmidt, "Optical characterization of thermal transport from the nanoscale to the macroscale," Ph.D. dissertation, Massachusetts Institute of Technology, 2008.

[125] B. C. Daly, H. J. Maris, K. Imamura, and S. Tamura, "Molecular dynamics calculation of the thermal conductivity of superlattices," *Physical review B*, vol. 66, no. 2, p. 024 301, 2002.

[126] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, 1996, ISSN: 10797114. DOI: 10.1103/PhysRevLett.77.3865. arXiv: 0927-0256(96)00008 [10.1016].

[127] T. Markussen, M. Palsgaard, D. Stradi, T. Gunst, M. Brandbyge, and K. Stokbro, "Electron-phonon scattering from Green's function transport combined with molecular dynamics: Applications to mobility predictions," *Physical Review B*, vol. 95, no. 24, 2017, ISSN: 24699969. DOI: 10.1103/PhysRevB.95.245210. arXiv: 1701.02883.

[128] K. W. Jacobsen, P. Stoltze, and J. K. Nørskov, "A semi-empirical effective medium theory for metals and alloys," *Surface Science*, vol. 366, pp. 394–402, 1996, ISSN: 0039-6028. DOI: http://dx.doi.org/10.1016/0039-6028(96)00816-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0039602896008163.

[129] E. B. Tadmor, R. S. Elliott, S. R. Phillpot, and S. B. Sinnott, "NSF cyberinfrastructures: A new paradigm for advancing materials simulation," *Curr. Opin. Solid State Mater. Sci.*, vol. 17, pp. 298–304, 2013, ISSN: 13590286. DOI: 10.1016/j.cossms.2013.10.004.

[130] OpenKIM, *Open Knowledgebase of Interatomic Models https://openkim.org/*, 2018. [Online]. Available: https://openkim.org/.

[131] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, "Density-functional method for nonequilibrium electron transport," *Physical Review B*, vol. 65, no. 16, p. 165 401, Mar. 2002.

[132] P. Ordejón, E. Artacho, and J. M. Soler, "Self-consistent order- Ndensity-functional calculations for very large systems," *Physical Review B*, vol. 53, no. 16, R10441–R10444, Apr. 1996.

[133] J. M. Soler, E. Artacho, J. D. Gale, *et al.*, "The SIESTA method for ab initio order-N materials simulation," *Journal of Physics: Condensed Matter*, vol. 14, no. 11, pp. 2745–2779, Mar. 2002.

[134] SIESTA, *Https://departments.icmab.es/leem/siesta/*, 2018. [Online]. Available: https://departments.icmab.es/leem/siesta/.

[135] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, Oct. 1996.

[136] J. Ravichandran, A. K. Yadav, R. Cheaito, *et al.*, "Crossover from incoherent to coherent phonon scattering in epitaxial oxide superlattices," *Nature materials*, vol. 13, no. 2, pp. 168–172, 2014.

[137] Z. D. McClure and A. Strachan, "Expanding materials selection via transfer learning for high-temperature oxide selection," *JOM*, vol. 73, no. 1, pp. 103–115, 2021.

[138] D. E. Farache, J. C. Verduzco, Z. D. McClure, S. Desai, and A. Strachan, "Active learning and molecular dynamics simulations to find high melting temperature alloys," *arXiv preprint arXiv:2110.08136*, 2021.

[139] T. M. Pollock and S. Tin, "Nickel-based superalloys for advanced turbine engines: Chemistry, microstructure and properties," *Journal of propulsion and power*, vol. 22, no. 2, pp. 361–374, 2006.

[140] G. Electric, "General Electric Annual Report," Tech. Rep., 2018.

[141] F. Falempin, E. Daniau, N. Getin, F. Bykovskii, and S. Zhdan, "Toward a continuous detonation wave rocket engine demonstrator," in *14th AIAA/AHI space planes and hypersonic systems and technologies conference*, 2006, p. 7956.

[142] J. Doychak and M. Hebsur, "Protective al 2 o 3 scale formation on nbal 3-base alloys," *Oxidation of Metals*, vol. 36, no. 1-2, pp. 113–141, 1991.

[143] J. Justin and A. Jankowiak, "Ultra high temperature ceramics: Densification, properties and thermal stability," *Journal Aerospace Lab*, vol. 3, pp. 1–11, 2011.

[144] J. L. Smialek and N. S. Jacobson, "Oxidation of high-temperature aerospace materials," *High temperature materials and mechanisms*, pp. 95–162, 2014.

[145] B. Cantor, I. Chang, P. Knight, and A. Vincent, "Microstructural development in equiatomic multicomponent alloys," *Materials Science and Engineering: A*, vol. 375, pp. 213–218, 2004.

[146] J.-W. Yeh, S.-K. Chen, S.-J. Lin, *et al.*, "Nanostructured high-entropy alloys with multiple principal elements: Novel alloy design concepts and outcomes," *Advancd Engineering Materials*, vol. 6, no. 5, pp. 299–303, 2004.

[147] S. Gorsse, J.-P. Couzinié, and D. B. Miracle, "From high-entropy alloys to complex concentrated alloys," *Comptes Rendus Physique*, vol. 19, no. 8, pp. 721–736, 2018.

[148] O. Senkov, G. Wilks, D. Miracle, C. Chuang, and P. Liaw, "Refractory high-entropy alloys," *Intermetallics*, vol. 18, no. 9, pp. 1758–1765, 2010.

[149] C.-H. Chang, M. S. Titus, and J.-W. Yeh, "Oxidation behavior between 700 and 1300 c of refractory tizrnbhfta high-entropy alloys containing aluminum," *Advanced Engineering Materials*, vol. 20, no. 6, p. 1 700 948, 2018.

[150] G. Smith and J. Fischer, "High temperature corrosion resistance of mechanically alloyed products in gas turbine environments," in *ASME 1990 International Gas Turbine and Aeroengine Congress and Exposition. Brussels, Bel-gium*, Citeseer, 1990, pp. 1–7.

[151] T. Butler, K. Chaput, J. Dietrich, and O. Senkov, "High temperature oxidation behaviors of equimolar nbtizrv and nbtizrcr refractory complex concentrated alloys (rccas)," *Journal of Alloys and Compounds*, vol. 729, pp. 1004–1019, 2017.

[152] R. Bedworth and N. Pilling, "The oxidation of metals at high temperatures," *J Inst Met*, vol. 29, no. 3, pp. 529–582, 1923.

[153] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *Jom*, vol. 65, no. 11, pp. 1501–1509, 2013.

