Machine Intelligence for Materials Science

N. M. Anoop Krishnan
Hariprasad Kodamana
Ravinder Bhattoo

Machine Learning for Materials Discovery

Numerical Recipes and Practical Applications

Springer

Machine Intelligence for Materials Science

Series Editor

N. M. Anoop Krishnan, Department of Civil Engineering, Yardi School of Artificial Intelligence (Joint Appt.), Indian Institute of Technology Delhi, New Delhi, India

This book series is dedicated to showcasing the latest research and developments at the intersection of materials science and engineering, computational intelligence, and data sciences. The series covers a wide range of topics that explore the application of artificial intelligence (AI), machine learning (ML), deep learning (DL), reinforcement learning (RL), and data science approaches to solve complex problems across the materials research domain.

Topical areas covered in the series include but are not limited to:

- AI and ML for accelerated materials discovery, design, and optimization
- Materials informatics
- Materials genomics
- Data-driven multi-scale materials modeling and simulation
- Physics-informed machine learning for materials
- High-throughput materials synthesis and characterization
- Cognitive computing for materials research

The series also welcomes manuscript submissions exploring the application of AI, ML, and data science techniques to following areas:

- Materials processing optimization
- Materials degradation and failure
- Additive manufacturing and 3D printing
- Image analysis and signal processing

Each book in the series is written by experts in the field and provides a valuable resource for understanding the current state of the field and the direction in which it is headed. Books in this series are aimed at researchers, engineers, and academics in the field of materials science and engineering, as well as anyone interested in the impact of AI on the field.

N. M. Anoop Krishnan · Hariprasad Kodamana · Ravinder Bhattoo

Machine Learning for Materials Discovery

Numerical Recipes and Practical Applications

Springer

N. M. Anoop Krishnan
Department of Civil Engineering, Yardi School of Artificial Intelligence (Joint Appt.)
Indian Institute of Technology Delhi
New Delhi, India

Hariprasad Kodamana
Department of Chemical Engineering, Yardi School of Artificial Intelligence (Joint Appt.)
Indian Institute of Technology Delhi
New Delhi, India

Ravinder Bhattoo
Indian Institute of Technology Delhi
New Delhi, India

ISSN 2948-1813 ISSN 2948-1821 (electronic)
Machine Intelligence for Materials Science
ISBN 978-3-031-44621-4 ISBN 978-3-031-44622-1 (eBook)
https://doi.org/10.1007/978-3-031-44622-1

To My Parents, Parents-in-law, Arathy, and Anant—NMAK

To My Parents, Susha, Avyay, and Adwait—HK

To My Parents, Pushpa, and Priyanka—RB

Foreword

Historically, science and engineering domains periodically experience revolutionary ideas that completely change the way we think about the domain. A few decades ago, it was the introduction of nanoengineering materials and the emergence of predictive multiscale modeling. Today we are in the cusp of a similar revolution. Advances in machine learning and artificial intelligence are now enabling tasks that would have seemed impossible just a short while ago, or would have been taking decades to achieve. Impacts of these new tools include the discovery of novel drugs, ultrahigh strength alloys, automated design of composites, efficient quantum accuracy simulations, bioinspired design, and knowledge transfer across domains, just to name a few. The new book by Prof. Krishnan, Prof. Kodamana, and Dr. Bhattoo provides an excellent introduction into the emerging field of machine learning for materials discovery. This book bridges a gap and acts as an enabler for the adoption of machine learning by material scientists, engineers, and students.

The book offers an excellent pedagogical approach towards the use of machine learning for materials discovery. The book is written in a lucid fashion, and accessible to audience ranging from undergraduate students to scientists. The book does not assume any prior knowledge in the domain of machine learning, and is self-sufficient. The second part of the book covers the basics of machine learning theory including supervised and unsupervised strategies with examples from the materials domain. An excellent feature of the book is that theory on machine learning is followed by codes that allows instructors, students, and practitioners to try the approaches in a hands-on fashion. The third section discusses a wide range of applications giving an overview of different avenues where machine learning can be used for materials discovery.

