JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

Doctoral Dissertation

AI development for capturing dynamic phenomena in materials: Insights and innovation through Deep learning model design and application

---

Author:
Tien-Sinh VU

Supervisor:
Professor Hieu-Chi DAM

A thesis submitted in fulfillment of the requirements for the degree of Knowledge science

in the

Dam Laboratory
School of Knowledge Science

Jan, 2025

Declaration of Authorship

I, Tien-Sinh VU, declare that this thesis titled, "AI development for capturing dynamic phenomena in materials: Insights and innovation through Deep learning model design and application" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Once you have read a book you care about, some part of it is always with you."

Louis L'Amour

JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

Abstract

Co-Creative Intelligence research area
School of Knowledge Science

Knowledge science

AI development for capturing dynamic phenomena in materials: Insights and innovation through Deep learning model design and application

by Tien-Sinh VU

This dissertation presents a domain-enriched deep learning framework for representation learning in materials science, addressing the challenges of capturing the complexity of dynamic, multidimensional material data where traditional descriptors are often insufficient. By embedding materials science knowledge within deep learning models, this research advances representation learning to support both predictive accuracy and scientific insight. The framework is applied to two key scenarios. First, in an unsupervised setting, it learns representations to reconstruct material images, capturing hidden structures and evolving patterns within the data and enabling discovery of material dynamic behaviors. Second, in a supervised learning context, it develops representations to predict material properties achieving both high accuracy and interpretability about structure-property relationship. This work highlights the impact of domain-guided representation learning, bridging deep learning with scientific principles to advance material discovery. Through case studies, it demonstrates that domain-enriched deep learning is not merely predictive but instrumental in generating insights, offering a versatile approach that strengthens the role of data-driven models in materials science innovation.

Keywords: Materials discovery, Data-driven approach, Deep Learning, Physics-informed, Materials Property, Materials Imaging

Acknowledgements

I would like to express my heartfelt appreciation to Professor Dam Hieu Chi for giving me the opportunity to engage in scientific research and experience the satisfaction that comes with it. Collaborating with him has taught me important lessons on professionalism, clarity of thought, devotion to science, and especially enjoying the journey.

I would also like to extend my gratitude to my lab colleagues, specifically M.S. D.D Anh, M.S. B. Adam, Dr. H.M. Quyet, and Dr. N.D. Nguyen, for their unwavering support and insightful discussions that have aided me in completing my work.

Lastly, I would like to dedicate this thesis to my family, my wife Ngoc Huyen, and my son Duc Tri, whose constant encouragement and guidance have been indispensable in my journey. I will always cherish their assistance.

Contents

List of Figures

List of Tables

List of Abbreviations

MAE	**M**ean **A**bsolute **E**rror
GA	**G**lobal **A**ttention
DE	**D**istance **E**mbedding
AE	**A**ngle **E**mbedding
CXDI	**C**oherent **X**-ray **D**iffraction **I**maging
RB	**R**efinement **B**lock
TB	**T**emporal **B**lock

List of Symbols

$\mathcal{D}$	Dataset
x	Representation variable of a data instance
y	Target variable of a data instance
$\hat{\mathbf{y}}$	Prediction output of model
σ	Activation function
$\mathcal{L}$	Loss function
$\mathcal{F}$	Fourier Transformation

For/Dedicated to/To my…

Chapter 1

Introduction

1.1 Modeling dynamics and complexity in Material Science

Since the inception of civilization, humankind's advancement has been characterized by an evolving relationship with materials. Each significant discovery—stone, bronze, iron, or silicon—has unlocked new possibilities, reshaping societies and propelling technological advancements. However, as our comprehension of these materials has deepened, so has the acknowledgment of their complexities. Materials are not merely inert or straightforward; they exhibit a dynamic and multifaceted nature that poses challenges for scientists and engineers (Agrawal and Choudhary, 2016; Ramprasad et al., 2017; Butler et al., 2018; Siriwardane et al., 2022). Materials frequently demonstrate intricate behaviors in response to environmental conditions, adapting and evolving in ways that reveal their fundamental characteristics and imply potential applications across diverse fields. These responses can range from molecular alterations to observable changes in structure or properties, providing critical insights into the nature of a material. Some transformations occur at the subatomic level, involving subtle forces and bonding interactions, while others manifest on a macroscopic scale, where their effects are directly observable and impactful. External influences such as stress, heat, or minor environmental changes can initiate a cascade of reactions within a material, underscoring its adaptability and versatility.

This complex interplay of responses presents both opportunities and challenges. The diversity and unpredictability of material behaviors necessitate detailed observation and innovative analytical frameworks. As researchers endeavor to elucidate these complexities, they often employ advanced representation techniques that impose order on the dynamic interactions observed. Such methodologies must be designed to capture the full spectrum of a material's behavior, advancing understanding of its underlying processes while establishing structured pathways for reasoning, analysis, and model development. In summary, achieving a comprehensive understanding of materials' complexity is pivotal to fostering innovation and revealing the vast potential inherent in each response and interaction. This continuum from understanding to application is essential to the future of materials science, where the capacity to capture and represent dynamic phenomena will be instrumental in unlocking the rich possibilities of materials that shape our world.

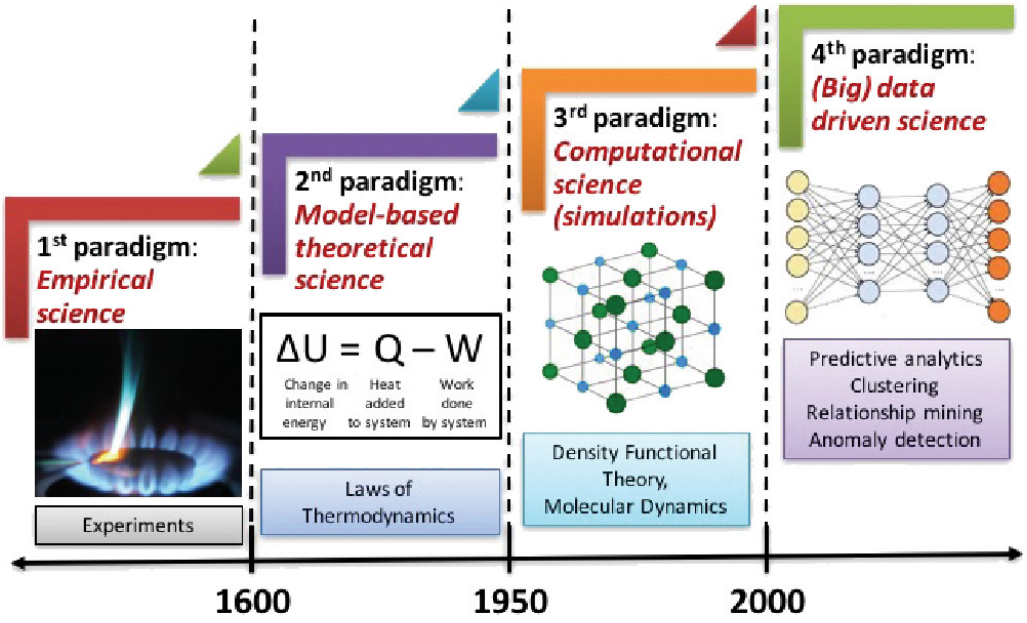


FIGURE 1.1: Illustration of four paradigms in materials science.

1.1.1 Material Informatics and Representation learning

Throughout human history, the quest for advanced materials has progressed with scientific and technological advancements. This journey is characterized by four critical paradigms that have significantly shaped the field of materials science. According to Agrawal (Agrawal and Choudhary, 2016), these paradigms—empirical science, theoretical science, computational science, and data-driven science—each offer unique approaches to discovering and understanding materials. Initially, knowledge in materials science was based on empirical observations and experiments, focusing on processes such as material extraction, purification, and processing. As mathematical concepts, particularly algebra and calculus, advanced, theoretical models emerged. These models, grounded in mathematical equations like the laws of thermodynamics, provided a basis for predicting material behavior. The subsequent significant development was the rise of computational science, which allowed for the simulation of complex material phenomena using advanced algorithms and methods, such as density functional theory and molecular dynamics. Most recently, a data-driven approach has emerged as the fourth paradigm in response to the data generated by experiments and simulations. This approach, known as materials informatics, aims to integrate theory, experimentation, and computational insights, thereby accelerating the discovery and understanding of materials through data-driven analysis. The significantly faster decision-making process is the primary advantage of data-driven approaches over experimental methods and materials simulations. Data-driven models typically evaluate a given material data within seconds, whereas simulations may take hours to days, and experiments can span days to months. Additionally, compared to the human summarization of physical rules, data-driven approaches excel in managing massive datasets and extracting highly complex and nonlinear relationships between multiple inputs and outputs.

A structured data mining process is necessary to tap into the potential of data-driven approaches in materials informatics (Fig 1.2). This process typically consists

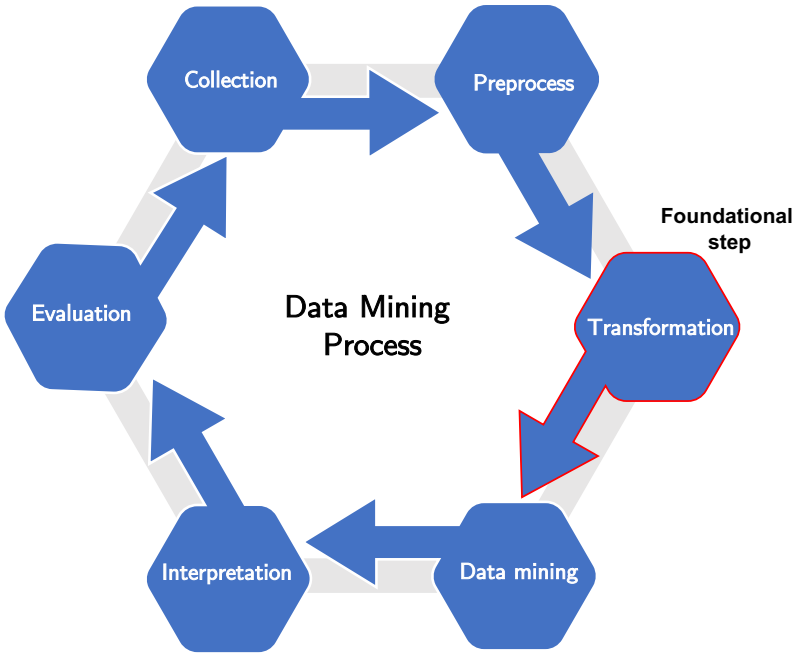


FIGURE 1.2: Schematic representation of the data mining process. The sequence begins with data collection, followed by preprocessing to clean and organize the raw data. The transformation step then converts data into suitable formats or representations for analysis. Next, data mining techniques are applied to extract patterns and insights. Finally, through interpretation, these findings are translated into actionable knowledge, contributing to advancements in materials science.

of several key stages: data collection, preprocessing, transformation, mining, interpretation, and knowledge generation. In these stages, transformation is a preliminary step in which raw and heterogeneous data are standardized to a machine-readable unified format that can be analyzed and interpreted effectively. Proper transformation ensures that the diverse and complex nature of data on materials is adequately treated, allowing for more accurate and meaningful data mining outcomes. Emphasizing transformation aids researchers in better handling the multi-scale and multi-dimensional aspects of materials data, providing the base for successful representation learning and subsequent analysis.

A successful application of materials informatics relies heavily on the effectiveness of representation in the transformation process (Yang et al., 2019). Accurately representing a material beyond merely cataloging its observable properties requires a deep understanding of its multi-scale complexities and the ability to encode them in a structured, machine-readable format that faithfully reflects real-world behaviors. Unlike simple datasets, material data are inherently diverse and multidimensional, encompassing atomic structures, phase compositions, chemical environments, and property measurements across various scales. Crafting representations to handle this diversity has often involved handcrafted features—carefully designed descriptors rooted in specific material attributes or known physical principles. However, these handcrafted features are typically limited in scope, capturing only narrow aspects of the material data and often requiring domain-specific expertise to design. Moreover, the challenge of representation design is exacerbated by the fact

that materials datasets are frequently specific, sparse, or fragmented, typically generated from experiments tailored to particular material classes or conditions. This sparsity obstructs the creation of generalized models that can reliably extend across diverse material systems. Additionally, the dynamics of materials, including their mechanical, thermal, and transport properties, play a crucial role in understanding their behavior under various conditions. Effective representations can significantly aid in capturing these dynamic phenomena, allowing for more accurate predictions of material performance over time. Furthermore, since fundamental physicochemical principles govern materials data, effective representations must adhere to these foundational laws, preserving essential relationships between properties and their underlying physics. Traditional numerical descriptors often fall short of this requirement, as they lack the flexibility necessary to encapsulate the intricate, nonlinear interactions present in natural materials, particularly when considering their dynamic responses (Hirn, Mallat, and Poilvert, 2017).

The limitations of handcrafted and fixed-feature representations underscore the urgent need for representation learning. This innovative approach empowers models to autonomously learn complex, data-driven representations without dependence on predefined features. In contrast to traditional descriptors, which may only capture limited dimensions of material behavior, representation learning facilitates the extraction of nuanced patterns and interactions within multidimensional material data, revealing relationships that would be challenging to identify through manually crafted features (Ramakrishnan et al., 2014; Himanen et al., 2019; Zhao et al., 2023). By learning representations directly from data, this approach effectively accommodates the complexity and diversity inherent in materials, including nonlinear dependencies, hierarchical structures, and varying scales of interaction, thereby offering a more comprehensive understanding of material behavior (Rupp et al., 2012). A notable example of the potential inherent in representation learning is the Materials Project. This initiative employs machine learning models to predict material properties by directly learning representations from large-scale material data. By establishing a continually expanding database of computed material properties, the project has enabled researchers to generate predictions for various materials, thereby advancing the discovery process and informing experimental efforts. The learned representations enhance generalization across different classes of materials, facilitating the prediction of novel compounds with desirable properties.

1.1.2 Challenges in Representation learning for Material science

As the field of materials science embraces data-driven approaches, representation learning faces distinctive challenges tied to the complexity and diversity of materials data. Unlike other domains, where data are often uniform and abundant, materials data present a spectrum of unique difficulties that require specialized approaches. Materials are governed by intricate physical, chemical, and structural properties that do not easily translate into standardized formats or simplified features. Instead, capturing the nuanced behaviors and interactions within materials requires representations that can span multiple scales, from atomic and molecular interactions to macroscopic properties. This inherent complexity requires representation learning models that can accommodate diverse and interconnected data while still delivering insights that are both accurate and interpretable.

An additional challenge lies in the nature of the data itself, which is often limited, sparse, or specialized. Unlike fields with extensive labeled datasets, materials science frequently depends on costly and time-consuming experimental data or

computational simulations that can introduce their own limitations and uncertainties. This scarcity poses a barrier for machine learning models, which typically benefit from large, well-labeled datasets for training. While techniques such as transfer learning and few-shot learning offer promising solutions, they require careful adaptation to fit the specific constraints of materials science, where data availability and diversity are far from uniform (Xie and Grossman, 2018).