[154] Citrine Informatics, *Citrination Database, citrination.com*, 2020. [Online]. Available: https://citrination.com/.

[155] G. Colaboratory, "Google colaboratory," *Google*, 2020. [Online]. Available: https://research.google.com/colaboratory/faq.html.

[156] S. T. Reeve, D. M. Guzman, L. Alzate-Vargas, B. Haley, P. Liao, and A. Strachan, "Online simulation powered learning modules for materials science," *MRS Advances*, pp. 1–16, 2019.

[157] M. F. Ashby and D. Cebon, "Materials selection in mechanical design," *Le Journal de Physique IV*, vol. 3, no. C7, pp. C7–1, 1993.

[158] M. Ashby, "Multi-objective optimization in material design and selection," *Acta materialia*, vol. 48, no. 1, pp. 359–369, 2000.

[159] M. Ashby, "Criteria for selecting the components of composites," *Acta metallurgica et materialia*, vol. 41, no. 5, pp. 1313–1335, 1993.

[160] V. Cutello, G. Narzisi, and G. Nicosia, "A class of pareto archived evolution strategy algorithms using immune inspired operators for ab-initio protein structure prediction," in *Workshops on Applications of Evolutionary Computation*, Springer, 2005, pp. 54–63.

[161] E. Van Der Giessen, P. A. Schultz, N. Bertin, *et al.*, "Roadmap on multiscale materials modeling," *Modelling and Simulation in Materials Science and Engineering*, vol. 28, no. 4, p. 043 001, 2020.

[162] E. D. Cubuk, A. D. Sendek, and E. J. Reed, "Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data," *The Journal of chemical physics*, vol. 150, no. 21, p. 214 701, 2019.

[163] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka, "Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization," *Physical review letters*, vol. 115, no. 20, p. 205 901, 2015.

[164] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Physical review letters*, vol. 114, no. 10, p. 105 503, 2015.

[165] W. M. Brown, S. Martin, M. D. Rintoul, and J.-L. Faulon, "Designing novel polymers with targeted properties using the signature molecular descriptor," *Journal of chemical information and modeling*, vol. 46, no. 2, pp. 826–835, 2006.

[166] C. J. Churchwell, M. D. Rintoul, S. Martin, *et al.*, "The signature molecular descriptor: 3. inverse-quantitative structure–activity relationship of icam-1 inhibitory peptides," *Journal of Molecular Graphics and Modelling*, vol. 22, no. 4, pp. 263–273, 2004.

[167] F. Lindemann, "The calculation of molecular vibration frequency," *Z. phys*, vol. 11, pp. 609–612, 1910.

[168] J. 1. Poirier, "Lindemann law and the melting temperature of perovskites," *Physics of the earth and planetary interiors*, vol. 54, no. 3-4, pp. 364–369, 1989.

[169] G. H. Wolf and R. Jeanloz, "Lindemann melting law: Anharmonic correction and test of its validity for minerals," *Journal of Geophysical Research: Solid Earth*, vol. 89, no. B9, pp. 7821–7835, 1984.

[170] W. R. Inc, "Mathematica, version 12.0," Champaign, IL, 2019. [Online]. Available: https://www.wolfram.com/wolfram-alpha-notebook-edition.

[171] A. Jain, G. Hautier, C. J. Moore, *et al.*, "A high-throughput infrastructure for density functional theory calculations," *Computational Materials Science*, vol. 50, no. 8, pp. 2295–2310, 2011.

[172] A. M. Deml, A. M. Holder, R. P. O'Hayre, C. B. Musgrave, and V. Stevanović, "Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides," *The journal of physical chemistry letters*, vol. 6, no. 10, pp. 1948–1953, 2015.

[173] J. T. Schick, A. M. Gopakumar, and A. M. Rappe, "Descriptors for thermal expansion in solids," *arXiv preprint arXiv:1701.03966*, 2017.

[174] C. W. Coley, W. Jin, L. Rogers, *et al.*, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical science*, vol. 10, no. 2, pp. 370–377, 2019.

[175] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, and P. W. Chung, "Applying machine learning techniques to predict the properties of energetic materials," *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.

[176] J. Ling, M. Hutchinson, E. Antono, S. Paradiso, and B. Meredig, "High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates," *Integrating Materials and Manufacturing Innovation*, vol. 6, no. 3, pp. 207–217, 2017.

[177] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[178] M. Blackman, "The specific heat of solids," in *HDP*, vol. 3, 1955, pp. 325–382.

[179] L. Ward, A. Dunn, A. Faghaninia, *et al.*, "Matminer: An open source toolkit for materials data mining," *Computational Materials Science*, vol. 152, pp. 60–69, 2018.

[180] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Computational Materials*, vol. 2, p. 16 028, 2016.

[181] B. Meredig, A. Agrawal, S. Kirklin, *et al.*, "Combinatorial screening for new materials in unconstrained composition space with machine learning," *Physical Review B*, vol. 89, no. 9, p. 094 104, 2014.

[182] W. D. Callister and D. G. Rethwisch, *Materials science and engineering: an introduction*. Wiley New York, 2018.

[183] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[184] B. Efron, *Model selection estimation and bootstrap smoothing. division of biostatistics*, 2012.

[185] S. Wager, T. Hastie, and B. Efron, "Confidence intervals for random forests: The jackknife and the infinitesimal jackknife," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1625–1651, 2014.

[186] R. K. Tripathy and I. Bilionis, "Deep uq: Learning deep neural network surrogate models for high dimensional uncertainty quantification," *Journal of computational physics*, vol. 375, pp. 565–588, 2018.

[187] M. Hutchinson, *Python package: Lolo*, URL: https://github.com/CitrineInformatics/lolo., 2016.

[188] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" In *International workshop on machine learning and data mining in pattern recognition*, Springer, 2012, pp. 154–168.

[189] J. Wang, Y. Zhou, X. Chong, R. Zhou, and J. Feng, "Microstructure and thermal properties of a promising thermal barrier coating: Ytao4," *Ceramics International*, vol. 42, no. 12, pp. 13 876–13 881, 2016.

[190] A. K. Bhattacharya, V. Shklover, W. Steurer, G. Witz, H.-p. Bossmann, and O. Fabrichnaya, "Ta2o5–y2o3–zro2 system: Experimental study and preliminary thermodynamic description," *Journal of the European Ceramic Society*, vol. 31, no. 3, pp. 249–257, 2011.