While research aspect of the topic is interesting, it is equally, if not more, important to train the next-generation materials scientists to be skilled in machine learning and artificial intelligence, especially to being able to critically discern the best modeling

strategies among a broad set of tools. I believe this book can give an impetus for the adoption of machine learning in materials science curricula across many universities. I hope you will enjoy reading this excellent book as much as I did.

Markus J. Buehler
Massachusetts Institute of Technology (MIT)
Cambridge, USA

Preface

The last decade in materials science has seen a major wave of change due to the advent of machine learning and artificial intelligence. While there have been significant advances in machine learning for the materials domain, a vast majority of students, researchers, and professionals working on materials still do not have access to the theoretical backgrounds of machine learning. This can be attributed partially to the intricate mathematical treatments commonly followed in many machine learning textbooks and the use of general examples that lack relevance to materials science-related applications.

This textbook aims to bridge this gap by providing an overview of machine learning in materials modeling and discovery. The textbook is well-suited for a diverse audience, including undergraduates, graduates, and industry professionals. The book is also structured as foundational and can be used as a textbook covering the basics and advanced techniques while giving hands-on examples using Python codes.

The book is structured into three parts. Part I gives an introduction to the evolution of machine learning in the materials domain. Part II focuses on building the foundations of machine learning, with various tailor-made examples accompanied by corresponding code implementations. In the part III, emphasis is given to several practical applications related to machine learning in the materials domain.

Although several use cases from the literature are covered, the book also integrates examples from the authors' research whenever possible. This deliberate choice is motivated by accessible data and first-hand details of available codes that might not readily exist in the literature. We believe such a treatment facilitates comprehensive information about practical implementation while striking a balance with the theoretical exposition.

The field of machine learning is growing at an exponential pace, and it is impossible to cover all the state-of-the-art methods. This book by no means is exhaustive. Rather, this book is an attempt to capture the essence of the basics of machine learning and make the readers aware of the foundations so that they can either delve into the deeper aspects of machine learning or focus on the applications to the materials domain using existing approaches to solve an impactful problem in the domain.

We hope you enjoy the book and find it useful for your journey in materials discovery.

New Delhi, India
August 2023

N. M. Anoop Krishnan
Hariprasad Kodamana
Ravinder Bhattoo

Acknowledgements

There are a lot of people who have contributed both actively and passively to the development of this book. First, we would like to thank our editor, Dr. Zachary Evenson, for initiating the idea of the book and encouraging us to complete it. It is indeed his motivation and support that resulted in the book. Thanks to Mohd Zaki for helping with the images and suggestions on graphics. Thanks are also due to Indrajeet Mandal, who painstakingly collected the copyrights for all the images used in the work. Special thanks to the research scholars of the M3RG at IIT Delhi for their comments, feedback, and proofreading that helped significantly improve the book. The authors also thank the support from IIT Delhi and specifically the Yardi School of Artificial Intelligence, Department of Civil Engineering, and Department of Chemical Engineering. The role played by the authors' family in the form of continuous support to complete the book cannot be emphasized enough. Boundless thanks to them for supporting us through this endeavor through thick and thin, COVID and many other uncertainties and challenges, and making this happen.

Contents

Acronyms

ADASYN	Adaptive synthetic sampling technique
AFM	Atomic force microscope
AI	Artificial intelligence
CNN	Convolutional neural network
CALPHAD	Calculation of phase diagrams
CSD	Cambridge structural database
DBSCAN	Density-based spatial clustering of applications with noise
EBSD	Electron backscatter diffraction
FEM	Finite element methods
GAN	Generative adversarial networks
GCN	Graph convolutional network
GNN	Graph neural network
GPR	Gaussian process regression
GRU	Gated recurrent units
IID	Independent and identically distributed
LAR	Least angle regression
LARS	Least angle regression and shrinkage
LMS	Least mean square
LSTM	Long short-term memory
LWR	Locally weighted linear regression
MD	Molecular dynamics
MC	Monte-Carlo
ML	Machine learning
MLP	Multi-layer perceptron
NLP	Natural language processing
NN	Neural network
OLS	Ordinary least square
OPTICS	Ordering points to identify cluster structure
PCA	Principal component analysis
QSPR	Quantitative structure–property relationships
RF	Random forest