One significant challenge lies in the complexity of materials data itself, which often includes high-dimensional features that are difficult to simplify without losing critical information. For instance, a single material's behavior may depend on electronic structures, atomic configurations, and environmental conditions, all of which interact in nonlinear and sometimes unpredictable ways. Traditional machine learning models can struggle to encapsulate these multifaceted relationships, leading to potential inaccuracies in predictions or generalizations. Furthermore, ensuring that these representations respect the physicochemical principles governing materials requires careful design, as oversimplifying these relationships can result in models that fail to capture essential properties accurately

In summary, representation learning in material informatics encounters substantial challenges due to the inherent complexity of materials data, limited availability of high-quality datasets and the need to respect physicochemical principles. Overcoming these obstacles is essential for building models that are not only robust and accurate but also capable of advancing our understanding and discovery of new materials.

1.2 Deep learning and data-driven representations

Deep learning (DL) has emerged as a transformative approach in data-driven science, offering unprecedented capability in learning complex, non-linear representations across high-dimensional data spaces. Unlike traditional methods, which typically rely on fixed mathematical functions or manually designed features, DL employs layered neural networks capable of autonomously extracting patterns and relationships from raw data. This structure allows DL models to learn representations directly from the data, adapting to its inherent complexities without requiring domain-specific descriptors. As a result, DL methods excel in tasks that involve intricate relationships and large feature spaces, where traditional approaches may struggle to capture the full depth of interactions (Tran et al., 2018).

The architecture of deep learning networks, often consisting of multiple hidden layers, enables these models to perform hierarchical learning. Lower layers capture simple features, while deeper layers progressively build on these to recognize complex patterns and higher-order relationships. This capacity for hierarchical abstraction allows deep learning to learn representations that not only fit the data but also generalize across diverse datasets and contexts. Additionally, advances in training techniques, optimization algorithms, and computational power have accelerated the adoption of DL, making it feasible to tackle high-dimensional data in fields ranging from natural language processing to computer vision (Krizhevsky, Sutskever, and Hinton, 2017, Vaswani et al., 2017).

1.2.1 Evolution of Deep learning architectures for enhanced data representation

The design of deep learning architectures has evolved from simple feedforward networks to more complex structures, which are tailored for specific data types and tasks. This reflects an ongoing process of enhancing the models' capabilities of learning rich, hierarchical representations of data, hence significantly improving performance across a wide range of applications. Basically, understanding this evolution is essential in highlighting the architectural innovations that have so far enabled deep learning models to capture increasingly complex patterns and relationships within data.

Early neural networks were relatively shallow-only a few layers with minimal capacity to model complex functions. Such networks struggled with learning representations that would generalize well to previously unseen data, primarily due to their shallow depth and inability to capture hierarchical features. The introduction of much deeper architectures overcame these limitations by stacking many layers together so that the models could learn representations incrementally at higher levels of abstraction. Each successive layer captures more complex features by building upon the representations learned in previous layers, facilitating a hierarchical learning process (Bengio, 2009).

Considering the development of deep learning architecture, one of the most important milestones was the convolutional neural network, which was designed for working with grid data, such as images. CNNs introduced convolutional layers using the idea of spatial hierarchies in data to capture local patterns and build them into global features effectively. Early models like LeNet-5 demonstrated the efficacy of CNNs for image recognition tasks (Lecun et al., 1998). Meanwhile, further architectures like AlexNet, VGG, and ResNet sharply increased depth. They introduced novel ideas concerning deeper convolutional layers and residual connections that resolved issues such as vanishing gradients (Krizhevsky, Sutskever, and Hinton, 2017,Simonyan and Zisserman, 2015, He et al., 2015). These advancements allowed CNN to learn more expressive representations and substantially improved performance for image classification and related tasks.

Recurrent neural networks were developed for sequential and temporal data with the aim of modeling dependencies across time. However, traditional RNNs suffered from the inability to learn long-term dependencies due to problems such as exploding and vanishing gradients. This limitation has driven the development of more complex architectures including Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) that introduce various forms of gated information flow and gradient propagation through time (Hochreiter and Schmidhuber, 1997, Chung et al., 2014). These architectures allowed models to learn representations that take into consideration both the short- and long-term patterns in sequential data, quite effectively model language and speech, among others.

Introducing attention mechanisms marked another significant milestone in deep learning architecture design. Attention allows models to focus on specific input parts when generating representations, improving the handling of dependencies and relationships within data. The Transformer architecture, which relies entirely on attention mechanisms without recurrent layers, revolutionized natural language processing by enabling models to capture global dependencies more effectively (Vaswani et al., 2017). Transformers learned highly expressive representations and formed the basis of recent models like BERT and GPT, which broke new records on many benchmarks across diverse language tasks.

These different architectural innovations together show how deep learning design is evolving to create models that can learn representations of ever-increasing sophistication. By overcoming the limitations of their predecessors and embedding mechanisms suited for different data structures and tasks, deep learning architectures have considerably widened the envelope of what is possible with data-driven representation learning. This evolution underlines how important architecture design is in deep learning. However, it also sets the stage for exploring how these principles can be put to work for the complex data in materials science.

1.2.2 Integrating domain knowledge into Deep learning architectures

The effective integration of domain knowledge has become a crucial strategy for enhancing the performance, interpretability, and generalization of deep learning models at large, especially in complex scientific domains. As much as deep learning models are excellent at learning features from the data themselves, knowledge of the specific expertise within the domain guides such learning processes to enable not only the model fitting of the data but also adherence to the underlying physical laws and principles. This synergy in data-driven learning and design informed by the domain leads to more robust and reliable models, especially necessary in areas where data may be scarce or noisy. (Metzler et al., 2018).

The main motives for incorporating domain knowledge include constraining the model's hypothesis space, a process that effectively reduces overfitting and improves generalization on unseen data. By embedding known relationships, symmetries, or invariances into the architecture, models focus their learning on aspects of the data that are truly novel or unexplained by existing theories. This is particularly useful in scientific fields where one has specific behaviors constrained by known laws, like conservation laws in physics or chemical bonding rules in chemistry.(Raissi, Perdikaris, and Karniadakis, 2019).

There are several ways in which domain knowledge can be incorporated into the framework of deep learning models. The common approaches involve introducing an explicitly defined loss function that penalizes deviations from known physical laws or constraints. Classic examples include physics-informed neural networks (PINNs), where physical laws represented by differential equations are directly used under appropriate forms in the loss function and thus guide the model toward solutions that satisfy those equations (Raissi, Perdikaris, and Karniadakis, 2019). Another popular strategy is to architect the network structure from the domain problem. Graph neural networks, for example, are particularly tailored to model entities and their interactions; hence, the graph-structured data like molecules or crystalline structures naturally model the interactions in accord with the knowledge of the domain at hand (Gilmer et al., 2017).

Embedding domain knowledge can also take the form of input feature engineering, whereby raw data are transformed using domain-specific operations before being fed into the model. This would relieve relevant patterns and relations critical to the task at hand. In materials science, for example, features like radial distribution functions or electronic density of states can be computed from raw structural data to provide the model with rich, physically meaningful inputs (Butler et al., 2018). Additionally, transfer learning techniques can leverage pretrained models from related domains, incorporating learned representations that already encode domain-relevant information.

While these methods have significant advantages, their integration with deep learning models further introduces a number of challenges. Basically, domain and

machine learning techniques are blended, which requires a deep knowledge of both the domain and machine learning techniques to perform effectively without introducing biases or constraints that may affect the model's ability to learn from data. Moreover, too rigid incorporation of domain knowledge can prevent the model from discovering new patterns that are outside conventional theories but still important. Guidance from domain knowledge must be balanced with flexibility in deep learning, an act that requires careful consideration of the two variables mentioned above (Xie and Grossman, 2018).

In short, the incorporation of domain knowledge into deep learning design represents a powerful approach toward improving model performance and reliability in complex scientific fields. By embedding expertise directly into models through customized architectures, loss functions, or feature engineering, we can obtain deep learning systems that not only learn from data but also respect the fundamental principles of the domain. This fusion of data-driven and knowledge-driven approaches clears the ground for further advances not only in places like materials science, where the understanding of behaviors should necessarily respect both rich data representations and adherence to physical laws.

1.3 The rising need for synergistic integration of Deep learning and Materials science

The convergence of deep learning and materials science shows a new era of co-creation, where domain knowledge and advanced computational methods reinforce mutual advances in both fields. Deep learning models become more capable of capturing complexities inherent in material data when informed by the rich, nuanced understanding of materials science. This can promote the learning of meaningful, accurate, and physically interpretable representations in such models.

Domain knowledge from materials science infused into deep learning architecture provides the models with a foundation of scientific principles. The physics integration could be explicitly expressed in network architecture, such as graph neural networks for crystalline materials or physics constraints baked into the learning process. Aligning model design to material behaviors and properties can thus guide the learning process to focus on relevant features, thereby improving both performance and interpretability.

Concurrently, Deep learning gives valuable feedback to the area of material science by uncovering hidden patterns and relationships within complex datasets. Sophisticated models may reveal insights into material behaviors that could not have been evident using traditional approaches in experiments or theory. For instance, deep learning can determine the factors that significantly affect material properties or predict new materials with specific properties, hence hastening the discovery process. This feedback loop enriches the domain knowledge of materials science, highlighting areas for further investigation and potential innovation.

The collaborative cycle between deep learning and materials science also supports a range of data mining tasks beyond predictive modeling. Deep learning's hierarchical feature extraction provides a powerful method for clustering materials based on intrinsic properties, anomaly detection indicative of potentially novel phenomena, and unsupervised learning where labeled data are scarce. These capabilities are beneficial within materials science, where the datasets can be heterogeneous and complex.

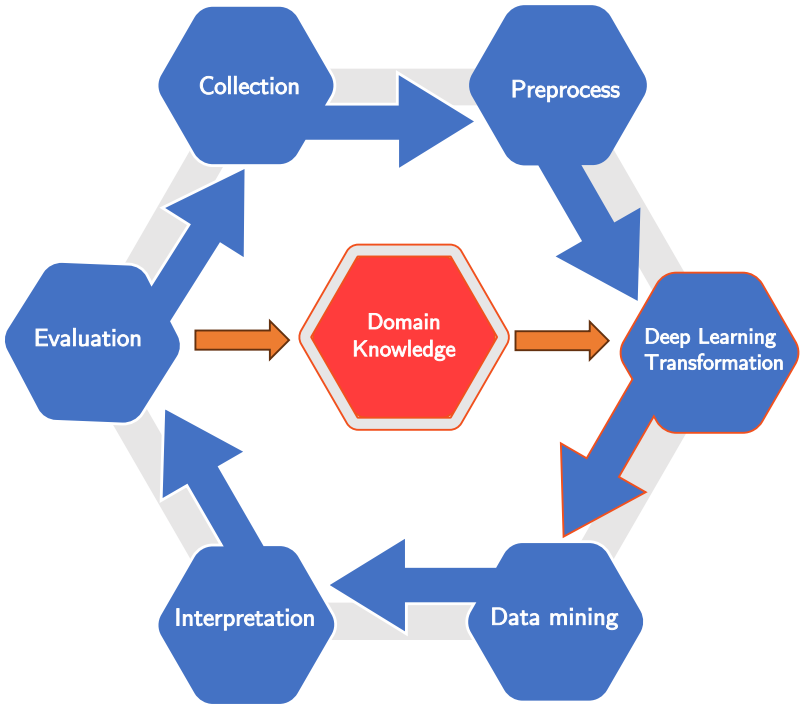


FIGURE 1.3: Synergistic Integration of Deep Learning and Domain Knowledge in Materials Science. The diagram illustrates the cyclical process of knowledge enhancement through deep learning in materials science. The enriched knowledge enhances the existing domain expertise, which then informs and refines subsequent deep learning models (DL), closing the loop. This iterative process fosters continuous advancement in materials science by coupling domain knowledge with deep learning capabilities.

This synergistic relationship forms a continuous cycle wherein domain knowledge leads and trains deep learning models, and the models, in return, provide new insights to expand and refine the knowledge base as shown in Fig 1.3. Deep learning serves as a transformative intermediary, translating complex material data into actionable information and hypotheses. The acceleration this creates for materials science allows quicker exploration of material spaces and innovation. Ultimately, co-creation by deep learning and materials science stands out as a bright example of what interdisciplinary collaboration is all about. Suppose deep learning strengths outstanding work with complex data and representation learning be combined with profound expertise present in materials science. In that case, one can push both deep learning and materials science beyond known limits. This would mean increased predictive performance and a fundamental understanding of material phenomena , enabling groundbreaking discoveries and technological developments.

1.4 Research objectives

In the previous section, I discussed the challenges encountered in designing effective representations for deep learning applications in materials science, highlighting the importance of selecting suitable approaches to address the complexities of this field. Representation learning in materials science is inherently challenging, given the diverse, high-dimensional nature of materials data and the need to capture nuanced

properties accurately. Developing deep learning models that can effectively learn from these data requires a careful balance between complexity and interpretability, and a robust understanding of the underlying material mechanisms is essential. While deep learning has shown significant promise in extracting patterns and reducing the complexity of materials data, several key limitations still impede its broader applicability:

- **Domain Knowledge Incorporation into Framework Design:** Designing representations that leverage domain-specific knowledge is essential for capturing the complex, multiscale interactions that govern material properties. Deep learning models must integrate principles from materials science to ensure representations are not only accurate but also interpretable and grounded in scientific understanding.
- **Prediction Interpretability:** Even when deep learning models achieve high predictive accuracy, accurately evaluating and communicating prediction interpretability remains a major challenge, especially in contexts requiring scientific insights.

To address these limitations, I have developed deep learning frameworks designed with embedded domain-specific knowledge to generate meaningful representations of materials that advance various tasks in materials science. By developing frameworks rooted in scientific principles, this work aims to achieve not only high predictive accuracy but also to provide scientifically valuable insights into material behaviors, thereby enhancing both interpretability and discovery potential in the field.

To demonstrate the applications of this framework, the thesis presents two major scenarios:

- **Unsupervised Representation Learning for Unknown Targets:** Developing representations from unlabeled data to discover patterns or classify materials based on implicit relationships.
- **Supervised Representation Learning for Known Targets:** Learning material representations with labeled data, targeting specific material properties or behaviors.

These scenarios are further illustrated through two in-depth case studies, each serving as a focal point for the application of these methodologies. These studies offer a roadmap for the thesis and showcase the potential of the proposed framework to achieve interpretability, accuracy, and discovery within the context of materials science. Our proposed methods are highly effective, as demonstrated by utilizing to uncover underlying mechanisms in materials (Chapter 3) and creating of an interpretable deep learning framework for predicting materials property (Chapter 4) and

Chapter 2

Representation Learning and Design with Domain Knowledge in Deep Learning

The following chapter explores the basic design methodologies for the application of deep learning to scientific research, with an emphasis on constructing domain-enriched representations for modeling complex phenomena. We detail the DL paradigm's essential ingredients, namely inputs, architectures, outputs, loss functions, and learning algorithms. Then, we will describe how each one of those elements can be tailored in an application-specific fashion given scientific requirements. Consequentially, the framework not only progresses the accuracy of material property predictions but furthermore extends the general scientific knowledge base in such a manner as to enable new material discoveries and unforeseen explanations of complex material phenomena.

Initially, the chapter provides a comprehensive overview of the foundational principles of deep learning. It introduces essential concepts such as neural networks, activation functions, and the architecture of deep learning models. The discussion also includes insights into the evolution of deep learning, highlighting significant milestones and breakthroughs that have shaped the field. This foundational knowledge sets the stage for more complex topics and advanced concepts as progress through the material science.