[191] J. L. Caslavsky and D. J. Viechnicki, "Melting behaviour and metastability of yttrium aluminium garnet (yag) and yalo 3 determined by optical differential thermal analysis," *Journal of materials science*, vol. 15, no. 7, pp. 1709–1718, 1980.

[192] S. Gu, S. Zhang, F. Liu, and W. Li, "New anti-ablation candidate for carbon/carbon composites: Preparation, composition and ablation behavior of y2hf2o7 coating under an oxyacetylene torch," *Journal of the European Ceramic Society*, vol. 38, no. 15, pp. 5082–5091, 2018.

[193] K. Kuribayashi, M. Yoshimura, T. Ohta, and T. SATA, "High-temperature phase relations in the system y2o3-y2o3 · wo3," *Journal of the American Ceramic Society*, vol. 63, no. 11-12, pp. 644–647, 1980.

[194] R. C. Reed, *The superalloys: fundamentals and applications*. Cambridge university press, 2008.

[195] S. Yamanaka, K. Kurosaki, T. Oyama, *et al.*, "Thermophysical properties of perovskite-type strontium cerate and zirconate," *Journal of the American Ceramic Society*, vol. 88, no. 6, pp. 1496–1499, 2005.

[196] Z. D. McClure and A. Strachan, *High temperature oxide property explorer*, 2020. DOI: doi:10.21981/6CXG-2Z62. [Online]. Available: https://nanohub.org/resources/htoxideprop.

[197] N. S. Eyke, W. H. Green, and K. F. Jensen, "Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening," *Reaction Chemistry & Engineering*, vol. 5, no. 10, pp. 1963–1972, 2020.

[198] L. Zhao, Z. Li, B. Caswell, J. Ouyang, and G. E. Karniadakis, "Active learning of constitutive relation from mesoscopic dynamics for macroscopic modeling of non-newtonian flows," *Journal of Computational Physics*, vol. 363, pp. 116–127, 2018.

[199] A. G. Kusne, H. Yu, C. Wu, *et al.*, "On-the-fly closed-loop materials discovery via bayesian active learning," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[200] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, and T. Lookman, "Accelerated search for materials with targeted properties by adaptive design," *Nature communications*, vol. 7, no. 1, pp. 1–9, 2016.

[201] K. Tran and Z. W. Ulissi, "Active learning across intermetallics to guide discovery of electrocatalysts for co2 reduction and h2 evolution," *Nature Catalysis*, vol. 1, no. 9, pp. 696–703, 2018.

[202] A. Jain, S. P. Ong, G. Hautier, *et al.*, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," *APL materials*, vol. 1, no. 1, p. 011 002, 2013.

[203] S. Desai, S. Clark, and A. Strachan, *Introduction to simtools*, 2020. [Online]. Available: https://nanohub.org/resources/introtosimtools,%20DOI:%2010.21981/4Z7T-X415.

[204] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, and T. Lookman, "Adaptive strategies for materials design using uncertainties," *Scientific reports*, vol. 6, no. 1, pp. 1–9, 2016.

[205] C. Kim, A. Chandrasekaran, A. Jha, and R. Ramprasad, "Active-learning and materials design: The example of high glass transition temperature polymers," *MRS Communications*, vol. 9, no. 3, pp. 860–866, 2019.

[206] J. C. Verduzco, E. E. Marinero, and A. Strachan, "An active learning approach for the design of doped LLZO ceramic garnets for battery applications," *Integrating Materials and Manufacturing Innovation*, pp. 1–12, 2021.

[207] J. Wise, A. G. de Barron, A. Splendiani, *et al.*, "Implementation and relevance of fair data principles in biopharmaceutical r&d," *Drug discovery today*, vol. 24, no. 4, pp. 933–938, 2019.

[208] B. Settles, *Synthesis Lectures on Artificial Intelligence and Machine Learning: Active Learning*. San Rafael, United States: Morgan & Claypool Publishers, 2011, ISBN: 978-1-60845-726-7. [Online]. Available: http://ebookcentral.proquest.com/lib/purdue/detail.action?docID=956875.

[209] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," *arXiv preprint arXiv:1707.05928*, 2017.

[210] P. Yoo, M. Sakano, S. Desai, M. M. Islam, P. Liao, and A. Strachan, "Neural network reactive force field for C, H, N, and O systems," *npj Computational Materials*, vol. 7, no. 1, pp. 1–10, 2021.

[211] A. G. Kusne, T. Gao, A. Mehta, *et al.*, "On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets," *Scientific reports*, vol. 4, no. 1, pp. 1–7, 2014.

[212] P. Nikolaev, D. Hooper, F. Webber, *et al.*, "Autonomy in materials research: A case study in carbon nanotube growth," *npj Computational Materials*, vol. 2, no. 1, pp. 1–6, 2016.

[213] F. Ren, L. Ward, T. Williams, *et al.*, "Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments," *Science advances*, vol. 4, no. 4, eaaq1566, 2018.

[214] M. M. Noack, K. G. Yager, M. Fukuto, G. S. Doerk, R. Li, and J. A. Sethian, "A kriging-based approach to autonomous experimentation with applications to x-ray scattering," *Scientific reports*, vol. 9, no. 1, pp. 1–19, 2019.

[215] S. Kiyohara, H. Oda, K. Tsuda, and T. Mizoguchi, "Acceleration of stable interface structure searching using a kriging approach," *Japanese Journal of Applied Physics*, vol. 55, no. 4, p. 045 502, 2016.

[216] A. M. Cuitiño, L. Stainier, G. Wang, *et al.*, "A multiscale approach for modeling crystalline solids," *Journal of computer-aided materials design*, vol. 8, no. 2, pp. 127–149, 2001.

[217] A. Strachan, T. Çağin, and W. A. Goddard III, "Phase diagram of mgo from density-functional theory and molecular-dynamics simulations," *Physical Review B*, vol. 60, no. 22, p. 15 084, 1999.

[218] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," *Physical review letters*, vol. 104, no. 13, p. 136 403, 2010.

[219] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," *Journal of Computational Physics*, vol. 285, pp. 316–330, 2015.

[220] J. Behler and M. Parrinello, "Generalized neural network representation of high dimensional potential energy surfaces," *Physical review letters*, vol. 98, no. 14, p. 146 401, 2007.

[221] T. M. Dieb, S. Ju, K. Yoshizoe, Z. Hou, J. Shiomi, and K. Tsuda, "Mdts: Automatic complex materials design using monte carlo tree search," *Science and technology of advanced materials*, vol. 18, no. 1, pp. 498–503, 2017.