RL	Reinforcement learning
RNN	Recurrent neural network
SARSA	State-action-reward-state-action
SEM	Scanning electron microscope
SMOTE	Synthetic minority oversampling technique
SVM	Support vector machine
SVR	Support vector regression
t-SNE	t-distributed stochastic neighbor embedding
XGBoost	Extreme gradient boosting

Part I
Introduction

Chapter 1
Introduction

Abstract Materials form the basis of human civilization. With the advance of computational algorithms, computational power, and cloud-based services, materials innovation is accelerating at a pace never witnessed by humankind. In this chapter, we briefly introduce the materials discovery approaches using AI and ML that has enabled some breakthrough in our understanding of materials. We list some publicly available databases on materials and some of the applications where AI and ML has been used to design and discover novel materials. The chapter concludes with a brief outline of the book.

1.1 Materials Discovery

The progress of human civilization has been closely related to the discovery and usage of new materials. Materials have shaped how we interact with the world, from the stone to the silicon age. This is exemplified by the fact that the different ages of human history have been named after the prominent materials used in those eras–the stone age, the bronze age, and the iron age. A surge of new materials such as glass, steel, ceramics, concrete, and polymers marked the period during and after the iron age. Thus, everything we see around us, from pins and pots to rockets and robots, has been made possible due to the discovery of materials. As we advance, materials are sure to play a crucial role in the sustainable development of humans, with the most negligible impact on the planet, in areas such as renewable energy, health care, agriculture, and even arts and culture.

However, the importance of materials discovery was formally accepted only in the 1950s with the proposition of materials as a separate engineering domain. During world war II and the ensuing cold war, countries realized that materials were the bottleneck in advancing military, space, and medical technologies. Thus, materials science emerged as the first discipline formed out of the fusion and collaborations of multiple disciplines from basic sciences and engineering, focusing on understanding material response leading to materials discovery. While the early focus of materials

N. M. A. Krishnan et al., *Machine Learning for Materials Discovery*,
Machine Intelligence for Materials Science,
https://doi.org/10.1007/978-3-031-44622-1_1