Consequently, this chapter offers a review of methods for the design of inputs and outputs compatible with scientific datasets, often characterized by high-dimensional, structured, or sparse data. This means addressing the challenges of data preprocessing and transformation and the appropriate selection of learning methods to enhance the generalization and robustness of models. When these components are aligned with domain-specific requirements, the scientific discovery facilitated by DL models allows researchers to interpret results with greater confidence and precision-meaningful patterns and relationships across varied scales and contexts.

2.1 Fundamentals of Deep Learning

Deep learning enables computational models composed of multiple processing layers to learn data representations at various levels of abstraction. In recent years, deep learning models have achieved remarkable success in numerous fields, including natural language processing, object detection and tracking (Vaswani et al., 2017; Ronneberger, Fischer, and Brox, 2015), image classification (Bertasius, Wang, and Torresani, 2021; Krizhevsky, Sutskever, and Hinton, 2017), and material property prediction (Chen et al., 2019; Xie and Grossman, 2018).

The success of deep learning (DL) relies on several critical factors, with four primary components—inputs, architectures, outputs, and loss functions—playing essential roles in maximizing the potential of DL models. Each component contributes uniquely to the model's ability to learn, generalize, and perform effectively across various tasks. Understanding these fundamental components is vital for designing and implementing successful deep learning models, especially in specialized fields such as materials science informatics. The following sections will explore these components, providing fundamental concepts and general applications.

2.1.1 Deep Learning models input

Deep learning models can process various types of data, each requiring a specific format to be effectively utilized by the network. Common data formats include:

- **Vectors** ($\mathbf{x} \in \mathbb{R}^n$): Represented as one-dimensional arrays, vectors are typically used for tabular data where each element corresponds to a distinct feature.
- **Matrices** ($\mathbf{X} \in \mathbb{R}^{m \times n}$): Two-dimensional arrays used for image data, where m and n correspond to the height and width of the image, respectively.
- **Tensors** ($\mathcal{X} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_k}$): Multi-dimensional arrays that extend beyond matrices, suitable for volumetric data or sequences of images.
- **Sequences** ($\mathbf{X} \in \mathbb{R}^{T \times d}$): For time-series or sequence data, inputs are arranged in temporal order to capture dependencies over time.

$$\mathbf{X} = [\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(T)}] \tag{2.1}$$

- **Graphs** ($\mathcal{G} = (\mathcal{V}, \mathcal{E})$): Consist of nodes ($\mathcal{V}$) and edges ($\mathcal{E}$), ideal for representing molecular structures or material lattices.

$$\mathcal{V} = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^n \tag{2.2}$$

$$\mathcal{E} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n} \tag{2.3}$$

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \tag{2.4}$$

Moreover, normalization is essential for ensuring that input features contribute equally to the learning process, preventing biases due to differing scales. Common normalization techniques include:

- **Min-Max Scaling**: Transforms features to a fixed range, typically $[0, 1]$.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2.5}$$

- **Z-Score Normalization**: Centers the data around zero with a standard deviation of one.

$$x' = \frac{x - \mu}{\sigma} \tag{2.6}$$

- **Batch Normalization**: Applied within neural network layers to stabilize and accelerate training by normalizing the inputs of each mini-batch.

$$\mathrm{BN}(x) = \gamma \left(\frac{x - \mu_{\text{batch}}}{\sqrt{\sigma^2_{\text{batch}} + \epsilon}} \right) + \beta \tag{2.7}$$

where μ_{batch} and σ^2_{batch} are the mean and variance of the batch, and *gamma* and β are learnable parameters.

2.1.2 Deep Learning network architectures

Deep learning architectures are fundamental to various applications, allowing models to learn intricate patterns and representations from data. Understanding these foundational architectures is crucial for developing effective DL models tailored to specific tasks in materials science informatics. This subsection will explore the basic architectures, including the perceptron, multilayer perceptron (MLP), fully connected networks, and activation functions. We will highlight their structures, functionalities, and roles in deep learning.

Core components of neural networks

Perceptron

The perceptron is the simplest neural network unit type and is the foundational building block for more complex architectures. Introduced by Frank Rosenblatt, 1958, the perceptron represents a single neuron with a binary output.

$$\hat{y} = \begin{cases} 1 & \text{if } \mathbf{w} \cdot \mathbf{x} + b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ is the input feature vector and $\mathbf{w} = [w_1, w_2, \ldots, w_n]^T$ is the weight vector. b is the bias term and $\hat{y}$ is the binary output.

The perceptron applies a linear combination of inputs and weights and a step activation function to produce a binary classification.

Multilayer Perceptron (MLP)

The multilayer perceptron (MLP) is a type of feedforward artificial neural network that consists of several layers of neurons. These layers include an input layer, one or more hidden layers, and an output layer. Each layer is fully connected to the next, and non-linear activation functions are applied to introduce complexity, allowing the network to learn non-linear relationships.

$$\mathbf{z}^{(l)} = \mathbf{W}^{(l)} \mathbf{a}^{(l-1)} + \mathbf{b}^{(l)} \tag{2.9}$$

$$\mathbf{a}^{(l)} = \sigma(\mathbf{z}^{(l)}) \tag{2.10}$$

Where l denotes the layer index, $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ are the weight matrix and bias vector for layer l. $\mathbf{a}^{(l-1)}$ is the activation from the previous layer and $\sigma(\cdot)$ is the activation function.

An MLP typically consists of an input layer, one or more hidden layers, and an output layer. The introduction of hidden layers allows the network to model complex, non-linear relationships between inputs and outputs through hierarchical feature extraction. While MLP are powerful, they can become computationally intensive and prone to overfitting, especially with large input dimensions. Techniques such as regularization and dropout are often employed to mitigate these challenges.

Activation functions

In addition to neural architecture, activation function play a vital role in incorporating non-linearity into neural networks, which allows them to learn complex patterns and representations alongside the neural architecture. Various activation functions are widely employed in deep learning architectures:

- **Sigmoid Function** The sigmoid function maps input values to a range between 0 and 1, making it suitable for binary classification tasks. $\sigma(x) = \frac{1}{1+e^{-x}}$ However, sigmoid functions can suffer from vanishing gradients, hindering the training of deep networks.
- **Hyperbolic Tangent (tanh) Function** The tanh function maps inputs to a range between -1 and 1, providing zero-centered outputs which can help in faster convergence. $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ Similar to the sigmoid function, tanh can also experience vanishing gradient issues.
- **Rectified Linear Unit (ReLU)** ReLU is widely used due to its simplicity and effectiveness in mitigating the vanishing gradient problem (Nair and Hinton, 2010). It outputs the input directly if it is positive; otherwise, it outputs zero. $\text{ReLU}(x) = \max(0, x)$ ReLU accelerates convergence and allows for the training of deeper networks but can suffer from the "dying ReLU" problem where neurons become inactive.
- **Softmax Function** Softmax is typically used in the output layer for multi-class classification tasks (Bridle, 1990). It converts logits into probabilities that sum to one. $\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ where K is the number of classes.

Popular Architectures

The following contents builds upon foundational concepts above to explore more advanced deep learning architectures, including Convolutional Neural Networks (CNNs), Autoencoders, Attention Mechanisms, and Transformers. These architectures have significantly enhanced the capabilities of deep learning models, allowing them to manage complex data structures and achieve state-of-the-art performance across various applications, including materials science informatics.

Convolutional Neural Networks (CNNs) architecture

Convolutional Neural Networks (CNNs) have transformed the field of deep learning by effectively capturing spatial hierarchies and patterns within data (Lecun et al., 1998). Originally developed for image processing tasks, CNNs have been successfully adapted for various applications in other fields, including image/videos based

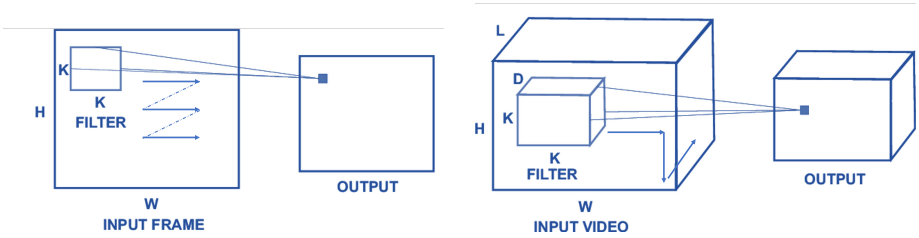


FIGURE 2.1: Illustration of 2D and 3D convolution operations for images and videos.

characterization, text classification, and property prediction. Figure 2.2 demonstrates the basic application of convolution operators on different kind of data. A typical CNN architecture consists of a series of convolutional layers, activation functions, pooling layers, and fully connected layers. Each component plays a specific role in extracting and processing features from the input data:

Convolutional layers Convolutional layers are the core building blocks of CNNs. They apply a set of learnable filters (kernels) to the input data to extract localized features such as edges, textures, and patterns.

$$\mathbf{Y}_{i,j}^{(k)} = \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{X}_{i+m,j+n} \cdot \mathbf{K}_{m,n}^{(k)} + b^{(k)} \tag{2.11}$$

where $\mathbf{X}$ is the input feature map, $\mathbf{K}^{(k)}$ is the k-th convolutional kernel of size $M \times N$, and $\mathbf{Y}^{(k)}$ is the output feature map after applying the k-th kernel. $b^{(k)}$ is the bias term for the k-th kernel and i, j denote the spatial position in the feature map.

The convolution operation enables the network to learn spatial hierarchies representation of data by detecting increasingly complex features in deeper layers.

Activation functions introduce non-linearity into the network, allowing CNNs to model complex relationships within the data. Common activation functions used in CNNs include ReLU, Leaky ReLU, or Swish (Ramachandran, Zoph, and Le, 2017), etc.

Pooling layers reduce the spatial dimensions of the feature maps, decreasing computational complexity and controlling overfitting. Common pooling operations include:

- **Max Pooling**:

$$\mathbf{Y}_{i,j}^{(k)} = \max \left\{ \mathbf{X}_{i,j}^{(k)}, \mathbf{X}_{i+1,j}^{(k)}, \ldots, \mathbf{X}_{i+m-1,j+n-1} \right\} \tag{2.12}$$

  Max pooling selects the maximum value within a defined window, preserving the most prominent features.

- **Average Pooling**:

$$\mathbf{Y}_{i,j}^{(k)} = \frac{1}{m \times n} \sum_{p=0}^{m-1} \sum_{q=0}^{n-1} \mathbf{X}_{i+p,j+q}^{(k)} \tag{2.13}$$

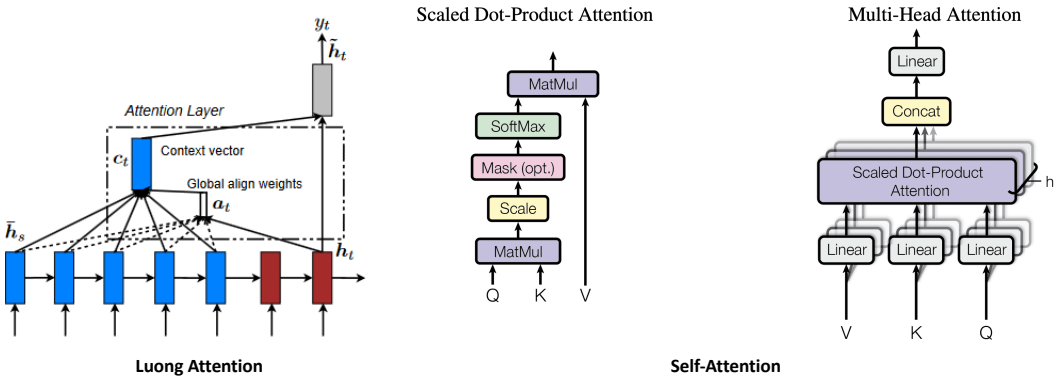


FIGURE 2.2: Illustration of two primary types of attention mechanism used in DL models.

Average pooling computes the average value within the window, providing a more generalized feature representation.

Fully Connected Layers After several convolutional and pooling layers, the high-level reasoning in the network is performed via fully connected layers. These layers connect every neuron in one layer to every neuron in the next layer, enabling the network to combine features and perform classification or regression.

$$\mathbf{y} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{2.14}$$

(2.15)

Attention Mechanisms and Transformers architecture

Attention mechanisms and Transformer architectures have significantly changed the field of deep learning by allowing models to better capture complex dependencies and contextual relationships within data. Originally developed to overcome challenges in sequence modeling tasks, these innovations have since found applications in various domains, including natural language processing (NLP), computer vision, and more. This subsection offers a thorough literature review of attention mechanisms and Transformers, explaining their underlying mechanisms, architectural components, and key applications.

The primary motivation behind attention mechanisms is to enable models to weigh the importance of different input elements dynamically. This approach contrasts with traditional sequence models, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), which process input data sequentially and often struggle with capturing long-term dependencies. There are three primary types of attention mechanisms used in natural language processing and machine learning: Bahdanau attention, Luong attention, and self-attention.

- **Bahdanau Attention (Additive Attention)**:

  Introduced by Bahdanau, Cho, and Bengio, 2014 for machine translation, additive attention computes attention scores using a feedforward neural network.

$$e_{ij} = \text{score}(h_i, s_j) = v_a^T \tanh(W_a h_i + U_a s_j) \tag{2.16}$$

where h_i is the encoder hidden state, s_j is the decoder hidden state, W_a and U_a are weight matrices, and v_a is a weight vector.

- **Luong Attention (Multiplicative Attention)**:

  Proposed by Luong, Pham, and Manning, 2015, multiplicative attention computes attention scores as the dot product between encoder and decoder hidden states.

$$e_{ij} = h_i^T W_a s_j \tag{2.17}$$

  This form is computationally efficient and forms the basis for later attention mechanisms.

- **Self-Attention**:

  Self-attention (Vaswani et al., 2017), or intra-attention, allows a sequence to interact with itself to compute a representation of the sequence. It is a fundamental component of Transformer architectures.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.18}$$

  where Q, K, and V are query, key, and value matrices, respectively, and d_k is the dimensionality of the keys.

Transformers, introduced by Vaswani et al., 2017 in the seminal paper "Attention is All You Need," leverage self-attention mechanisms to model dependencies irrespective of their distance in the input sequence. This architecture has since become the backbone of numerous state-of-the-art models in various domains. Models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) utilize Transformer architectures to achieve high performance in tasks such as question answering, text generation, and sentiment analysis. As research continues to tackle existing challenges and explore new applications, attention and Transformers are expected to remain at the forefront of deep learning advancements.

Encoder-Decoder architectures

Encoder-Decoder architectures are essential neural network models that play a crucial role in various deep learning applications, such as machine translation, image segmentation, and generative modeling. In the field of materials informatics, these models aid in tasks like predicting material properties, generating structures, and detecting anomalies by effectively capturing and reconstructing complex data representations. This subsection will explore the core components of Encoder-Decoder architectures, their operational mechanisms, and different variations and concepts associated with them. The bidirectional flow enables the model to learn meaningful representations that capture the essential features of the input data, facilitating accurate and interpretable predictions.

$$\text{Input} \xrightarrow{\text{Encoder}} \text{Latent z} \xrightarrow{\text{Decoder}} \text{Output} \tag{2.19}$$

The Encoder-Decoder framework consists of three primary components:

Encoder: Compresses the input data into a latent representation. It typically consists of multiple layers that progressively extract higher-level features from the raw input.

$$\mathbf{z} = \text{Encoder}(\mathbf{x}) = f(\mathbf{x}; \theta_{\text{enc}}) \tag{2.20}$$

where $\mathbf{x}$ is the input data and $f(\cdot)$ represents the function learned by the encoder with θ_{enc} are the parameters of the encoder.