[222] T. M. Dieb, S. Ju, J. Shiomi, and K. Tsuda, "Monte carlo tree search for materials design and discovery," *MRS Communications*, vol. 9, no. 2, pp. 532–536, 2019.

[223] T. K. Patra, T. D. Loeffler, and S. K. Sankaranarayanan, "Accelerating copolymer inverse design using monte carlo tree search," *Nanoscale*, vol. 12, no. 46, pp. 23 653–23 662, 2020.

[224] T. D. Loeffler, S. Banik, T. K. Patra, M. Sternberg, and S. K. Sankaranarayanan, "Reinforcement learning in discrete action space applied to inverse defect design," *Journal of Physics Communications*, vol. 5, no. 3, p. 031 001, 2021.

[225] Y. Lu, Y. Dong, S. Guo, *et al.*, "A promising new class of high-temperature alloys: Eutectic high-entropy alloys," *Scientific reports*, vol. 4, no. 1, pp. 1–5, 2014.

[226] Z. Xu, Z. Ma, M. Wang, Y. Chen, Y. Tan, and X. Cheng, "Design of novel low-density refractory high entropy alloys for high-temperature applications," *Materials Science and Engineering: A*, vol. 755, pp. 318–322, 2019.

[227] S. Praveen and H. S. Kim, "High entropy alloys: Potential candidates for high temperature applications: An overview," *Advanced Engineering Materials*, vol. 20, no. 1, p. 1 700 645, 2018.

[228] O. Senkov, G. Wilks, D. Miracle, C. Chuang, and P. Liaw, "Refractory high-entropy alloys," *Intermetallics*, vol. 18, no. 9, pp. 1758–1765, 2010, ISSN: 0966-9795.

[229] W. Wang, L. Hu, S. Luo, L. Meng, D. Geng, and B. Wei, "Liquid phase separation and rapid dendritic growth of high-entropy CoCrCuFeNi alloy," *Intermetallics*, vol. 77, pp. 41–45, 2016.

[230] F. He, Z. Wang, Q. Wu, J. Li, J. Wang, and C. Liu, "Phase separation of metastable CoCrFeNi high entropy alloy at intermediate temperatures," *Scripta Materialia*, vol. 126, pp. 15–19, 2017.

[231] Q.-J. Hong, J. Schroers, D. Hofmann, S. Curtarolo, M. Asta, and A. van de Walle, "Theoretical prediction of high melting temperature for a Mo–Ru–Ta–W HCP multiprincipal element alloy," *npj Computational Materials*, vol. 7, no. 1, pp. 1–4, 2021.

[232] J. R. Morris and X. Song, "The melting lines of model systems calculated from coexistence simulations," *The Journal of chemical physics*, vol. 116, no. 21, pp. 9352–9358, 2002.

[233] D. Farkas and A. Caro, "Model interatomic potentials and lattice strain in a high-entropy alloy," *Journal of Materials Research*, vol. 33, no. 19, pp. 3218–3225, 2018.

[234] S.-N. Luo, T. J. Ahrens, T. Çağın, A. Strachan, W. A. Goddard III, and D. C. Swift, "Maximum superheating and undercooling: Systematics, molecular dynamics simulations, and dynamic experiments," *Physical Review B*, vol. 68, no. 13, p. 134 206, 2003.

[235] D. Farkas and A. Caro, "Model interatomic potentials for fe–ni–cr–co–al high-entropy alloys," *Journal of Materials Research*, vol. 35, no. 22, pp. 3031–3040, 2020.

[236] A. S. Lemak and N. K. Balabaev, "On The Berendsen Thermostat," *Molecular Simulation*, vol. 13, no. 3, pp. 177–187, 1994.

[237] P. J. Daivis and D. J. Evans, "Comparison of constant pressure and constant volume nonequilibrium simulations of sheared model decane," *J. Chem. Phys*, vol. 100, 1994.

[238] A. Stukowski, "Visualization and analysis of atomistic simulation data with ovito–the open visualization tool," *Modelling and simulation in materials science and engineering*, vol. 18, no. 1, p. 015 012, 2009.

[239] X.-G. Li, C. Chen, H. Zheng, Y. Zuo, and S. P. Ong, "Complex strengthening mechanisms in the nbmotaw multi-principal element alloy," *npj Computational Materials*, vol. 6, no. 1, pp. 1–10, 2020.

[240] S. Yin, Y. Zuo, A. Abu-Odeh, *et al.*, "Atomistic simulations of dislocation mobility in refractory high-entropy alloys and the effect of chemical short-range order," *Nature communications*, vol. 12, no. 1, pp. 1–14, 2021.

[241] T. Kostiuchenko, F. Körmann, J. Neugebauer, and A. Shapeev, "Impact of lattice relaxations on phase transitions in a high-entropy alloy studied by machine-learning potentials," *npj Computational Materials*, vol. 5, no. 1, pp. 1–7, 2019.

[242] Q.-J. Hong and A. van de Walle, "Solid-liquid coexistence in small systems: A statistical method to calculate melting temperatures," *The Journal of chemical physics*, vol. 139, no. 9, p. 094 114, 2013.

[243] Z. Wu, H. Bei, F. Otto, G. Pharr, and E. George, "Recovery, recrystallization, grain growth and phase stability of a family of fcc-structured multi-component equiatomic solid solution alloys," *Intermetallics*, vol. 46, pp. 131–140, 2014, ISSN: 0966-9795.

[244] J. J. Valencia and P. N. Quested, "Thermophysical properties," *NIST*, 2013, Table 5: Thermal expansion of selected pure metals at temperatures close to melting.

[245] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[246] M. Hutchinson, *Python package: Lolopy*, Version 1.1.1. URL: https://pypi.org/project/lolopy/1.1.1/, 2020.

[247] D. Jha, L. Ward, A. Paul, *et al.*, "Elemnet: Deep learning the chemistry of materials from only elemental composition," *Scientific reports*, vol. 8, no. 1, pp. 1–13, 2018.

[248] Y. Zhang, C. Wen, C. Wang, *et al.*, "Phase prediction in high entropy alloys with a rational selection of materials descriptors and machine learning models," *Acta Materialia*, vol. 185, pp. 528–539, 2020.

[249] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, and Y. Yang, "Machine learning guided appraisal and exploration of phase design for high entropy alloys," *npj Computational Materials*, vol. 5, no. 1, pp. 1–9, 2019.

[250] S. P. Ong, W. D. Richards, A. Jain, *et al.*, *Python materials genomics (pymatgen): A robust, open-source python library for materials analysis*, 2013.