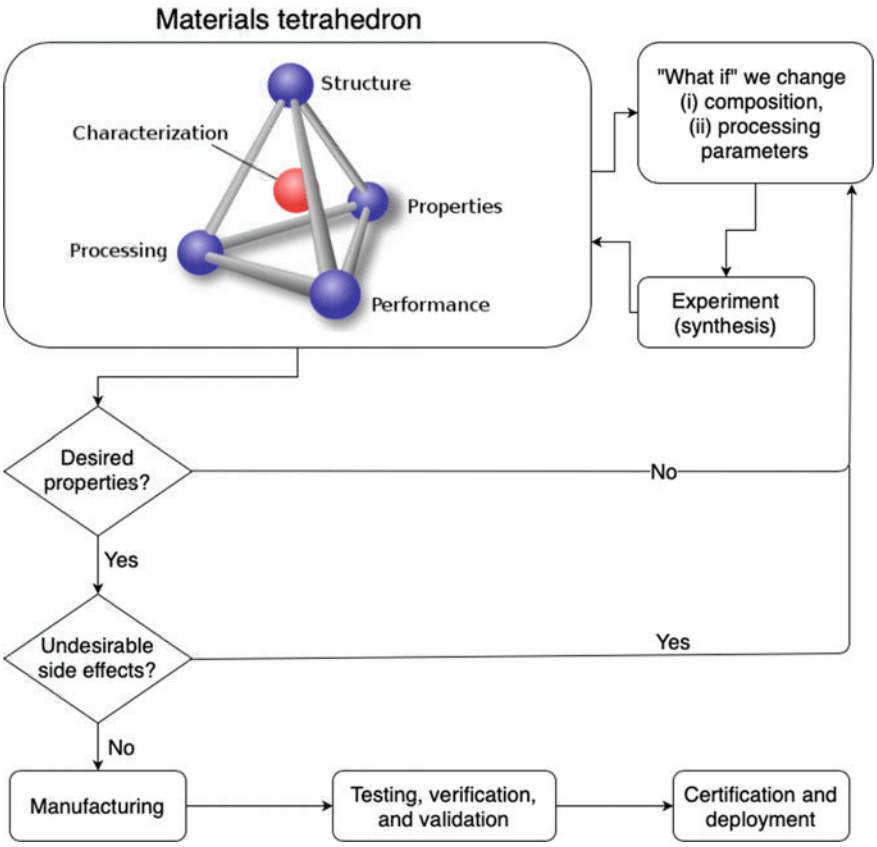

Fig. 1.1 Flow chart of traditional materials discovery based on *what-if* scenarios. Intuition and expert knowledge is used to cleverly pose the *what-if* questions that can potentially lead to the discovery of novel materials

science remained in metallurgy, it was soon expanded to other domains such as ceramics, polymers, and later to composites, nano-materials, and bio-materials.

The earlier approaches for materials discovery relied on trial-and-error approaches driven either by physics or strong intuition developed through years of experience. In such cases, the idea of what-if scenarios was used for discovering materials with tailored properties, as shown in Fig. 1.1. This approach would start from a "what-if" question on one or more aspects of the tetrahedron materials: processing, structure, property, and performance. A set of candidate solutions would be proposed based on the available knowledge and intuitions. These solutions would be tested using experimental synthesis and characterizations. If a candidate solution meets the expected performance, the new material is manufactured, verified, validated, certified, and deployed in the industry. If any of the candidate solutions do not meet the expected performance, the iteration is continued until a desired candidate is discov-

ered. As a trivial example, consider the following. Carbon can improve the hardness and strength of steel–what if we increase the carbon content of steel? Experimental studies reveal that the increase in carbon content improves steel's hardness and strength. However, higher carbon content makes steel brittle and less weldable! Thus, although the new candidate meets the expected performance in terms of strength, it induces some undesirable *side-effects* on other properties. Hence, the candidate may not be accepted. Thus, the what-if scenarios required detailed and time-consuming experimental characterization and analysis of new materials, significantly increasing the cost and time required for materials discovery. In these cases, the typical timescale associated with the discovery of new material was 20–30 years from the initial research to its first use.

The invention of computers and in-silico approaches came as a breakthrough in materials discovery in the second half of the twentieth century. Monte Carlo (MC) algorithms and molecular dynamics (MD) simulations, both proposed in the 1950s, became valuable tools for understanding materials response under different scenarios. These approaches reduced the number of actual experiments to be carried out, accelerating materials discovery. At the same time, slowly but steadily, researchers also started realizing the importance of compiling and documenting the materials data generated by the experiments and simulations. The first attempts to this extent were the Cambridge Structural Database (CSD) and Calculation of Phase Diagrams (CALPHAD) around the 1970s. These databases enabled the development of a quantitative structure-property relationships (QSPR) approach in materials. The QSPR approaches primarily relied on correlations and simple linear or polynomial regressions that allowed the discovery of patterns from the available data, which ultimately provided insights into materials response.