Latent representation: Latent representation of input from encoder or sampling distribution. In standard Encoder-Decoder models, the encoder produces a deterministic latent vector $\mathbf{z}$.

$$\mathbf{z} = \text{Encoder}(\mathbf{x}) = f(\mathbf{x}; \theta_{\text{enc}}) \tag{2.21}$$

On the other hand, in Variational Autoencoders (VAEs), the encoder outputs parameters of a probability distribution, typically a Gaussian, from which the latent vector $\mathbf{z}$ is sampled.

$$\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I}) \tag{2.22}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ are the mean and variance vectors output by the encoder. The $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mathbf{I})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\sigma}^2 \mathbf{I}$.

The stochastic nature of $\mathbf{z}$ in VAEs allows for the generation of diverse outputs and smooth interpolation in the latent space.

Decoder: Reconstructs the output data from the latent representation. It often mirrors the encoder's architecture but operates in reverse to generate the desired output format.

$$\hat{\mathbf{y}} = \text{Decoder}(\mathbf{z}) = g(\mathbf{z}; \theta_{\text{dec}}) \tag{2.23}$$

where $\hat{\mathbf{y}}$ is the reconstructed output and $g(\cdot)$ represents the function learned by the decoder with parameters θ_{dec} of the decoder.

Encoder-decoder architectures include various neural network structures designed for input and output data types. A notable example is U-Net (Ronneberger, Fischer, and Brox, 2015), a convolutional neural network (CNN) created for biomedical image segmentation. It features a symmetric design with a contracting path (encoder) that captures context through convolutional and pooling layers and an expansive path (decoder) for precise localization using upsampling and skip connections. This enables U-Net to effectively utilize high-level and low-level features, making it ideal for tasks requiring detailed spatial information.

Beyond U-Net, convolutional encoder-decoder models are widely used for image-to-image translation tasks like denoising, super-resolution, and style transfer, leveraging the powerful feature extraction of CNNs. Models like Long Short-Term Memory (LSTM) networks and Transformers are crucial in sequential data. LSTM-based models capture temporal dependencies and are suitable for machine translation and speech recognition applications. In contrast, transformer-based models enhance natural language processing through self-attention mechanisms, allowing for efficient

parallel processing and greater scalability. Their versatility extends to various tasks, including image captioning and multimodal data processing.

2.1.3 Deep learning loss functions and learning strategies

Loss functions and learning strategies are fundamental components of DL frameworks, playing pivotal roles in guiding the training process and optimizing model performance. This subsection provides a comprehensive literature review of various DL loss functions and learning strategies, elucidating their mechanisms, applications, and significance in developing effective neural network models.

Loss functions Loss functions also known as cost functions or objective functions, quantify the discrepancy between the predicted outputs of a neural network and the true target values. They serve as the foundation for the optimization process, guiding the adjustment of model parameters to minimize prediction errors. Some common loss functions used in various application of DL models are

- **Mean Squared Error (MSE) Loss**

  MSE loss is widely used for regression tasks, measuring the average squared difference between predicted and actual values.

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.24}$$

  where N is the number of samples, $y_i$is the true value, and $\hat{y}_i$ is the predicted value.

- **Cross-Entropy Loss**

  Cross-entropy loss is commonly employed for classification tasks, particularly binary and multi-class classifications. It measures the dissimilarity between the true label distribution and the predicted probability distribution.

  - **Binary Cross-Entropy Loss**:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{2.25}$$

  - **Categorical Cross-Entropy Loss**:

$$\mathcal{L}_{\text{CCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{2.26}$$

    where C is the number of classes, $y_{i,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for observation i, and $\hat{y}_{i,c}$ is the predicted probability for class c.

- **Dice Loss**: Commonly used in image segmentation tasks to maximize the overlap between predicted and ground truth masks.

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i + \epsilon}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i + \epsilon} \tag{2.27}$$

where ϵ is a small constant to prevent division by zero.

- **Perceptual Loss**: measures the difference between high-level feature representations of the predicted and ground truth images, rather than raw pixel values. This loss is particularly effective in tasks like style transfer and super-resolution, where maintaining perceptual similarity is more important than pixel-wise accuracy.

$$\mathcal{L}_{\text{Perceptual}} = \sum_{j=1}^{M} \|\phi_j(y) - \phi_j(\hat{y})\|^2 \tag{2.28}$$

where ϕ_j represents the activation of the j-th layer in a pre-trained network (e.g., VGG), and M is the number of selected layers.

- **Adversarial Loss**: Adversarial loss is employed in Generative Adversarial Networks (GANs) to train the generator to produce realistic images that can fool the discriminator (Goodfellow et al., 2014). It encourages the generator to create images indistinguishable from real ones.

$$\mathcal{L}_{\text{Adv}} = \mathbb{E}_{y \sim p_{\text{data}}}[\log D(y)] + \mathbb{E}_{\hat{y} \sim p_{\text{gen}}}[\log(1 - D(\hat{y}))]$$

where D is the discriminator network, p_{data} is the distribution of real images, and p_{gen} is the distribution of generated images. In generative modeling, $\mathbb{E}_{y \sim p_{\text{data}}}$ denotes taking the average (expected value) over real data samples, while $\mathbb{E}_{\hat{y} \sim p_{\text{gen}}}$ refers to the average over samples drawn from the model's generated distribution, thereby enabling comparisons between real and generated data.

Learning strategies encompass the methodologies and approaches employed during the training of deep learning models. These strategies significantly influence the model's ability to learn effectively, generalize to unseen data, and achieve optimal performance. The primary learning strategies include supervised learning, unsupervised learning, and transfer learning.

- Supervised learning involves training models on labeled datasets, where each input image is paired with a corresponding target label. This strategy is widely used in tasks such as image classification, object detection, and segmentation.
- Unsupervised learning involves training models on unlabeled data, allowing the model to discover inherent structures and patterns within the data. This strategy is beneficial for tasks where labeled data is scarce or expensive to obtain.
- Transfer learning leverages pre-trained models on large-scale datasets and adapts them to specific target tasks. This strategy is particularly effective when the target dataset is limited in size.

The learning process of deep learning (DL) models depends on optimization algorithms that minimize the chosen loss function by iteratively adjusting the model parameters. Common optimization algorithms, such as Stochastic Gradient Descent (Ruder, 2016), and Adam (Kingma and Ba, 2017), are crucial for ensuring convergence while balancing speed and stability, and for avoiding local minima. Learning strategies often include techniques like regularization, dropout, and data augmentation to improve model generalization and prevent overfitting. Understanding and selecting the right learning strategies and optimization algorithms is essential for designing robust and high-performing deep learning models.

By carefully selecting suitable loss functions tailored to specific tasks and utilizing effective learning strategies such as supervised, unsupervised, and transfer learning, researchers can improve model performance, ensure effective convergence, and enhance generalization to unseen data. Understanding the relationship between these components is crucial for designing robust and efficient neural network architectures that can tackle complex challenges.

This section explored the foundational architectures and components of deep learning, including perceptrons, multilayer perceptrons, CNNs, and Transformer models. We discussed the importance of activation functions in introducing nonlinearity and the role of attention mechanisms in capturing complex data dependencies. Additionally, we examined loss functions for image tasks and various learning strategies such as supervised, unsupervised, and transfer learning that optimize model performance. With this foundation, the next section will focus on incorporating domain knowledge into deep learning inputs to enhance model accuracy and interpretability.

Figure 2.3 summarizes visually how material domain knowledge can be integrated into every stage of the deep learning framework to highlight interactions among scientific hypotheses with model components. By developing the DL models with incorporations from domain-specific knowledge, researchers can produce predictive and interpretable representations reflecting the underlying scientific principles. This enables deeper insights into material behaviors while supporting iterative refinement in the models and underlying scientific theories.

2.2 Designing Deep learning inputs and outputs with domain knowledge

Designing inputs and outputs is particularly important in domain-enriched representation learning to enable deep learning models to utilize scientific data effectively. Unlike generic applications, scientific data often involve complex relations controlled by fundamental physics, chemistry, and biology principles. Therefore, the design of inputs and outputs that encapsulate this domain-specific knowledge will be essential to obtain models that are more accurate, interpretable, and compatible with scientific hypotheses. In Figure 2.4, we summarize the commonly seen material input and output data formats could be used in DL models for materials science. This subsection reviews the methodologies used to design inputs and outputs to be compatible with scientific data, emphasizing the incorporation of domain knowledge to enhance model performance and relevance.

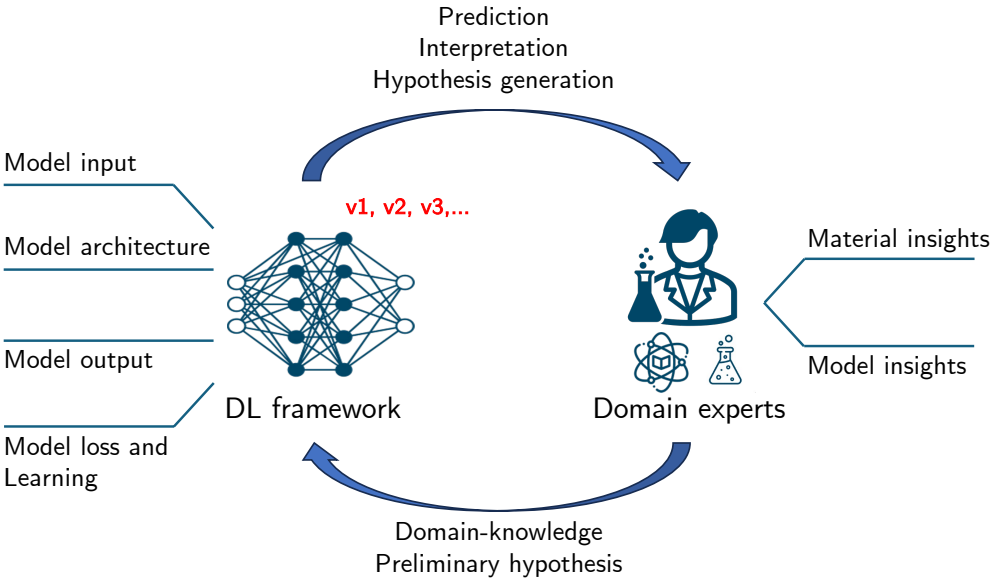


FIGURE 2.3: Schematic of Deep learning framework designing strategy for domain-enriched representation learning in material science.

2.2.1 Designing Deep learning outputs aligned with scientific objectives

The process starts with clearly identifying the targets and outputs that one wants to achieve through the DL framework, which is intrinsically linked to the scientific questions or hypotheses at hand. Some standard outputs in the materials informatics area are material property prediction, material type classification, material synthesis process suggestion, and segment/reconstructing material images. These outputs need to be designed based on a clear understanding of the scientific questions being addressed and the specific requirements for targets of downstream applications. In the case of regression tasks, this may involve continuous variables related to properties like band gap energy, thermal conductivity, or tensile strength. In the case of classification tasks, it could relate to material classification on structural phase, stability, or suitability for application in specific uses. This can be further extended to design the generative models to propose new material compositions or structures, thus aligning the output with the objective of discovering new materials with specific properties.

Besides, in scientific applications, it is useful to predict results and quantify the uncertainty associated with these results. It is thus vital to design outputs for models that can include estimates of uncertainty so that the scientists can have an informed idea about the reliability of the model's prediction (Jha et al., 2018). Bayesian neural networks or ensembling methods can be used to produce probabilistic outcomes and measure the confidence level of each prediction. Additionally, improvement in the interpretability of the results is necessary to extract scientific insights, which are the ultimate target in material science. Techniques such as feature importance analysis, attention mapping, and layer-wise relevance propagation may be incorporated into model design output to provide an understanding of which input features most greatly drive the predictions. This itergration, in turn, allows a more fundamental understanding of the underlying materials' behavior and further helps validate scientific hypotheses and insights from data.

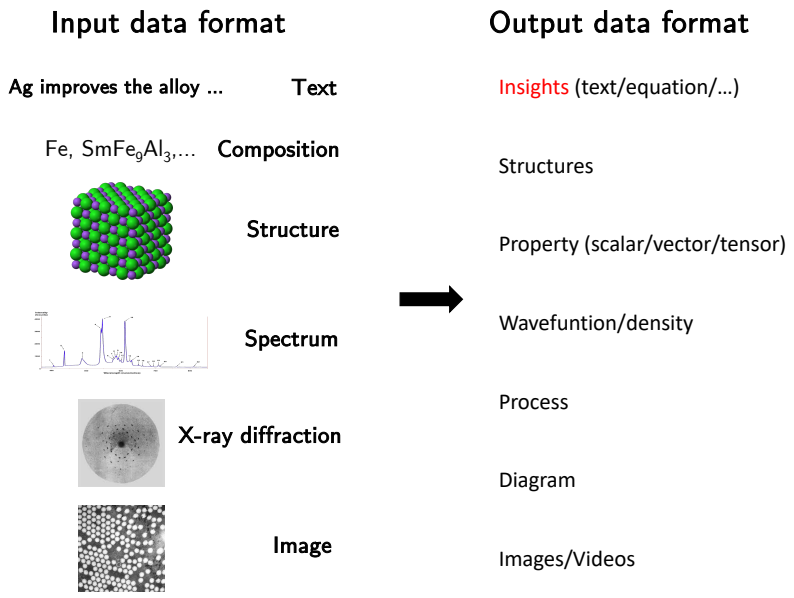


FIGURE 2.4: Illustration of material data formats and target data formats utilized in data-driven models for materials science. The insights regarding physicochemical behavior in materials are the primary target of application-driven models.

2.2.2 Designing Deep learning inputs with domain knowledge embedding

After defining the outputs, the next step is to design inputs that would carry the necessary information to produce those outputs. In scientific applications, inputs must codify the relevant features of an application influencing the desired outcome. This often involves incorporating domain knowledge into the input representations to capture the underlying physical and chemical interactions. Traditional methods are based on hand-designed features, typically from domain knowledge, such as atomic coordinates or bond lengths, electronic configuration, or optical systems. However, there could be potential limitations within these features to precisely represent the complexity of material behaviors; therefore, more intelligent methods must be designed to prepare inputs.

In many cases, this input transformation depends on certain scientific hypotheses that pre-determine the most relevant features to predict the desired output. This hypothesis-driven approach reinforces the input representations as data-driven and informed about appropriate scientific principles. For instance, one might hypothesize that the arrangement of atoms and the presence of defects in a material play a critical role in determining its mechanical properties; hence, input features can be transformed to highlight structural motifs, defect densities, and stress distributions (Chen et al., 2019). Examples of such transformations include applying dimensionality reduction, feature scaling, or domain-specific transformations that emphasize exciting aspects of the data.

The interplay between the input design and output requirements is a feedback loop that continuously refines both aspects to capture scientific data's complexities better. Models are trained and evaluated, and their outputs provide insights that might inform adjustments of the input representations to improve the performance iteratively. For instance, some input features may prove to be highly predictive of

the output variation; emphasis or expansion of these features could be made in subsequent input designs. On the other hand, if the model shows poor predictive performance for specific outputs, this may imply the inclusion of additional or alternative input features that capture the phenomena more precisely. This thus initiates a dynamic process of co-evolution between inputs and outputs that ensures deep learning models are informed by the scientific objectives at hand and adjusted to new insights emerging. By integrating feedback from model performance into their design of inputs, researchers can develop more robust and interpretable models that advance the understanding and discovery of materials.