[251] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[252] E. Vazquez and J. Bect, "Convergence properties of the expected improvement algorithm with fixed mean and covariance functions," *Journal of Statistical Planning and inference*, vol. 140, no. 11, pp. 3088–3095, 2010.

[253] J. W. Yeh, Y. L. Chen, S. J. Lin, and S. K. Chen, "High-entropy alloys–a new era of exploitation," in *Materials Science Forum*, Trans Tech Publ, vol. 560, 2007, pp. 1–9.

[254] E. Pickering and N. Jones, "High-entropy alloys: A critical assessment of their founding principles and future prospects," *International Materials Reviews*, vol. 61, no. 3, pp. 183–202, 2016.

[255] D. B. Miracle and O. N. Senkov, "A critical review of high entropy alloys and related concepts," *Acta Materialia*, vol. 122, pp. 448–511, 2017.

[256] O. Senkov, J. Miller, D. Miracle, and C. Woodward, "Accelerated exploration of multi-principal element alloys with solid solution phases," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[257] M. C. Troparevsky, J. R. Morris, P. R. Kent, A. R. Lupini, and G. M. Stocks, "Criteria for predicting the formation of single-phase high-entropy alloys," *Physical Review X*, vol. 5, no. 1, p. 011 041, 2015.

[258] A. Sharma, R. Singh, P. K. Liaw, and G. Balasubramanian, "Cuckoo searching optimal composition of multicomponent alloys by molecular simulations," *Scripta Materialia*, vol. 130, pp. 292–296, 2017.

[259] P. Singh, A. Sharma, A. V. Smirnov, *et al.*, "Design of high-strength refractory complex solid-solution alloys," *npj Computational Materials*, vol. 4, no. 1, pp. 1–8, 2018.

[260] S. Gorsse, M. Nguyen, O. Senkov, and D. Miracle, "Database on the mechanical properties of high entropy alloys and complex concentrated alloys," *Data in Brief*, vol. 21, pp. 2664–2678, 2018, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2018.11.111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234091831504X.

[261] S. M. Lundberg, G. Erion, H. Chen, *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature machine intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[262] C. K. Borg, C. Frey, J. Moh, *et al.*, "Expanded dataset of mechanical properties and observed phases of multi-principal element alloys," *Scientific Data*, vol. 7, no. 1, pp. 1–6, 2020.

[263] C. Varvenne, A. Luque, and W. A. Curtin, "Theory of strengthening in fcc high entropy alloys," *Acta Materialia*, vol. 118, pp. 164–176, 2016.

[264] F. Maresca and W. A. Curtin, "Theory of screw dislocation strengthening in random bcc alloys from dilute to "high-entropy" alloys," *Acta Materialia*, vol. 182, pp. 144–162, 2020.

[265] F. Maresca and W. A. Curtin, "Mechanistic origin of high strength in refractory bcc high entropy alloys up to 1900k," *Acta Materialia*, vol. 182, pp. 235–249, 2020.

[266] J. Rickman, H. Chan, M. Harmer, *et al.*, "Materials informatics for the screening of multi-principal elements and high-entropy alloys," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[267] P. Sarker, T. Harrington, C. Toher, *et al.*, "High-entropy high-hardness metal carbides discovered by entropy descriptors," *Nature communications*, vol. 9, no. 1, pp. 1–10, 2018.

[268] R. Chen, G. Qin, H. Zheng, *et al.*, "Composition design of high entropy alloys using the valence electron concentration to balance strength and ductility," *Acta Materialia*, vol. 144, pp. 129–137, 2018.

[269] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.

[270] L. S. Shapley, "A value for n-person games," *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[271] S. Guo, C. Ng, J. Lu, and C. Liu, "Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys," *Journal of applied physics*, vol. 109, no. 10, p. 103 505, 2011.

[272] Z. D. McClure and A. Strachan, *Feature selection for cca strength models*, 2021. DOI: doi:10.21981/K4Z5-K344. [Online]. Available: https://nanohub.org/resources/ccaoptimizer.

[273] M. S. Farnell, Z. D. McClure, S. Tripathi, and A. Strachan, "Modeling environment-dependent atomic-level properties in complex-concentrated alloys," *The Journal of Chemical Physics*, 2022.

[274] D. Miracle and O. Senkov, "A critical review of high entropy alloys and related concepts," *Acta Materialia*, vol. 122, pp. 448–511, 2017.

[275] O. N. Senkov, D. B. Miracle, K. J. Chaput, and J.-P. Couzinie, "Development and exploration of refractory high entropy alloys—a review," *Journal of Materials Research*, vol. 33, no. 19, pp. 3092–3128, 2018.

[276] W. Nohring and W. Curtin, "Correlation of microdistortions with misfit volumes in high entropy alloys," *Scripta Materialia*, vol. 168, pp. 119–123, 2019.

[277] G. Kim, H. Diao, C. Lee, *et al.*, "First-principles and machine learning predictions of elasticity in severely lattice-distorted high-entropy alloys with experimental validation," *Acta Materialia*, vol. 181, pp. 124–138, 2019.

[278] W. Huang, P. Martin, and L. Z. Zhuang Houlong, "Machine-learning phase prediction of high-entropy alloys," *Acta Materialia*, vol. 169, pp. 225–236, 2019.

[279] X. Liu, J. Zhang, M. Eisenbach, and Y. Wang, *Machine learning modeling of high entropy alloy: The role of short-range order*, 2019.

[280] W. Chen, X. Ding, Y. Feng, *et al.*, "Vacancy formation enthalpies of high-entropy fecocrni alloy via first-principles calculations and possible implications to its superior radiation tolerance," *Journal of Materials Science & Technology*, vol. 34, pp. 355–364, 2018.

[281] R. Machaka, "Machine learning-based prediction of phases in high-entropy alloys," *Computational Materials Science*, vol. 188, p. 110 244, 2021, ISSN: 0927-0256. DOI: https://doi.org/10.1016/j.commatsci.2020.110244. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025620307357.

[282] K. Kaufmann and K. S. Vecchio, "Searching for high entropy alloys: A machine learning approach," *Acta Materialia*, vol. 198, pp. 178–222, 2020, ISSN: 1359-6454. DOI: https://doi.org/10.1016/j.actamat.2020.07.065. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359645420305814.

[283] J. M. Rickman, G. Balasubramanian, C. J. Marvel, H. M. Chan, and M.-T. Burton, "Machine learning strategies for high-entropy alloys," *Journal of applied physics*, vol. 128, no. 22, 2020, ISSN: 0021-8979.