List of Publically Available Materials Databases

1. CSD: Cambridge Structural Database
2. CALPHAD
3. Granta Design
4. Pauling File
5. ICSD: Inorganic Crystal Structure Database
6. ESP: Electronic Structure Project
7. AFLOW: Automatic-Flow for Materials Discovery
8. MatNavi
9. AIST: National Institute of Advanced Industrial Science and Technology Databases
10. COD: Crystallography Open Database
11. MatDL: Materials Digital Library
12. The Materials Project
13. CMR: Computational Materials Repository
14. Springer Material
15. OpenKIM
16. NREL CID: NREL Center for Inverse Design
17. MGI: Materials Genome Initiative

18. MatWeb
19. MATDAT
20. CEPDB: The Clean Energy Project Database
21. CMD: Computational Materials Network
22. Catalysis Hub
23. OQMD: Open Quantum Materials Database
24. Open Material Databases
25. NREL MatDB
26. Citrine Informatics
27. Exabyte.io
28. NOMAD: Novel Materials Discovery Laboratory
29. Marvel
30. Thermoelectrics Design Lab
31. MaX: Materials Design at the Exascale
32. CritCat
33. Khazana
34. Material Data Facility
35. MICCOM: Midwest Integrated Center for Computational Materials
36. MPDS: Materials Platform for Data Science
37. CMI2: Center for Materials Research by Information Integration
38. HTEM: High Throughput Experimental Materials Database
39. JARVIS: Joint Automated Repository for Various Integrated Simulations
40. OMDB: Organic Materials Database
41. aNANt
42. Atom Work Adv
43. FAIR Data Infrastructure
44. Materiae
45. Materials Zone
46. MolDis
47. QCArchive: The Quantum Chemistry Archive
48. PyGGi: Python for Glass Genomics

The next breakthrough in materials discovery could be attributed to the *internet revolution*, which democratized the access to data for everyone. This period also saw a surge in the development of experimental and computational databases which started serving as information repository and *cook-book* for material synthesis. Figure 1.2 shows the databases that are available and their geographical distribution. The availability of these databases also inspired the automated search for correlations in composition–structure–processing–property relationships of the materials. Thus, the stage was set for the use of machine learning for materials discovery with all the relevant ingredients in place, namely,

1. availability of large amounts data,
2. computational power to process and "learn" the data, and
3. extremely non-linear composition–structure–property relationships along with the poor understanding of physics governing these relationships in materials.

Fig. 1.2 Material database timeline and geographical region of origin. Reprinted with permission from [1]

1.2 Physics- and Data-Driven Modeling

Models are simplified replicas of real-world scenarios with attention to the features or phenomena of interest. For example, a ball-and-stick model of atoms aims to show the relative atomic positions for a given lattice, while completely ignoring the dynamics, electronic structure, and other details of an atomic system. Figure 1.3 shows the ball and stick model for benzene with the chemical formula C_6H_6. Note that the black balls represent the carbon atom while the white ones represent hydrogen atoms. Further, the alternating single and double bonds are represented beautifully by single and double sticks connecting the carbon atoms. Such models can be very useful for giving a quick understanding of complex molecular structures and are hence used commonly for teaching purposes.

While a ball-and-stick model is a physical model, phenomenons are typically expressed through mathematical models. Traditional models in materials and engineering disciplines have relied on mathematical equations derived based on physical theories or laws. This approach has been widely accepted for centuries and has stood the test of time. Some of the widely used mathematical models in materials science include laws of thermodynamics, Fick's laws, Avrami equation, Arrhenius equation, Gibbs–Thomson equation, Bragg's law, and Hooke's law. Thus, the physical models are derived based on existing theories and can be explained using reasoning to understand the phenomenon. However, the physical models have traditionally been limited to simple systems. The extremely complex and non-linear nature of advanced materials have remained elusive to physical models as well as in-silico models. Understanding the response of these materials require high-fidelity high-

Fig. 1.3 Ball and stick model of benzene (C_6H_6)

throughput experiments and simulations, which are highly prohibitive in terms of cost and manpower.