Well-designed inputs and outputs extend far beyond predictive modeling and facilitate data mining tasks crucial to scientific exploration. Some common tasks that can take advantage of domain-enriched representations include clustering, classification, anomaly detection, and dimensionality reduction. In many cases, clustering algorithms might group materials that share similar properties, thus enabling the identification of novel material families or the discovery of trends that would have gone unnoticed. Similarly, anomaly detection can underline those materials that show unusual behaviors, triggering further investigations on their specific properties or synthesis conditions. Practical input and output design also enhances the interpretability of these data mining tasks, enabling the scientists to drive actionable insights from the model's predictions. By coupling the representations with domain knowledge, researchers can more readily correlate the model findings with existing scientific theories and hypotheses about the material systems under study.

2.3 Designing Deep learning architectures with domain knowledge

The architecture of deep learning models is critical in defining their performance in terms of meaningful representation learning from complex data. Integrating domain knowledge into model architectures becomes crucial for raising performance, interpretability, and generalization within scientific domains like materials science. By designing architectures that integrate domain-specific insights, models would be better positioned to capture the underlying physical, chemical, and structural principles governing material behaviors. This section will revisit some methodologies and techniques to inject domain knowledge into deep learning architecture to enhance representation learning and facilitate scientific discovery.

2.3.1 Incorporating Physical and Chemical Constraints

One of the fundamental ways to integrate domain knowledge into deep learning architectures is by embedding physical and chemical constraints directly into the model. This approach ensures that the learned representations adhere to known scientific laws and principles, enhancing the model's reliability and interpretability. For example, Physics-Informed Neural Networks incorporate into the loss function differential equations describing physical laws that guide the network to output solutions satisfying these equations (Raissi, Perdikaris, and Karniadakis, 2019). Such embedded constraints enable PINNs to model complex phenomena such as fluid dynamics or material deformation where conventional neural networks may lose physical plausibility.

Most importantly, geometrical constraints play a crucial role in materials science, wherein the spatial arrangement of atoms and molecules significantly affects material properties. By incorporating geometrical constraints into deep learning models, one can ensure that the learned representations respect the materials' inherent symmetries and structural invariances. For instance, rotational and translational symmetries are fundamental in crystal structures. By constructing convolutional layers that are equivariant to these transformations, models can ensure predictable consistency independent of the orientations or positions of the input data, as stated by Fuchs et al., 2020. This process not only generalizes a model across different configurations of materials but also simplifies the learning task by infusing prior knowledge about the geometry of materials.

Activation functions play a critical role in determining the behavior and stability of deep learning models. In materials informatics, activation constraints can be used to enforce physical limits and ensure that the model outputs remain within scientifically plausible ranges. For instance, certain material properties, such as density or thermal conductivity, cannot exceed specific physical bounds. By designing activation functions that inherently respect these limits—such as using bounded activation functions like sigmoid or tanh for properties with known upper and lower bounds—models can avoid unphysical predictions (Yao et al., 2022a). Additionally, enforcing non-negativity constraints on outputs where applicable (e.g., energy densities) can be achieved by employing activation functions like ReLU or its variants, which naturally restrict outputs to non-negative values.

Embedding scientific equations directly into neural network layers is another effective strategy for incorporating domain knowledge. This method involves designing custom layers that perform specific calculations based on physical or chemical laws. For example, in modeling material deformation, a layer can be designed to compute stress-strain relationships based on Hooke's Law or more complex constitutive models. By integrating these equations into the network architecture, the model is guided to produce outputs that are consistent with established physical theories (Fuchs et al., 2020; Anderson, Hy, and Kondor, 2019). This approach not only enhances the interpretability of the model by making its internal computations align with known scientific principles but also improves predictive accuracy by leveraging domain-specific relationships.

In some materials science applications, analyzing data in the frequency domain can reveal patterns and relationships that are not easily discernible in the spatial or time domains. Incorporating Fourier transformations into deep learning models allows the network to learn representations that capture frequency-based features. For example, in studying vibrational properties of materials, Fourier transforms can convert time-series data of atomic vibrations into frequency spectra, which can then be used as inputs for convolutional or recurrent neural networks (Metzler et al., 2018; Shamshad, Abbas, and Ahmed, 2019). By leveraging frequency domain representations, models can more effectively capture periodicities and oscillatory behaviors inherent in material properties, enhancing their ability to predict dynamic phenomena accurately.

Materials show behavior from atomic level interactions to properties at the macro scale. Integrating multiscale constraints within deep learning architecture allows models to reveal the hierarchical nature of material properties (Rupp et al., 2012). This can be enabled by utilizing multiscale neural networks comprising interconnected modules to process information on various scales. For example, a multiscale network could have separate branches for atomic configurations and structures at the mesoscale; each branch would learn scale-specific representations that

later merge to predict properties at the macroscopic scale. This hierarchical approach enables models to incorporate constraints and interactions at each relevant scale, affording a holistic understanding of material behavior.

Incorporating physical and chemical constraints into deep learning architectures allows for constructing accurate but scientifically meaningful models in materials science. This embedding can be done through geometrical constraints, activation constraints, incorporation of domain-specific equations, derivative constraints, modeling noisy data, Fourier transformations, and multiscale constraints; thus, deep learning models will make sure to respect fundamental scientific principles and capture complicated and multifaceted natures of material behaviors. This will enhance the reliability and interpretability of model predictions and facilitate the discovery of new materials by unmasking patterns and relationships that are somewhat complicated yet consistent with established scientific knowledge. As deep learning continues to evolve, integrating domain knowledge and such constraints will remain a cornerstone for advancing materials informatics and driving scientific innovation.

2.3.2 Utilizing specialized layers

Another strategy for embedding domain knowledge involves using specialized layers and modules within deep learning architectures. These components are designed to capture specific aspects of the data pertinent to the domain. For instance, Convolutional Neural Networks (CNNs) incorporate convolutional layers that exploit spatial hierarchies in data, making them well-suited for image-based tasks. In materials science, similar principles can be applied by designing layers that capture spatial relationships between atoms or molecules.

Graph Neural Networks (GNNs) exemplify this approach by using graph-based layers to model the interactions between atoms in a material. GNNs represent atoms as nodes and bonds as edges, allowing the network to naturally encode the structural information inherent in molecular and crystalline systems (Karamad et al., 2020; Rahaman and Gagliardi, 2020). This architectural choice leverages the relational nature of material data, enabling the model to learn representations that reflect the underlying atomic interactions and spatial configurations critical for predicting material properties.

Attention mechanisms represent another advanced module that significantly enhances the capability of deep learning models to handle complex correlations within data (Fuchs et al., 2020; Vaswani et al., 2017). Attention models allow the network to dynamically focus on different parts of the input data when making predictions, effectively modeling intricate dependencies and interactions. In materials science, attention mechanisms can be utilized to emphasize critical atomic or molecular interactions that substantially impact material properties. For example, in predicting the thermal conductivity of a material, an attention layer can prioritize interactions between specific atom pairs that contribute most to heat transfer.

Furthermore, incorporating autoencoder architectures facilitates unsupervised representation learning by compressing input data into lower-dimensional latent spaces and then reconstructing the original data from these representations (Cherukara et al., 2020; Wengrowicz et al., 2020). In materials science, autoencoders can be employed to learn compact and meaningful representations of complex material structures, enabling tasks such as anomaly detection, clustering, and dimensionality reduction. These learned representations can reveal hidden patterns and relationships within the data, providing valuable insights for material discovery and characterization.

2.3.3 Feedback mechanisms and iterative refinement

The integration of domain knowledge into deep learning architectures is often an iterative process, where feedback from model performance informs subsequent architectural adjustments. Feedback mechanisms, such as attention layers or iterative refinement modules, allow models to dynamically adjust their focus based on the relevance of different features or interactions. In materials science, attention mechanisms can help models prioritize atomic interactions that are most influential in determining material properties, thereby enhancing the efficiency and accuracy of representation learning (Vaswani et al., 2017).

Additionally, iterative refinement processes enable models to progressively enhance their representations by revisiting and refining previously learned features. This approach aligns with the scientific method, where hypotheses and models are continuously tested and improved based on new data and insights. By incorporating such feedback loops, deep learning architectures can evolve to better capture the complexities of material systems, leading to more accurate and insightful representations.

Active learning is another crucial component of the feedback loop in representation learning. By actively selecting the most informative data points for training, models can focus their learning efforts on areas where they are most uncertain or where additional data would yield the greatest improvement in performance (Nguyen et al., 2023). In materials science, this might involve identifying and synthesizing new material samples that exhibit unique or extreme properties, thereby expanding the model's understanding of the material space. Active learning not only optimizes the use of limited experimental resources but also ensures that the model continuously evolves to incorporate new and diverse data, enhancing its generalizability and robustness

Designing deep learning architectures with embedded domain knowledge is a pivotal strategy for advancing representation learning in materials science. By integrating physical and chemical constraints, utilizing specialized layers, embedding symmetries, capturing hierarchical structures, leveraging transfer learning, and incorporating feedback mechanisms, researchers can develop models that are both powerful and scientifically meaningful. These domain-enriched architectures not only enhance the accuracy and interpretability of predictions but also facilitate the discovery of new materials by uncovering intricate patterns and relationships inherent in complex material data. As deep learning continues to evolve, the synergistic integration of domain knowledge will remain a cornerstone for unlocking the full potential of data-driven approaches in materials science.

2.4 Designing Deep learning loss functions and evaluation metrics with domain knowledge

In deep learning, the selection of appropriate loss functions and evaluation metrics is paramount to the successful training and assessment of models. Loss functions guide the optimization process by quantifying the discrepancy between the predicted outputs and the true values, thereby enabling the model to learn from data effectively. Evaluation metrics, on the other hand, provide a means to assess the performance and generalizability of the trained models. In the context of materials science, where the objectives often extend beyond mere prediction accuracy to include adherence to physical laws and scientific interpretability, the design of

custom loss functions and the selection of domain-relevant metrics become crucial components of representation learning

2.4.1 Custom loss functions for incorporating domain knowledge

Traditional loss functions, such as mean squared error (MSE) or cross-entropy loss, are widely used in various deep learning applications due to their simplicity and effectiveness. However, in materials science, these standard loss functions may not sufficiently capture the intricate relationships and physical constraints inherent to material properties. To address this limitation, custom loss functions are developed to embed domain-specific knowledge directly into the learning process. For instance, Physics-Informed Neural Networks (PINNs) incorporate differential equations that represent physical laws into the loss function, ensuring that the model's predictions adhere to these fundamental principles (Raissi, Perdikaris, and Karniadakis, 2019). By penalizing deviations from known physical behaviors, such custom loss functions enhance the model's ability to produce physically plausible and scientifically meaningful representations.

Another approach involves the use of regularization terms that enforce specific constraints related to material properties. For example, in predicting mechanical properties, constraints can be added to ensure that the predicted values respect conservation laws or symmetry properties observed in real materials. These domain-informed regularizations not only improve the model's predictive accuracy but also enhance its interpretability by aligning the learned representations with established scientific theories

In materials science, image reconstruction tasks, such as scanning electron microscopy (SEM) image enhancement or tomography, benefit significantly from custom loss functions that incorporate physical constraints. For example, in SEM image reconstruction, loss functions can be designed to enforce spatial consistency and fidelity to the underlying material structure. By integrating constraints that preserve edge sharpness or specific texture patterns characteristic of certain materials, models can achieve higher-quality reconstructions that are both visually accurate and scientifically relevant (Haan et al., 2019).

Furthermore, in tomography, custom loss functions can incorporate geometric constraints derived from the physics of imaging systems. These constraints ensure that reconstructed images adhere to the principles of projection geometry and material density distributions. By embedding such domain-specific knowledge, deep learning models can produce more accurate and reliable reconstructions, facilitating better analysis and interpretation of material structures (Wang, Ye, and De Man, 2020).

Generative models, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), are powerful tools for creating novel material structures or simulating material behavior under various conditions. In materials science, ensuring the chemical validity of generated structures is paramount. Custom loss functions can be designed to enforce chemical rules, such as valency constraints or stoichiometric balances, during the generation process. For instance, in the generation of new alloy compositions, loss functions can penalize configurations that violate known chemical bonding rules, ensuring that the synthesized materials are chemically feasible (Dan et al., 2020).

Additionally, for applications like molecular generation, loss functions can incorporate constraints that maintain molecular stability and functional group integrity. By embedding these chemical principles into the loss function, generative models

can produce realistic and functional molecular structures that are more likely to exhibit desired properties, thereby accelerating the discovery of new materials with targeted functionalities.

Accurate similarity measurements between materials are crucial for tasks such as clustering, classification, and recommendation in materials informatics. Custom loss functions can be designed to incorporate structural invariants and domain-specific similarity metrics that reflect the true relational nature of materials. For example, in graph-based representations of materials, loss functions can be tailored to preserve topological features and connectivity patterns that are indicative of similar material properties (Wang et al., 2024).

By embedding structural invariants into the loss function, models can learn representations that maintain meaningful similarities between materials, even when traditional numerical descriptors fail to capture these relationships. This leads to more effective clustering of materials based on their intrinsic properties and facilitates the discovery of new materials by identifying those that are structurally similar to known high-performance materials.

Regularization techniques are integral to preventing overfitting and enhancing the generalization capabilities of deep learning models. In materials informatics, custom regularization terms can be crafted to enforce smoothness, sparsity, or other domain-specific properties within the learned representations. For example, in predicting material properties, a smoothness regularization term can be introduced to ensure that small changes in input features lead to gradual and consistent changes in the output predictions, reflecting the continuous nature of material behaviors (Kimanius et al., 2024).

Custom loss functions play a vital role in embedding domain knowledge into deep learning models, ensuring that the learned representations are both accurate and scientifically meaningful in materials science. This integration not only improves predictive performance but also enhances interpretability and trustworthiness, paving the way for more effective and insightful applications of deep learning in materials informatics. As the field continues to evolve, the development of increasingly sophisticated custom loss functions will remain essential for advancing the synergy between deep learning and materials science, driving forward innovation and discovery.

2.4.2 Evaluation metrics from Deep learning and Materials science perspectives

Evaluation metrics in materials informatics must bridge the gap between traditional deep learning performance indicators and domain-specific requirements. While metrics such as accuracy, precision, recall, and F1-score are essential for classification tasks, materials science often demands additional metrics that evaluate the physical relevance and scientific validity of the predictions. For regression tasks, metrics like root mean squared error (RMSE) and mean absolute error (MAE) are commonly used; however, in materials science, it is also important to consider metrics that assess the adherence to physical laws and the consistency of predicted properties with known material behaviors.

Moreover, materials-specific metrics may include the deviation of predicted properties from experimentally measured values, the ability to generalize across different material classes, and the robustness of predictions under varying environmental conditions. Incorporating these specialized metrics ensures that the deep learning models not only perform well statistically but also provide valuable insights that are

actionable within the scientific domain. Materials often exhibit different behaviors under varying environmental conditions such as temperature, pressure, or chemical environments. Metrics that assess the robustness of model predictions under such variations are crucial. Sensitivity analysis metrics, which measure how changes in input conditions affect the outputs, can be employed to evaluate model stability and reliability (Peng et al., 2022). A model that maintains consistent performance across a range of conditions is more valuable for practical applications.