[284] J. M. Sanchez, F. Ducastelle, and D. Gratias, "Generalized cluster description of multi-component systems," *Physica A: Statistical Mechanics and its Applications*, vol. 128, no. 1-2, pp. 334–350, 1984.

[285] S. Hao, L.-D. Zhao, C.-Q. Chen, V. P. Dravid, M. G. Kanatzidis, and C. M. Wolverton, "Theoretical prediction and experimental confirmation of unusual ternary ordered semiconductor compounds in sr–pb–s system," *Journal of the American Chemical Society*, vol. 136, no. 4, pp. 1628–1635, 2014.

[286] I. Toda-Caraballo, J. Wróbel, S. Dudarev, D. Nguyen-Manh, and P. Rivera-Díaz-del-Castillo, "Interatomic spacing distribution in multicomponent alloys," *Acta Mat.*, vol. 97, pp. 156–169, 2015.

[287] J. Sanchez, "Foundations and practical implementations of the cluster expansion," *Journal of Phase Equilibria and Diffusion*, vol. 38, no. 3, pp. 238–251, 2017.

[288] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016.

[289] A. Shapeev, "Accurate representation of formation energies of crystalline alloys with many components," *Computational Materials Science*, vol. 139, pp. 26–30, 2017.

[290] S. Daigle and D. Brenner, "Statistical approach to obtaining vacancy formation energies in high-entropy crystals from first principles calculations: Application to a high-entropy diboride," *Physical Review Materials*, vol. 4, no. 12, p. 123 602, 2020.

[291] J. Behler, "Perspective: Machine learning potentials for atomistic simulations," *Journal of Chemical Physics*, vol. 4, p. 053 208, 2016.

[292] K. Gubaev, E. V. Podryabinkin, G. L. Hart, and A. V. Shapeev, "Accelerating high-throughput searches for new alloys with active learning of interatomic potentials," *Computational Materials Science*, vol. 156, pp. 148–156, 2019.

[293] A. P. Bartok, R. Kondor, and G. Csanyi, "On representing chemical environments," *Physical Review B*, 2013.

[294] A. Zunger, S.-H. Wei, L. Ferreira, and J. E. Bernard, "Special quasirandom structures," *Physical Review Letters*, vol. 65, no. 3, p. 353, 1990.

[295] D. Tsai, "The virial theorem and stress calculation in molecular dynamics," *The Journal of Chemical Physics*, vol. 70, no. 3, pp. 1375–1382, 1979.

[296] C. Rycroft, "Voro++: A three-dimensional voronoi cell library in c++," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2009.

[297] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations," *Physical Review Letters*, vol. 125, no. 16, p. 166 001, 2020.

[298] Y. Zuo, C. Chen, X. Li, *et al.*, "Performance and cost assessment of machine learning interatomic potentials," *The Journal of Physical Chemistry A*, vol. 124, no. 4, pp. 731–745, 2020.

[299] A. Roy and G. Balasubramanian, "Predictive descriptors in machine learning and data-enabled explorations of high-entropy alloys," eng, *Computational materials science*, vol. 193, 2021, ISSN: 0927-0256.

[300] nanoHUB, *Jupyter notebook*, 2016. DOI: doi:10.21981/W6TE-1750. [Online]. Available: https://nanohub.org/resources/jupyter.

[301] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[302] F. Chollet *et al.*, *Keras*, 2015. [Online]. Available: https://github.com/fchollet/keras.

[303] W. Zhang, R. Mazzarello, M. Wuttig, and E. Ma, "Designing crystallization in phase-change materials for universal memory and neuro-inspired computing," *Nature Reviews Materials*, vol. 4, no. 3, pp. 150–168, 2019.

[304] Z. Wang, H. Wu, G. W. Burr, *et al.*, "Resistive switching materials for information processing," *Nature Reviews Materials*, vol. 5, no. 3, pp. 173–195, 2020.

[305] N. Yamada, E. Ohno, N. Akahira, K. Nishiuchi, K. Nagata, and M. Takao, "High speed overwritable phase change optical disk material," *Japanese Journal of Applied Physics*, vol. 26, no. S4, p. 61, 1987.

[306] N. Yamada, E. Ohno, K. Nishiuchi, N. Akahira, and M. Takao, "Rapid-phase transitions of gete-sb2te3 pseudobinary amorphous thin films for an optical disk memory," *Journal of Applied Physics*, vol. 69, no. 5, pp. 2849–2856, 1991.

[307] H. Horii, J. Yi, J. Park, *et al.*, "A novel cell technology using n-doped gesbte films for phase change ram," in *2003 Symposium on VLSI Technology. Digest of Technical Papers (IEEE Cat. No. 03CH37407)*, IEEE, 2003, pp. 177–178.

[308] S. Raoux, M. Salinga, J. L. Jordan-Sweet, and A. Kellock, "Effect of al and cu doping on the crystallization properties of the phase change materials sbte and gesb," *Journal of applied physics*, vol. 101, no. 4, p. 044 909, 2007.

[309] Q. Hubert, C. Jahan, A. Toffoli, *et al.*, "Lowering the reset current and power consumption of phase-change memories with carbon-doped ge2sb2te5," in *2012 4th IEEE International Memory Workshop*, IEEE, 2012, pp. 1–4.

[310] W. Song, L. Shi, X. Miao, and T. Chong, "Phase change behaviors of sn-doped ge–sb–te material," *Applied Physics Letters*, vol. 90, no. 9, p. 091 904, 2007.

[311] S. Privitera, E. Rimini, and R. Zonca, "Amorphous-to-crystal transition of nitrogen-and oxygen-doped ge 2 sb 2 te 5 films studied by in situ resistance measurements," *Applied physics letters*, vol. 85, no. 15, pp. 3044–3046, 2004.

[312] J.-H. Eom, Y.-G. Yoon, C. Park, *et al.*, "Global and local structures of the ge-sb-te ternary alloy system for a phase-change memory device," *Physical Review B*, vol. 73, no. 21, p. 214 202, 2006.

[313] J. Hegedüs and S. Elliott, "Microscopic origin of the fast crystallization ability of ge–sb–te phase-change memory materials," *Nature materials*, vol. 7, no. 5, pp. 399–405, 2008.

[314] S. Caravati, M. Bernasconi, T. Kühne, M. Krack, and M. Parrinello, "Coexistence of tetrahedral-and octahedral-like sites in amorphous phase change materials," *Applied Physics Letters*, vol. 91, no. 17, p. 171 906, 2007.