An alternate approach that has emerged recently is the data-driven approach. Here, the data is used to first identify the model and then fit the parameters of the model. Data-driven models are not based on physical theories and hence are occasionally termed as "black-box" models. It is interesting to note that although, data-driven models such as machine learning was first proposed at the same time as MC and MD simulations in 1950s, it has started finding wide-spread applications in materials engineering only for the past two decades. The inertia to not accept data-driven models, despite their fast, accurate, and efficient ability to learn patterns from data, could be attributed to their black-box nature. In other words, the data-driven models cannot be explained using known physics, they can only be tested for unknown scenarios. However, the advances in machine learning coupled with the availability of large-scale data on materials have shown the potential of data-driven approaches for materials discovery. In addition, the development of explainable machine learning algorithms, which allows the interpretation of black-box models, has allowed domain experts to interpret the black-box models. This allows the interpretation of the features "learned" by the model, thereby, giving insights into the inner workings of the models. Overall, data-driven approaches have shown significant potential to

accelerate materials discovery and reduce the discovery-to-deployment period from 20 years to 10 years or even lesser.

1.3 Introduction to Machine Learning

Machine learning (ML) refers to the branch of study which focuses on developing algorithms that "learns" the hidden patterns in the data. In contrast to physics-based models, ML uses the data for both model development and model training. Further, it improves the model in a recursive fashion using a predictor-corrector approach without being explicitly programmed to do the specific task. As such, large amounts of data is required for the ML models to learn the patterns reasonably–the more the data, the better the ML model is. ML has already been widely used in our day-to-day life for several applications such as face recognition, email spam detection, personal assistants, automated chat-bots, and fraud detection. To achieve these tasks, ML uses different classes of algorithms as detailed below.

Algorithms in ML can be broadly classified into supervised, unsupervised, and reinforcement learning. Supervised learning refers to those which learns the function that maps a set of input-output data. The examples of this approach include predicting the Young's modulus or density of an alloy based on the composition and processing or classifying a set of materials into conductor or insulator. It may be noticed in the first task the output Young's modulus can take continuous values as a function of the composition and processing and hence, is known as regression. Whereas in the second task, the output can either be conductor or insulator, and hence is a classification task. Note that the classification problems can be multi-class as well having more than two classes, for example, conductor, insulator, superconductor, and semi-conductor. The crucial aspect in supervised learning is the availability of a labeled dataset on which the model can be trained. The accuracy of the model depends highly on the accuracy of the dataset among other factors. Some commonly used supervised models are linear and polynomial regressions, logistic regression, decision trees, random forest (RF), XGBoost, support vector (SVR), neural network (NN), and Gaussian process regression (GPR).

In unsupervised learning, the algorithm tries to find out patterns from the features of the data. In this case, there is no labeled training set that is used. Some of the main approaches in unsupervised learning include clustering and anomaly detection. Clustering refers the automated grouping of materials based on their similarity to each other based on the features provided. Clustering may be used to remove an outlier in the data, or to identify subgroups in the data. Some of the unsupervised models include k-means, DBSCAN, OPTICS (inspired from DBSCAN), t-SNE, and principal component analysis (PCA).

Reinforcement learning, although holds a great potential, is relatively less explored in materials discovery. Reinforcement learning relies on a carrot-and-stick policy where an agent is trained to take actions that maximizes the cumulative reward. Thus, reinforcement learning tries to combine the existing knowledge and exploration in

a judicious fashion to maximize the reward. Reinforcement learning can be used to identify optimal process parameters for material synthesis and characterization, and also to explore novel materials with superior properties such as room temperature superconductors or ultra-stable glasses. Some of the reinforcement learning algorithms include Q-learning, and State-action-reward-state-action (SARSA). In this book, we will be primarily focusing only on supervised and unsupervised algorithms. These algorithms are discussed in detail in Part II. Reinforcement learning is briefly outlined in Sect. 8.7.

1.4 Machine Learning for Materials Discovery

Machine learning has found applications in accelerating the discovery of a variety of materials as well as to gain deep insights into the material response [2–6]. Here, we briefly review some of the applications where ML has successfully solved some of the open problems or has outperformed classical approaches. These applications are discussed in detail in Part III.