2.4.3 Learning strategies: Supervised, Unsupervised, and Hybrid approaches

The choice of learning strategy significantly influences the design of loss functions and evaluation metrics in materials informatics. Supervised learning, where models are trained on labeled data, is prevalent in tasks such as property prediction and classification of material types. In this paradigm, the loss function is directly tied to the accuracy of the predictions against known labels, making it straightforward to apply traditional and custom loss functions (Butler et al., 2018).

Unsupervised learning, on the other hand, focuses on uncovering hidden patterns and representations within unlabeled data. Techniques such as clustering, dimensionality reduction, and autoencoders are employed to explore the intrinsic structure of material data. In this context, loss functions may emphasize reconstruction accuracy or the preservation of data manifold structures, while evaluation metrics might include measures of cluster cohesion and separation or the quality of the learned embeddings (Himanen et al., 2019).

Hybrid approaches that combine supervised and unsupervised learning paradigms are increasingly being adopted to leverage the strengths of both strategies. For example, semi-supervised learning can utilize a small amount of labeled data alongside a large volume of unlabeled data, enhancing the model's ability to generalize and reducing the dependency on extensive labeled datasets. Such strategies necessitate the design of composite loss functions that balance supervised objectives with unsupervised regularizations, as well as the development of multifaceted evaluation metrics that capture performance across different learning dimensions.

2.4.4 Balancing predictive accuracy and scientific interpretability

A critical challenge in designing loss functions and evaluation metrics for materials informatics is achieving a balance between predictive accuracy and scientific interpretability. Highly accurate models that lack interpretability may offer limited utility in scientific discovery, as they do not provide insights into the underlying material mechanisms. Conversely, models that prioritize interpretability may sacrifice some degree of accuracy, potentially limiting their predictive power.

To address this challenge, researchers employ multi-objective optimization techniques that simultaneously optimize for both accuracy and interpretability. For example, regularization terms can be introduced to encourage sparsity or simplicity in the learned representations, making them easier to interpret without significantly compromising predictive performance. Additionally, visualization techniques and feature importance analyses are integrated into the evaluation process to provide intuitive explanations of the model's predictions, thereby enhancing the interpretability of complex deep learning models (Butler et al., 2018).

One effective strategy for balancing accuracy and interpretability is the gradual addition of features to the model. Starting with a minimal set of well-understood, domain-relevant features allows researchers to build a foundation of interpretable

models. As the model's performance on these core features is established, additional features can be incrementally introduced to capture more complex interactions and nuances within the data. This staged approach helps prevent the model from becoming overly complex too quickly, which can obscure the underlying mechanisms and reduce interpretability.

By carefully selecting and adding features, researchers can maintain control over the model's complexity, ensuring that each new feature contributes meaningfully to the predictive performance without overwhelming the model with unnecessary information. Techniques such as feature selection algorithms, which identify the most impactful features based on statistical criteria or domain knowledge, can aid in this process. Furthermore, dimensionality reduction methods like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbor Embedding (t-SNE) can be employed to distill high-dimensional data into more manageable representations, preserving essential information while enhancing interpretability

Another crucial aspect of balancing accuracy and interpretability involves the careful design of model parameters, progressing from simple to more complex configurations. Initially, models can be trained with a smaller number of parameters and simpler architectures, ensuring that the fundamental relationships within the data are captured without introducing excessive complexity. As understanding of the data deepens and initial models demonstrate reliable performance, additional parameters and more intricate architectural elements can be incorporated to refine the model's capabilities.

This incremental approach not only facilitates a clearer interpretation of how each parameter and architectural choice influences the model's predictions but also mitigates the risk of overfitting. By progressively expanding the model's capacity, researchers can monitor performance improvements and maintain a balance between model sophistication and interpretability. Techniques such as cross-validation and regularization can further support this strategy by ensuring that the model generalizes well to unseen data while avoiding unnecessary complexity (Bishop, 2006).

Another approach to maintaining interpretability is model simplification, which involves reducing the complexity of deep learning models without significantly sacrificing their performance. Techniques such as pruning, where less important weights or neurons are systematically removed from the network, can streamline the model while preserving its core predictive capabilities. Pruned models are generally easier to interpret, as they contain fewer parameters and simpler structures, making it easier to identify and understand the key factors driving the model's predictions (Han, Mao, and Dally, 2016).

In the context of materials informatics, simplified models can facilitate the identification of critical material features and their interactions, providing clearer insights into the mechanisms governing material properties. Additionally, simpler models often require less computational resources, enabling faster training and evaluation, which is beneficial for iterative scientific research processes

Figure 2.5 presents a schematic overview of the comprehensive model design strategy for material representation and analysis, which seamlessly integrates raw material data, scientific insights, and domain knowledge to inform and optimize the design of model inputs, architectures, outputs, and loss functions. The framework emphasizes enriched feature selection, the incorporation of physical and chemical constraints, geometric representation learning, and the utilization of physics-informed operators to create robust, scientifically grounded representations. Outputs are guided by custom loss functions and domain-specific evaluation metrics, while transfer learning techniques enable adaptability across diverse material classes.

An iterative feedback and refinement process ensures that the model design remains aligned with material hypotheses and experimental validations, thereby balancing predictive accuracy with scientific interpretability and fostering a synergistic advancement in materials informatics.

Building upon this comprehensive design strategy, the subsequent sections delve into the practical applications of deep learning in materials science. By adhering to the outlined framework and effectively embedding domain knowledge, these applications demonstrate how deep learning can address specific challenges in materials discovery, property prediction, and structural analysis. Through detailed case studies and examples, we illustrate the integration of enriched feature selection, domain-informed architectures, and custom loss functions to solve complex materials science problems, thereby showcasing the transformative potential of deep learning in advancing the field.

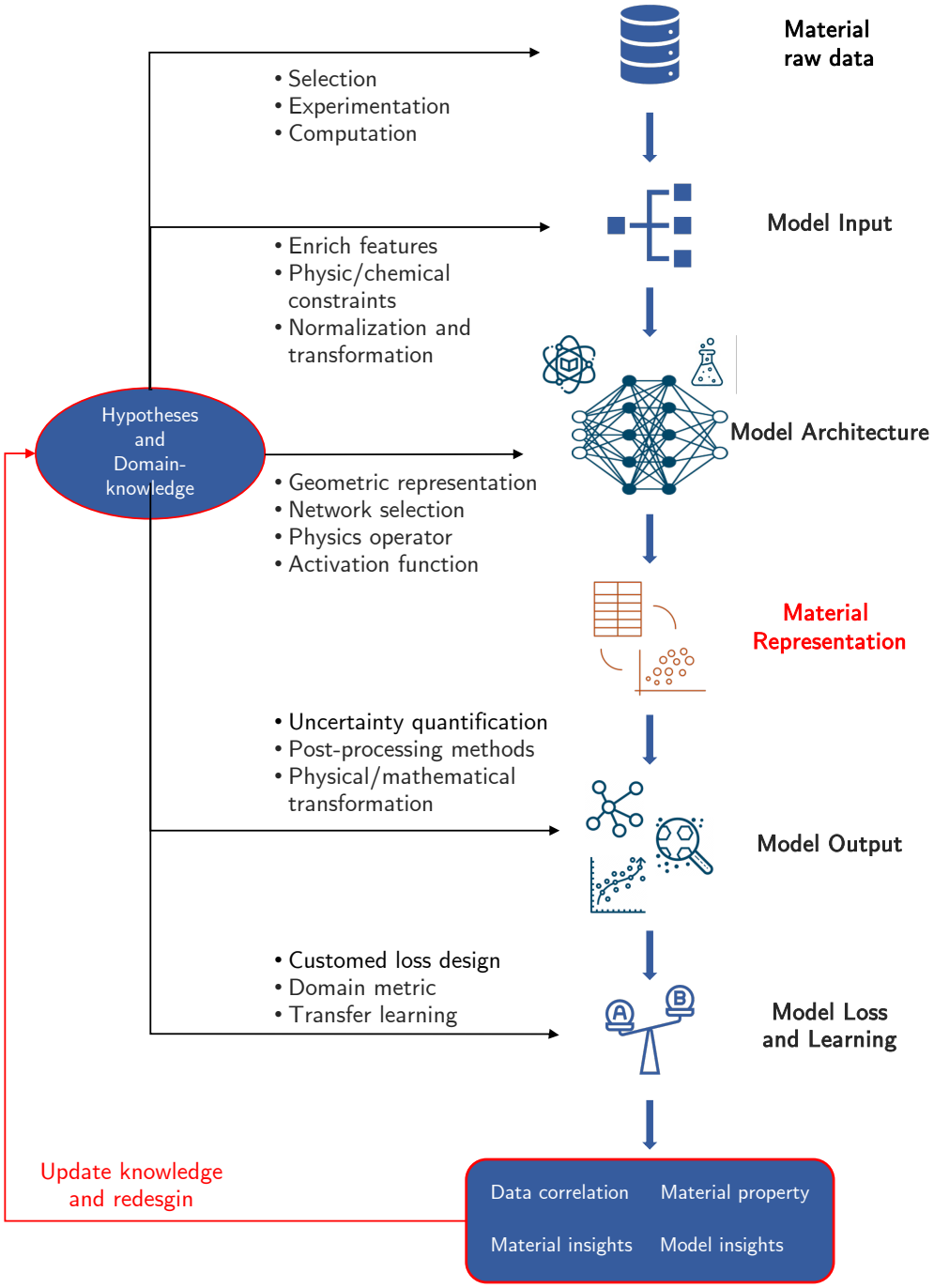


FIGURE 2.5: Schematic overview of the model design strategy for material representation and analysis. The framework integrates material raw data, insights, and domain knowledge to inform the design and optimization of model inputs, architecture, outputs, and loss functions. Key components include: enriched feature selection, incorporation of physical and chemical constraints, geometric representation learning, and physics-informed operators. The outputs are guided by custom loss designs, domain-specific metrics, and transfer learning techniques. The iterative process of updating knowledge and redesigning ensures alignment with material hypotheses and experimental validation.

Chapter 3

Deep Learning Framework Design for Unsupervised Representation Learning in Image Reconstruction from Diffraction Data

3.1 Introduction

X-ray imaging offers unparalleled insights into the internal structures of materials without causing damage, making it indispensable in materials science and engineering (Wang et al., 2008; Withers et al., 2021; Johnson et al., 2023). X-ray diffraction, which is central to techniques such as X-ray crystallography, facilitates detailed studies of crystalline materials at the atomic or molecular levels. Coherent X-ray diffraction imaging (CXDI) advances these capabilities, enabling the visualization of materials (without the requirement of the crystallinity) with high spatial resolution, surpassing traditional lens-based approaches (Chapman and Nugent, 2010; Keller, 1957) and the capabilities of conventional methods such as electron and probe microscopy, which usually require extensive sample preparation and operate in special environments, offering only excellent surface resolution without providing internal structural information. Thus, CXDI enables non-destructive imaging of both the surface and internal structures of materials in their natural state, with exceptional spatial resolution (Miao et al., 2015a).

In CXDI, a well-defined coherent X-ray beam is directed onto an sample, producing a diffraction image that includes intensity information but lacks phase details (Chapman et al., 2006; Robinson and Harder, 2009; Sun et al., 2012; Abbey, 2013; Ulvestad et al., 2015; Yau et al., 2017; Pedersen et al., 2020; Takazawa et al., 2021). Reconstruction of the sample image necessitates sophisticated phase retrieval analysis, which recovers the phase information from the measured diffraction image. The inverse problem typically uses iterative computational methodologies that employ forward diffraction simulation and image-reconstructing algorithms based on error reduction, hybrid input–output (HIO) (Marchesini et al., 2003), or difference map (Latychevskaia, 2018; Fienup, 1982; Kliuiev et al., 2016; Marchesini, 2004). However, the inherent ambiguity in phase retrieval analysis, in which a single diffraction image with only intensity information may correspond to multiple potential sample images, complicates image reconstruction and evaluation of the reliability of the constructed images. Moreover, it limits real-time applications crucial for in-situ and operando experiments (Kourousias et al., 2018; Datta et al., 2019).

Scanning CXDI, commonly known as X-ray ptychography (Pfeiffer, 2018), effectively images extensive samples by systematically scanning a coherent beam across

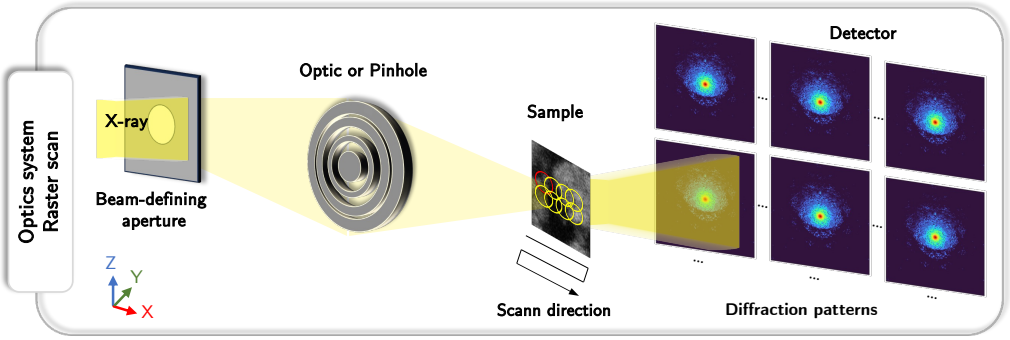


FIGURE 3.1: Schematic illustration of ptychographic setup for the overlapping raster scan.

them, necessitating overlapping spatial sequences to ensure each section contributes multiple diffraction images (Fig 3.1). This redundancy is vital as it provides additional spatial constraints to verify the accuracy of the phase retrieval analysis, enhancing the convergence of the analysis. Despite their advantages, phase retrieval methods using iterative algorithms such as the extended ptychographical iterative engine (ePIE) (Maiden and Rodenburg, 2009), which is widely recognized for its high convergence rate and noise robustness, often struggle without static overlapping illumination areas and are highly sensitive to parameter selections (Cherukara et al., 2020). These challenges significantly compromise accuracy and limit capabilities in imaging dynamic phenomena (Miao et al., 2015b; Rodenburg and Maiden, 2019; Nashed et al., 2014).

Scanning CXDI has broader issues, especially in capturing high temporal and spatial resolution of samples with dynamic processes. In particular, the scanning speed of the coherent X-ray beam moving across the whole sample for collecting diffraction images directly affects the ability to capture fast-evolving phenomena. The CXDI is typically performed in a step-scan manner, where the beam stops at discrete points following a previously designed scenario. On-the-fly scanning is also developed, which attempts to speed up the measurement process by moving the beam across the sample without stopping and the data is collected continuously as the beam moves. Regardless of the method, a slow scanning speed may result in temporal inconsistencies between imaging regions of the sample, as simultaneous imaging of every local region is desired. Conversely, a faster scanning speed shortens the exposure time for each illuminated local area, reducing diffraction signals and lowering image quality. Additionally, the assumption that each sample section is consistently illuminated multiple times and contributes multiple diffraction images may not hold during dynamic changes in the sample, potentially reducing the accuracy of the phase retrieval analysis and negatively affecting the spatial resolution of the reconstructed image.