[315] T. Lee and S. Elliott, "Ab initio computer simulation of the early stages of crystallization: Application to ge 2 sb 2 te 5 phase-change materials," *Physical review letters*, vol. 107, no. 14, p. 145 702, 2011.

[316] F. Yang, Y. Tao, L. Zhang, *et al.*, "Ab initio study on the fast reversible phase transitions of ge2sb2te5," *Journal of Applied Physics*, vol. 130, no. 2, p. 025 106, 2021.

[317] F. Zipoli and A. Curioni, "Reactive potential for the study of phase-change materials: Gete," *New Journal of Physics*, vol. 15, no. 12, p. 123 006, 2013.

[318] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, "Neural network interatomic potential for the phase change material gete," *Physical Review B*, vol. 85, no. 17, p. 174 103, 2012.

[319] F. C. Mocanu, K. Konstantinou, T. H. Lee, *et al.*, "Modeling the phase-change memory material, ge2sb2te5, with a machine-learned interatomic potential," *The Journal of Physical Chemistry B*, vol. 122, no. 38, pp. 8998–9006, 2018.

[320] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," *Physical Review B*, vol. 99, no. 1, p. 014 104, 2019.

[321] S. Lorenz, A. Groß, and M. Scheffler, "Representing high dimensional potential energy surfaces for reactions at surfaces by neural networks," *Chemical Physics Letters*, vol. 395, no. 4-6, pp. 210–215, 2004.

[322] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chemistry of Materials*, vol. 31, no. 9, pp. 3564–3572, 2019.

[323] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Physical review letters*, vol. 108, no. 5, p. 058 301, 2012.

[324] H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, and A. Tkatchenko, "Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces," *The Journal of chemical physics*, vol. 150, no. 11, p. 114 102, 2019.

[325] A. Vitek, M. Stachon, P. Krömer, and V. Snáel, "Towards the modeling of atomic and molecular clusters energy by support vector regression," in *2013 5th International Conference on Intelligent Networking and Collaborative Systems*, IEEE, 2013, pp. 121–126.

[326] M. S. Daw and M. I. Baskes, "Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals," *Physical Review B*, vol. 29, no. 12, p. 6443, 1984.

[327] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and E. Weinan, "Active learning of uniformly accurate interatomic potentials for materials simulation," *Physical Review Materials*, vol. 3, no. 2, p. 023 804, 2019.

[328] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, "Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning," *Physical Review B*, vol. 99, no. 6, p. 064 114, 2019.

[329] N. Bernstein, G. Csányi, and V. L. Deringer, "De novo exploration and self-guided learning of potential-energy surfaces," *npj Computational Materials*, vol. 5, no. 1, pp. 1–9, 2019.

[330] C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," *The Journal of Chemical Physics*, vol. 153, no. 10, p. 104 105, 2020.

[331] P. E. Blöchl, "Projector augmented-wave method," *Physical Review B*, vol. 50, no. 24, pp. 17 953–17 979, Dec. 1994.

[332] S. Grimme, "Semiempirical gga-type density functional constructed with a long-range dispersion correction," *Journal of computational chemistry*, vol. 27, no. 15, pp. 1787–1799, 2006.

[333] G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," *Computational Materials Science*, vol. 6, no. 1, pp. 15–50, Jul. 1996.

[334] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initiototal-energy calculations using a plane-wave basis set," *Physical Review B*, vol. 54, no. 16, pp. 11 169–11 186, Oct. 1996.

[335] B. Kooi and J. T. M. De Hosson, "Electron diffraction and high-resolution transmission electron microscopy of the high temperature crystal structures of ge x sb 2 te 3+ x (x= 1, 2, 3) phase change material," *Journal of applied physics*, vol. 92, no. 7, pp. 3584–3590, 2002.

[336] *Materials Project*. [Online]. Available: http://www.materialsproject.org.

[337] P. Urban, M. N. Schneider, L. Erra, S. Welzmiller, F. Fahrnbauer, and O. Oeckler, "Temperature dependent resonant x-ray diffraction of single-crystalline ge 2 sb 2 te 5," *CrystEngComm*, vol. 15, no. 24, pp. 4823–4829, 2013.

[338] J. Akola and R. O. Jones, "Density functional study of amorphous, liquid and crystalline $Ge_2Sb_2Te_5$: homopolar bonds and/or AB alternation?," vol. 20, no. 46, p. 465 103, 2008. DOI: 10.1088/0953-8984/20/46/465103. [Online]. Available: https://doi.org/10.1088/0953-8984/20/46/465103.

[339] S. Caravati, M. Bernasconi, T. D. Kühne, M. Krack, and M. Parrinello, "First-principles study of crystalline and amorphous $Ge_2Sb_2Te_5$ and the effects of stoichiometric defects," *Journal of Physics: Condensed Matter*, vol. 21, no. 25, p. 255 501, 2009. DOI: 10.1088/0953-8984/21/25/255501. [Online]. Available: https://doi.org/10.1088/0953-8984/21/25/255501.

[340] J. Sun, S. Mukhopadhyay, A. Subedi, T. Siegrist, and D. J. Singh, "Transport properties of cubic crystalline $Ge_2Sb_2Te_5$: A potential low-temperature thermoelectric material," *Applied Physics Letters*, vol. 106, no. 12, p. 123 907, 2015. DOI: 10.1063/1.4916558. eprint: https://doi.org/10.1063/1.4916558. [Online]. Available: https://doi.org/10.1063/1.4916558.

[341] T. Nonaka, G. Ohbayashi, Y. Toriumi, Y. Mori, and H. Hashimoto, "Crystal structure of GeTe and $Ge_2Sb_2Te_5$ meta-stable phase," *Thin Solid Films*, vol. 370, no. 1, pp. 258–261, 2000, ISSN: 0040-6090. DOI: https://doi.org/10.1016/S0040-6090(99)01090-1. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040609099010901.

[342] N. Yamada and T. Matsunaga, "Structure of laser-crystallized $Ge_2Sb_2 + xTe_5$ sputtered thin films for use in optical memory," *Journal of Applied Physics*, vol. 88, no. 12, pp. 7020–7028, 2000. DOI: 10.1063/1.1314323. eprint: https://doi.org/10.1063/1.1314323. [Online]. Available: https://doi.org/10.1063/1.1314323.

[343] T. Matsunaga, N. Yamada, and Y. Kubota, "Structures of stable and metastable Ge2Sb2Te5, an intermetallic compound in GeTe–Sb2Te3 pseudobinary systems," *Acta Crystallographica Section B*, vol. 60, no. 6, pp. 685–691, 2004. DOI: 10.1107/S0108768104022906. [Online]. Available: https://doi.org/10.1107/S0108768104022906.