1.4.1 Property Prediction

One of the most commonly used application of ML is property prediction. This is a major problem for almost all materials such as alloys, ceramics, glasses, polymers, and nanomaterials as the property of a material can be a complex, non-convex function of composition, structure, and processing [3, 7–14]. For some properties such as hardness, it can also be a function of the testing method and testing parameters [15]. To predict material properties, first a clean dataset of input features and output property of interest need to be prepared. Note that the input features can be simple chemical composition, or more complex features such as the periodic table based descriptors. The input features can also be a combination of multiple features engineered using additional unsupervised ML techniques. Once a clean dataset is prepared, supervised ML is used to train models that can predict the property of interest.

Figure 1.4 shows the predicted values of density, Young's modulus, Vicker's hardness, and shear modulus of oxide glasses with respect to the experimental values [16]. The dataset consists of 50,000 oxide glasses with multiple components. We observe that the the predicted values for this large dataset exhibit a good agreement with respect to the experimental values for all the properties. In addition, the 95% confidence interval of the error histogram shown in the inset confirms that the predictions indeed exhibit a very low error in comparison to the range of values considered. Similar approaches have been widely used for the prediction of properties of several

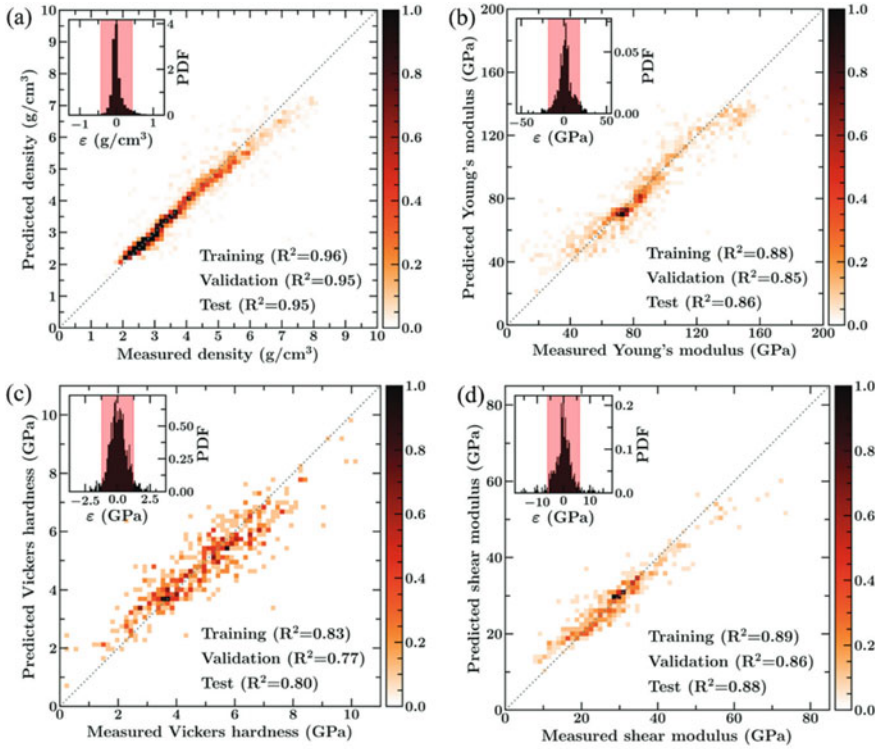

Fig. 1.4 Predicted values of **a** density, **b** Young's modulus, **c** Vicker's hardness, and **d** shear modulus of oxide glasses with respect to the experimental values. The R^2 values of training, validation, and test are shown. The inset shows the histogram of error in the prediction along with the 95% confidence interval

materials including ceramics, metal alloys, metallic glasses, 2D materials, polymers, and even proteins.

1.4.2 Materials Discovery

While property prediction allows one to explore the properties of hitherto unknown composition, it necessarily does not directly provide a recipe of new materials. Materials discovery a more challenging problem having constraints on multiple properties and components. For instance, a desired alloy for automotive applications should be light-weight, hard, strong, tough, ductile, and easily weldable. Many of these properties are conflicting. Effectively, this problem translates to solving the inverse of property prediction. Here, we need to predict the candidate composition and processing parameters corresponding to a target property. To this extent, surrogate model