Several non-scanning CXDI methods, which use single-shot imaging to capture multiple frames, have been proposed to depict the dynamic process with excellent temporal resolution (Abbey et al., 2008; Zhang et al., 2016; Khakurel et al., 2017; Lo et al., 2018; Levitan et al., 2020; Takayama et al., 2021a). However, these remain in the proof-of-principle stage, and comprehensive studies on the dynamics of actual samples in the hard X-ray region must be undertaken. One of the challenges in phase retrieval analysis for the single-shot CXDI is the lack of additional spatial constraints, which complicates image reconstruction and affects the reliability of the results. Additionally, low signal-to-noise ratios (SNR) present another challenge,

as the single-shot CXDI requires short exposure times per frame to capture fast-evolving phenomena, leading to fewer photons being detected and higher noise in the measurements. Recent advancements have led to the proposal of practical methods for single-shot CXDI, which can effectively and comprehensively elucidate material dynamics and functions (Takazawa et al., 2021; Kang et al., 2021; Takazawa et al., 2023).

Deep learning (DL) can be applied to address complex imaging problems like those encountered in holography and phase retrieval analysis, enhancing accuracy and computational efficiency (Cherukara et al., 2020; Wu et al., 2021a; Yao et al., 2022b; Bohra et al., 2023; Welker et al., 2022; Cha et al., 2021; Lee et al., 2021; Zhuang et al., 2022; Zhang et al., 2021a; Ye, Wang, and Lun, 2023; Gugel and Dekel, 2022; Cherukara, Nashed, and Harder, 2018; Ye, Wang, and Lun, 2022). These techniques often face challenges such as the accurate reconstruction of phase information from intensity measurements and handling the inherent noise and artifacts in the data. Among the notable advancements, PtychNet, a convolutional neural network (CNN), facilitates the direct reconstruction of Fourier ptychography data. Similarly, PtychoNN (Cherukara et al., 2020) and SSPNet, which are encoder–decoder deep neural network (DNN) architectures, are designed explicitly for single-shot CXDI. These networks streamline the imaging process by directly reconstructing sample images without explicitly reconstructing phase information, thus simplifying the imaging process (Kappeler et al., 2017; Wengrowicz et al., 2020; Cherukara et al., 2020). Deep PtychShamshad, Abbas, and Ahmed, 2019 employs generative models to regularize the phase problem and achieve superior reconstruction quality; however, it requires the identification of an applicable prior model. Conversely, prDeep (Metzler et al., 2018) optimizes an explicit minimization objective using a generic denoising DNN trained for a specific target data distribution.

Recent advancements in DL for video processing have increased our ability to capture the semantics and context of dynamic scenes (Bertasius, Wang, and Torresani, 2021). Leveraging sophisticated neural network architectures, DL methods can extract meaningful temporal variations and patterns from sequential image frames, surpassing the results of traditional image analysis. This capability demonstrates the potential of DL in the CXDI of dynamic phenomena. By analyzing temporal context and correlations in diffraction images and reconstructed images, DL is expected to facilitate high-resolution image reconstructions of dynamically changes within the sample and learn the underlying mechanisms of these phenomena. This learning may, in turn, enhance the quality of phase retrieval and image reconstruction, enabling a reciprocal relationship in which each aspect reinforces the other.

Despite remarkable advances in phase retrieval analysis that benefit from DL, DL-based methods are associated with numerous challenges. A notable issue is the reliance on supervised learning, which necessitates the use of ground-truth data for training. However, obtaining such data in real-world phase retrieval scenarios is often unfeasible, leading to unreliable image reconstruction models when only simulated data are used. Several unsupervised DL methods, such as AutoPhaseNN (Yao et al., 2022a) and DeepMMSE (Chen et al., 2022a), have been proposed to address this issue. These methods do not require ground-truth data for constructed images (Tripathi, McNulty, and Shpyrko, 2014) but need appropriate constraints tailored to the optical settings. Nonetheless, they often trade time efficiency for accuracy, which compromises their suitability for in-situ experiments. Incorporating physical insights and mathematical constraints into DL-based methods can enhance their performance on complex problems (Zhang et al., 2016; Lucas et al., 2018; Kang et al., 2021).

This study introduces a DL-based method specifically designed to address the inherent challenges of phase retrieval analysis for single-shot CXDI, aiming to significantly enhance the visualization of local nanostructural dynamics within the sample. This method adopts a unsupervised learning strategy, where the model is trained using the data itself to generate supervisory signals, rather than relying on external labels provided by humans. The DL architecture is composed of a measurement-informed refined neural network block (RB), designed to integrate optical settings and mathematical constraints into the learning process, and a temporal neural network block (TB), designed to capture the temporal correlations between diffraction images and those of the corresponding reconstructed images via the learning process. Additionally, a single-shot CXDI optical system was constructed to compare the proposed method with state-of-the-art methods in real-world experiments.

3.2 Integrating Domain Knowledge into Model Design Strategies

3.2.1 Principle of coherent X-ray diffraction

The common coherent X-ray system uses a monochromatic X-ray beam, which is shaped by an aperture to generate X-ray beam illuminating the sample. The exit wavefield $\psi(\mathbf{r})$ of the sample can be represented as

$$\psi(\mathbf{r}) = P(\mathbf{r}) \times O(\mathbf{r}), \tag{3.1}$$

where $\mathbf{r}$ denotes the real-space coordinate vector, and $P(\mathbf{r})$ is the illumination probe function. The complex object function $O(\mathbf{r}) = A(\mathbf{r}) * e^{i\phi(\mathbf{r})}$ is the mathematical representation of the sample in real-space, which comprises amplitude $A(\mathbf{r})$ and phase $\phi(\mathbf{r})$ components.

This exit wavefield produces a diffraction images in the far field, captured as a two-dimensional intensity pattern by a downstream detector. The intensity of the diffraction images can be expressed as

$$I(\mathbf{q}) = |\Psi(\mathbf{q})|^2 = |\mathcal{F}[\psi(\mathbf{r})]|^2, \tag{3.2}$$

where $\mathbf{q}$ denotes the reciprocal-space coordinate vector, and $\Psi(\mathbf{q})$ is the wavefront of the $\psi(\mathbf{r})$ in the detector plane. $\mathcal{F}$ represents the Fourier transform operator. The objective of phase retrieval in a CXDI experiment is to inversely derive a unique complex object function O from I diffraction pattern.

In practice, the measured diffraction intensity $I(\mathbf{q})$ on the detector is represented as a two-dimensional image, usually a square, denoted as $I \in \mathbb{N}^{m\times m}$ (m is the measurement size of the image in pixels). The illumination probe function $P(\mathbf{r})$ is a two-dimensional matrix $P \in \mathbb{C}^{m\times m}$ and is assumed constant during the experiment in a fixed optical system with a coherent X-ray (stable illumination source). Such probe functions are often estimated in advance via scanning CXDI and are assumed to remain constant. The complex object function $O(\mathbf{r})$ is represented by two-dimensional matrices $A \in \mathbb{R}^{m\times m}$ and $\phi \in \mathbb{R}^{m\times m}$ expressing the amplitude and phase information, respectively.

For scanning CXDI with N positions, the observed diffraction intensity image at the t^{th} frame is denoted as I_t. Typically, the conventional phase-retrieval algorithm iteratively refines the object function O_t based on the reciprocal–space constraint. Starting from an initial guess O^0, the estimation of the object function in the iteration

$(i+1)^{th}$ is expressed as follows:

$$\begin{aligned} O_t^{i+1} &= O_t^i + \alpha U^i \\ &= O_t^i + \alpha \frac{\bar{P}(\psi'_t - \psi_t)}{|P|^2_{\max}}, \end{aligned} \tag{3.3}$$

where $\bar{P}$ is the complex conjugate of P. The scalar α is a feedback parameter of the updating object function U and $\psi_t = P \times O_t^i$ is the wavefield of the sample reconstructed from the previous iteration i^{th}. Meanwhile, ψ'_t is the revision of the wavefield ψ_t satisfying the constraint in the reciprocal space:

$$\psi'_t = \mathcal{F}^{-1}\left[\sqrt{I_t}\frac{\Psi_t}{|\Psi_t|^2}\right], \tag{3.4}$$

where $\Psi_t = \mathcal{F}[\psi_t]$. $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operator. Through the iterative process, the refinement algorithm is expected to converge to a solution of the complex object function O_t that corresponds to I_t.

However, due to the complexity of both the object and the illumination in diffraction experiments, it is necessary to ensure that the overlapping condition is met by having at least 50% overlap between scanning positions (Pfeiffer, 2018). This substantial overlap is crucial for accurately reconstructing the object from the collected diffraction patterns, as it provides sufficient redundancy and constraints to solve the phase retrieval problem by the iterative algorithm. The requirement for such a high degree of overlap introduces challenges in dynamic CXDI, particularly when attempting to capture time-resolved changes in materials.

3.2.2 Designing output targets for image reconstruction

The ultimate objective in phase retrieval for dynamic coherent X-ray diffraction imaging (CXDI) is to accurately reconstruct the evolving object and gain insights into the underlying dynamic phenomena within materials. In an unsupervised learning context, this involves designing a model that can learn from unlabeled diffraction data to produce meaningful reconstructions without explicit ground truth references.

To align the model outputs with both the physical principles of diffraction and the goals of dynamic imaging, the designs of outputs need to include:

- Reconstructed Object: A spatial representation of the material's structure at each time point, capturing its dynamic changes over time.
- Computed Diffraction Patterns: Simulated diffraction data obtained by applying the forward Fourier Transform to the reconstructed object, which can be directly compared with the experimentally observed diffraction patterns.

By including both the reconstructed object and the computed diffraction patterns as outputs, the model can iteratively adjust the object's representation to minimize the discrepancy between the computed and experimental diffraction data. This design ensures that the reconstruction adheres to the mathematical formulation of diffraction, specifically the Fourier relationship between the object and its diffraction pattern.

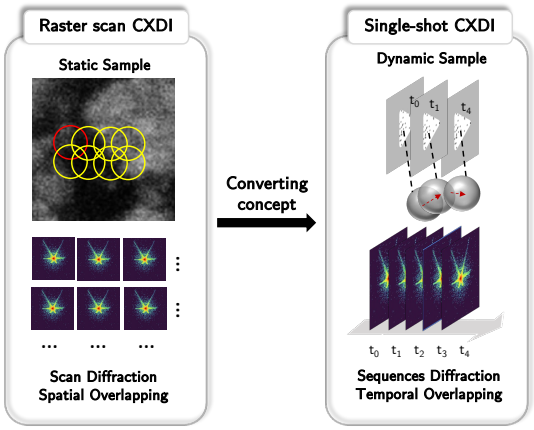


FIGURE 3.2: Overlapping concept in raster scan CXDI for static object and single-shot CXDI for dynamic object sample.

Moreover, it is crucial that the reconstructed object exhibits consistency over time to accurately reflect the material's dynamics. While the iterative methods independently reconstruct single object at a time, there is no guarantee the stable and consistent of objects' representation overtime. To overcome this, the output design need to incorporates temporal coherence by enforcing smoothness or continuity constraints across consecutive time frames.

3.2.3 Designing input transformations aligned with reconstruction goals

In the context of dynamic coherent X-ray diffraction imaging (CXDI), the requirement for consistent reconstruction of the object over time directly influences how we prepare and transform the input data for the model. To achieve outputs that accurately reflect the temporal evolution of the material, we must design the input transformation to align with these objectives.

Understanding the physical movement of the object provides critical domain knowledge that informs our input transformation strategy. In many dynamic CXDI experiments, the object's motion between consecutive time frames is minimal due to the high frame rates and relatively slow dynamics of the material processes being studied .

- Converting Spatial Overlap to Temporal Overlap: Traditionally, ensuring sufficient spatial overlap in scanning positions is necessary for accurate phase retrieval in static imaging. However, in dynamic CXDI, the minimal movement of the object over short time intervals means that overlapping information can be obtained through the temporal dimension. This effectively transforms the requirement of spatial overlap into one of temporal overlap (Fig 3.2).

Based on the need for temporal consistency in the reconstructed object, the input diffraction data should be arranged as sequences of diffraction patterns corresponding to consecutive time frames, rather than processing each diffraction image independently. By presenting the data as a temporal sequence, the model can learn the dependencies and correlations between adjacent time points, enabling it to capture the dynamic behavior of the object more effectively.

By embedding knowledge of the diffraction process and object motion into the input transformation, we ensure that the model's learning is guided by both data

and physical principles. This includes normalizing and preprocessing the diffraction data in ways that preserve critical physical information.

3.2.4 Model architecture design incorporating diffraction principles

The design of the model architecture is critically informed by domain knowledge from materials science and the physics of diffraction. Based on the input and output requirements, the model must effectively learn material representations in both spatial and Fourier domains while considering temporal information between adjacent frames. Additionally, it must integrate physical constraints such as the Fourier Transform (FFT) and align with experimentally measured data. Dynamic CXDI involves reconstructing a sequence of object states over time. The model architecture must therefore consider temporal dependencies between adjacent frames to accurately capture the material's dynamic behavior.

To archive that, the model design need to include:

- Temporal Modeling Components: Architectural elements such as recurrent connections, temporal convolutional layers, or attention mechanisms that can model sequential data and capture temporal dependencies.
- Consistency Constraints: Mechanisms that enforce temporal coherence in the reconstructed object, such as loss terms that penalize unrealistic changes between frames or modules that model expected physical motions.
- Fourier Transform Relationship: The model must ensure that the reconstructed object's Fourier Transform matches the experimentally measured diffraction patterns. This can be enforced by incorporating the FFT operation within the model and comparing the computed diffraction patterns with the observed data.
- Physics-Informed Networks: Incorporate physical laws and constraints directly into the network architecture. By embedding equations governing diffraction and material behavior in model design, the model is guided by known physics, reducing the solution space and improving convergence.

Designing the model architecture with domain knowledge at its core enables the model to effectively reconstruct dynamic material structures from diffraction data. By learning representations in both spatial and Fourier domains, considering temporal information, and integrating physical constraints, the model aligns closely with the underlying physics of the problem. This alignment enhances the model's ability to produce accurate, physically meaningful reconstructions that provide valuable insights into the dynamic phenomena within materials.

3.2.5 Designing loss functions and evaluation metrics for image reconstruction

In unsupervised learning for diffraction image reconstruction, designing an appropriate loss function and evaluation metrics is crucial for effective model training and validation. The nature of diffraction data, which follows a Poisson distribution due to the counting statistics of photon detection, necessitates careful consideration in defining the loss function between the computed diffraction patterns and the experimentally measured data. Additionally, incorporating temporal consistency into

the loss function is important for accurately reconstructing dynamic processes over time.

Diffraction measurements are inherently subject to Poisson noise because the detected intensities correspond to photon counts, which are discrete and probabilistic. To account for this, the loss function should reflect the statistical properties of the data. Moreover, to ensure that the reconstructed object maintains temporal coherence across sequential frames, temporal regularization terms are required to design and add to the loss function.

Evaluating the quality of the reconstructed object without ground truth data requires a multifaceted approach that combines quantitative assessments with expert validation.

- Visual Inspection: The direct and straightforward evaluation is from experts reviews about reconstructed images to identify realistic features and artifacts.
- Simulations comparison : If simulations of the expected dynamics are available, compare them with the reconstruction as a test case with known object and behaviors. This will bring the insights about the model design settings and modification need for improving in real scenarios.
- Consistency with Experimental Settings: Ensure that the reconstruction aligns with known experimental conditions, such as applied stresses, size of the materials or temperature changes.