[344] W. K. Njoroge, H.-W. Wöltgens, and M. Wuttig, "Density changes upon crystallization of ge 2 sb 2.04 te 4.74 films," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 20, no. 1, pp. 230–233, 2002.

[345] K. Hansen, F. Biegler, R. Ramakrishnan, *et al.*, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," *The journal of physical chemistry letters*, vol. 6, no. 12, pp. 2326–2331, 2015.

[346] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, "Wacsf: Weighted atom-centered symmetry functions as descriptors in machine learning potentials," *The Journal of chemical physics*, vol. 148, no. 24, p. 241 709, 2018.

[347] J. Behler, "Four generations of high-dimensional neural network potentials," *Chemical Reviews*, vol. 121, no. 16, pp. 10 037–10 072, 2021.

[348] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials," *The Journal of chemical physics*, vol. 148, no. 24, p. 241 730, 2018.

[349] M. W. Mahoney and P. Drineas, "Cur matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[350] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error cur matrix decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 2, pp. 844–881, 2008.

[351] M. Gastegger and P. Marquetand, "High-dimensional neural network potentials for organic reactions and an improved training algorithm," *Journal of chemical theory and computation*, vol. 11, no. 5, pp. 2187–2198, 2015.

[352] A. Singraber, T. Morawietz, J. Behler, and C. Dellago, "Parallel multistream training of high-dimensional neural network potentials," *Journal of chemical theory and computation*, vol. 15, no. 5, pp. 3075–3092, 2019.

[353] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[354] A. Singraber, J. Behler, and C. Dellago, "Library-based lammps implementation of high-dimensional neural network potentials," *Journal of chemical theory and computation*, vol. 15, no. 3, pp. 1827–1840, 2019.

[355] P. H. Berens, D. H. Mackay, G. M. White, and K. R. Wilson, "Thermodynamics and quantum corrections from molecular dynamics for liquid water," *The Journal of chemical physics*, vol. 79, no. 5, pp. 2375–2389, 1983.

[356] J. Orava, A. á. Greer, B. Gholipour, D. Hewak, and C. Smith, "Characterization of supercooled liquid ge2sb2te5 and its crystallization by ultrafast-heating calorimetry," *Nature materials*, vol. 11, no. 4, pp. 279–283, 2012.

[357] Z. D. McClure, S. Desai, and A. Strachan, *High entropy alloy melting point calculation*, URL: https://nanohub.org/resources/meltheas, DOI: 10.21981/W5VD-T039, 2020.

[358] J. C. Verduzco Gastelum, D. E. Farache, Z. D. McClure, S. Desai, and A. Strachan, *Active learning workflow for mpcas*, URL: https://nanohub.org/resources/activemeltheas, DOI: 10.21981/NK7E-HA16, 2021.

[359] J. C. V. Gastelum, Z. D. McClure, and A. Strachan, *Visualization dashboard for mpcas*, 2022. DOI: doi:10.21981/A6GT-V318. [Online]. Available: https://nanohub.org/resources/meltdashboard.

[360] M. Farnell, Z. McClure, and A. Strachan, *Machine learning for high entropy atomic properties*, 2021. DOI: https://doi.org/10.21981/FMF8-AM68. [Online]. Available: https://nanohub.org/tools/mlatomprop.

[361] nanoHUB, *Group: Hands-on learning modules on data science and machine learning in engineering*. [Online]. Available: https://nanohub.org/groups/mlmodules.

[362] Z. D. McClure and A. Strachan, *Module 2: Querying materials data repositories*, 2020. [Online]. Available: https://nanohub.org/resources/34288.

[363] Z. D. McClure and A. Strachan, *Querying data repositories*, 2020. DOI: doi:10.21981/39S9-TG18. [Online]. Available: https://nanohub.org/resources/matdatarepo.

[364] Z. D. McClure and A. Strachan, 2020. DOI: doi:10.21981/J09R-4Q37. [Online]. Available: https://nanohub.org/resources/featureselect.

[365] A. Johnston, A. Strachan, C. Wang, *et al.*, *Matlab coding and data analysis in the context of radiation hardening*, 2020. DOI: doi:10.21981/GYKR-5Z02. [Online]. Available: https://nanohub.org/resources/matlabrad.

[366] K. Nykiel, A. Leichty, Z. D. McClure, *et al.*, *Matlab data analysis using jupyter notebooks*, 2020. DOI: doi:10.21981/8XRN-X942. [Online]. Available: https://nanohub.org/resources/matlabdata.

[367] M. N. Sakano, A. Hamed, E. M. Kober, *et al.*, "Unsupervised learning-based multiscale model of thermochemistry in 1, 3, 5-trinitro-1, 3, 5-triazinane (rdx)," *The Journal of Physical Chemistry A*, vol. 124, no. 44, pp. 9141–9155, 2020.

[368] S. Desai, S. T. Reeve, and J. F. Belak, "Implementing a neural network interatomic model with performance portability for emerging exascale architectures," *Computer Physics Communications*, vol. 270, p. 108 156, 2022.

[369] S. Goswami, C. Anitescu, S. Chakraborty, and T. Rabczuk, "Transfer learning enhanced physics informed neural network for phase-field modeling of fracture," *Theoretical and Applied Fracture Mechanics*, vol. 106, p. 102 447, 2020.

VITA

Zachary McClure graduated with his Ph.D. in Materials Engineering from Purdue University in the spring of 2022. His education was preceded by a B.S. in Chemical Engineering in 2017 from Oregon State University. During his time as a graduate student he focused on applications of computational modeling solutions for materials screening, experimental investigation, and interatomic potential development. Leveraging information from electronic structure theory and dynamic simulations, his work helped guide new pathways for materials design and modeling within the fields of nanoscale metallics, high entropy alloys, and phase change memory devices. His primary interests have aligned with using machine learning and modern cyberinfrastructure to explore new ways of combining tools within the field of materials engineering. In addition to his work as an academic he has mentored multiple undergraduate students leading to peer-reviewed publications, and development open-source computational tools for reduced entry barriers to the field. Along with three of his colleagues, his efforts during the Covid-19 pandemic in creating educational resources and hands-on workshops was recognized by the university with the Boilermaker Changemaker Award.

To continue broadening his horizons after his time at Purdue, he joined the Life Sciences Division at Nvidia Corp. as a Computational Chemist investigating molecular dynamic simulation scaling for drug discovery and genomics initiatives.