Designing an effective loss function and metric strategy is essential for unsupervised learning in diffraction image reconstruction, especially in the absence of ground truth data. By modeling the Poisson nature of diffraction measurements in the loss function and incorporating temporal consistency terms, the model is guided toward producing physically plausible and temporally coherent reconstructions. The evaluation metrics combine quantitative assessments of data fidelity and temporal consistency with qualitative analyses by experts and cross-validation with experimental settings. This comprehensive approach ensures that the reconstructed images are not only mathematically consistent with the measured data but also meaningful representations of the dynamic phenomena within the material.

Based on all of these design strategies, the subsequent methodology and experiment sections will provide detailed explanations of our unsupervised learning approach for dynamic coherent X-ray diffraction imaging (CXDI). The methodology section will elaborate on the specific implementation of the model architecture, including how we integrate data-driven methods with physics-informed networks, process input transformations, and design loss functions that reflect the Poisson nature of diffraction data while incorporating temporal consistency and physical constraints. The experiment section will showcase the application of our model to real diffraction datasets, demonstrating its effectiveness in reconstructing dynamic material structures and capturing their temporal evolution. By aligning our model design with domain knowledge from materials science and diffraction physics, we aim to achieve accurate, physically consistent reconstructions that offer valuable insights into the dynamic phenomena occurring within materials.

3.3 Methodology

3.3.1 Dynamic phase retrieval in single-shot CXDI

Figure 3.3a illustrates a single-shot CXDI optics system for dynamic object imaging Takazawa et al., 2021; Takazawa et al., 2023. The system uses a monochromatic X-ray beam, which is shaped by a rounded triangular aperture and a Fresnel zone plate (FZP) to generate a rounded triangular X-ray beam illuminating the sample.

To investigate the dynamic behavior of objects included in a sample, we captured the dynamic changes by imaging the sample a total of N consecutive frames, with each frame recorded over an exposure time Δt. The observed diffraction intensity image at the t^{th} frame is denoted as I_t. The objective of phase retrieval in a dynamic CXDI experiment is to inversely derive a unique complex object function O_t from I_t for each t^{th} frame. Typically, the conventional phase-retrieval algorithm iteratively refines the object function O_t based on the reciprocal–space constraint.

In this study, we adopt the mixed-state reconstruction approach (Thibault and Menzel, 2013) that applies a set of illumination probe functions $P^{(k)} \in \mathbb{C}^{m\times m}$ with mode $k \in [1, 2, \cdots, K]$ for single-shot CXDI. Such probe functions are often estimated in advance via scanning CXDI and are assumed to remain constant. Starting from an initial guess O^0, the estimation of the object function in the iteration $(i+1)^{th}$ is expressed as follows:

$$\begin{aligned} O_t^{i+1} &= O_t^i + \alpha U^i \\ &= O_t^i + \alpha \frac{\sum_k \bar{P}^{(k)}(\psi_t'^{(k)} - \psi_t^{(k)})}{\sum_k |P^{(k)}|^2_{\max}}, \end{aligned} \tag{3.5}$$

where $\bar{P}^{(k)}$ is the complex conjugate of $P^{(k)}$. The scalar α is a feedback parameter of the updating object function U and $\psi_t^{(k)} = P^{(k)} \times O_t^i$ is the wavefield of the sample reconstructed from the previous iteration i^{th}. Meanwhile, $\psi_t'^{(k)}$ is the revision of the wavefield $\psi_t^{(k)}$ satisfying the constraint in the reciprocal space:

$$\psi_t'^{(k)} = \mathcal{F}^{-1}\left[\sqrt{I_t}\frac{\Psi_t^{(k)}}{\sqrt{\sum_k |\Psi_t^{(k)}|^2}}\right], \tag{3.6}$$

where $\Psi_t^{(k)} = \mathcal{F}[\psi_t^{(k)}]$. $\mathcal{F}^{-1}$ denotes the inverse Fourier transform operator. Through the iterative process, the refinement algorithm is expected to converge to a solution of the complex object function O_t that corresponds to I_t. However, it should be noted that the diffraction intensity images I_t are observed with noise, and the iterative algorithm and the condition of corresponding I_t may result in multiple solutions for the complex object function O_t.

3.3.2 PID3Net: Physics-informed unsupervised learning framework

Borrowing the concept from the iterative refinement algorithm, this study employed a unsupervised learning strategy to develop a neural network: the Physics-Informed Deep learning Network for Dynamic Diffraction imaging (PID3Net), tailored to phase

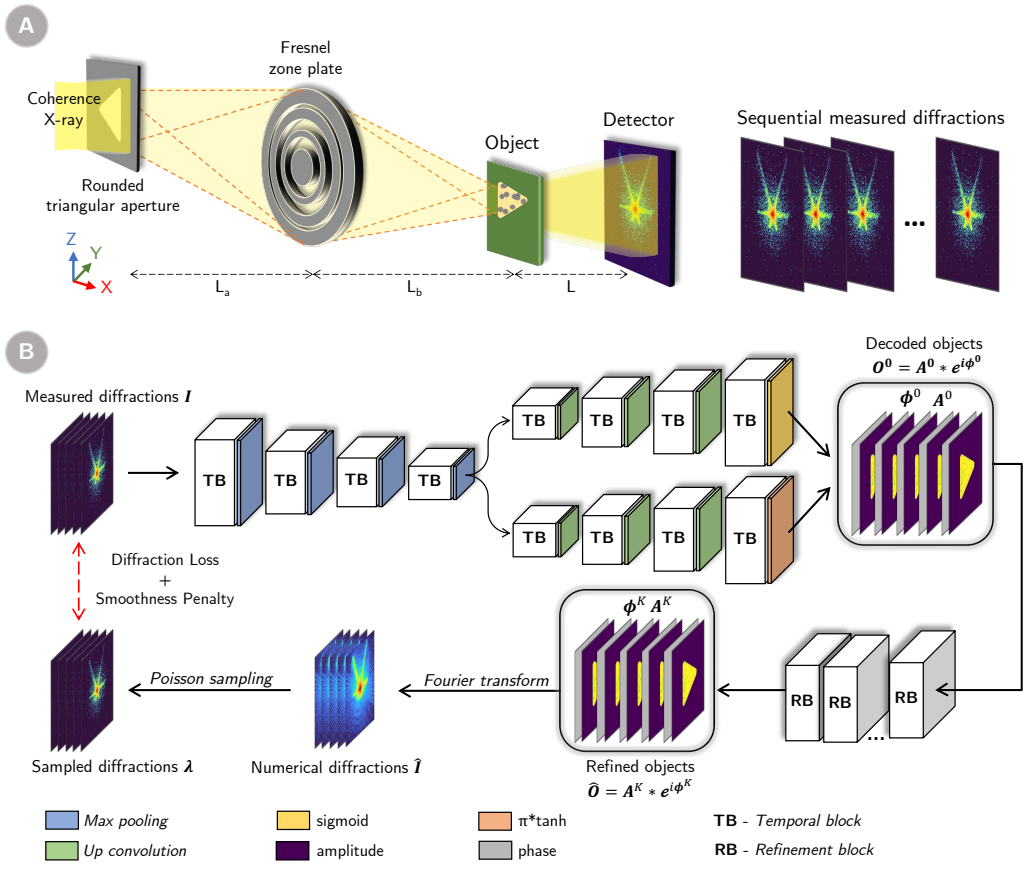


FIGURE 3.3: (a) Schematic diagram of single-shot CXDI optical system with a triangular aperture and a Fresnel zone plate. (b) Overview of the PID3Net framework for dynamic phase retrieval in single-shot CXDI.

retrieval in single-shot CXDI for dynamic object imaging. Figure 3.3b shows the design of the proposed network. Given a sequence of T consecutive diffraction intensity images $\boldsymbol{I} = [I_1, I_2, \ldots, I_T]$, which are extracted from the total observed N consecutive diffraction intensity images, the network was designed to reconstruct simultaneously a corresponding set of T object functions, $\boldsymbol{O} = [O_1, O_2, \ldots, O_T]$. The objective was to ensure that the calculated diffraction images $\hat{\boldsymbol{I}} = [\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_T]$ derived from these object functions closely matched the experimental diffraction images $\boldsymbol{I}$. Consequently, the challenge of phase retrieval in CXDI experiments was formulated as a gradient-descent optimization problem (Tripathi, McNulty, and Shpyrko, 2014) to minimize the discrepancies between two sets of diffraction images, $\boldsymbol{I}$ and $\hat{\boldsymbol{I}}$. This self-supervised learning strategy enabled our method to learn directly from intensity data while avoiding the need for images of the sample and ensuring consistency in the inverse estimation of O_i within T consecutive frames ($i \in [1, 2, \cdots, T]$).

3.3.3 Encoder-Decoder Block for Learning Diffraction Representations

To learn effective representations of diffraction patterns and reconstruct the amplitude and phase representations of the object, we employ an encoder-decoder block. This architecture is designed to facilitate the convergence of phase retrieval in single-shot CXDI for dynamic object imaging by introducing smoothness constraints in both spatial and temporal domains. The smoothness constraints are implemented

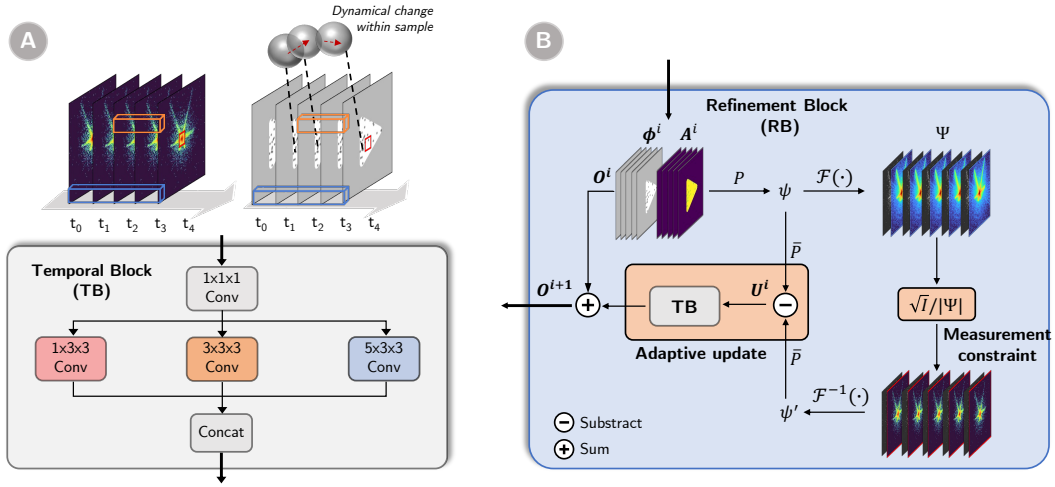


FIGURE 3.4: Designs of constraint-integrated blocks: (a) Temporal Block (TB) with 3D CNN layers for integrating smoothness constraints in both spatial and temporal domains, and (b) Measurement-informed refinement block (RB) for incorporating optical settings and mathematical constraints.

to ensure temporal coherence and physical plausibility in the reconstructed object functions across sequential frames.

To enable spatiotemporal feature learning directly from a sequence of diffraction frames, we adopted the spatiotemporal convolution (Tran et al., 2018 (or 3D-CNN)), which is widely applied for video understanding tasks. This block processes spatiotemporal information by learning features that span both spatial regions and adjacent time frames. The spatiotemporal convolutional block forms the core of the phase retrieval network and is critical for capturing dynamic behaviors in the diffraction data.

PID3Net first inputs the set of diffraction images $\boldsymbol{I}$ to encoder–decoder modules, including one encoder and two decoders (Fig. 3.3b). These modules utilize temporal blocks (TBs), composed of three 3D-CNN layers, to encode temporal information and reconstruct real-space amplitude and phase representations of the object (Fig. 3.4a). The TBs are designed to enhance the temporal coherence of the reconstructed objects across sequential frames.

In the encoder, temporal information is captured at three hierarchical levels: the diffraction image itself, three adjacent diffraction images, and five adjacent diffraction images. This multi-level encoding ensures that temporal dependencies across varying timescales are effectively captured. During decoding, shared temporal information is reconstructed through TBs, enhancing coherence over time. We represented the kernel as F×K×W×W, where F represents the number of filters, and K and W are the kernel sizes for the temporal and spatial spaces, respectively. As shown in Figure 3.4a, one TB consists of three CNN layers: F×1×3×3 + (F/2)×3×3×3 + (F/2)×5×3×3, which accumulate information in both spatial and temporal spaces to encode the input diffraction images. Following each TB layer, a $1 \times 2 \times 2$ max pool layer was applied to reduce the image size by half while preserving the temporal information. The encoded diffraction representation at the last layer of encoder module would be used for reconstructing the object representation in the decoder module.

In the decoding phase, two separate decoder channels reconstruct amplitude and

phase information. These decoders use TB layers similar to those in the encoder and incorporate $1 \times 2 \times 2$ transpose convolution layers for adaptive upsampling. This approach ensures that fine-grained spatial and temporal details are preserved during reconstruction, outperforming simpler interpolation methods. A final $1 \times 3 \times 3$ convolution layer generates the outputs for each decoder channel. For the amplitude and phase outputs, the sigmoid and $\pi \times$ tanh activation functions are used, respectively, to ensure normalized and physically interpretable results. The decoders leverage the encoded diffraction representation to produce sequential real-space amplitude $\boldsymbol{A}^0 = [A_0^0, A_1^0, \cdots, A_T^0]$ and phase $\boldsymbol{\phi}^0 = [\phi_0^0, \phi_1^0, \cdots, \phi_T^0]$ images for the T object functions.

3.3.4 Measurement-Informed Refinement Block: Integrating Physical Constraints into the Model

The phase-retrieval process often causes the problems of twin image ambiguity, translation, and initialization, complicating the convergence of the inversion process Li et al., 2016; Guizar-Sicairos and Fienup, 2012. To address the challenges, we integrated optical settings and mathematical constraints via a second block, the measurement-informed refinement block (RB). The block refines the phase $\boldsymbol{\phi}^0$ and intensity $\boldsymbol{A}^0$ information using a hybrid approach that combines DL methods with the iterative process (Fig. 3.4b).

At each refinement step $(i+1)^{th}$, the amplitude $\boldsymbol{\phi}^i$ and intensity $\boldsymbol{A}^i$ information from the previous step is first combined to complex object functions $\boldsymbol{O}^i = [O_1^i, O_2^i, \ldots, O_T^i]$, in which $O_t^i = A_t^i * e^{\mathrm{i}\phi_t^i}$. The updating information is defined using CNN blocks, by adopting the updating Eq. 3.5 from the iterative phase retrieval algorithm. However, rather than individually updating each object function, we use the TB layer to learn the updates for all T object functions simultaneously, preserving overlapping information shared with other sample reconstructions in real space (Fig. 3.4b). We denote $F_A, F_\phi : \mathbb{R}^{T\times m\times m} \rightarrow \mathbb{R}^{T\times m\times m}$ as the CNN blocks used to determine the revision for the amplitude and phase information. Each network comprises one TB followed by an output convolution layer of size $1 \times 3 \times 3$. Activation functions in F_A and F_ϕ are sigmoid and $\pi*$tanh to limit the amplitude and phase output to $[0, 1]$ and $[-\pi,\pi]$, respectively. The updating process is expressed as follows:

$$\boldsymbol{A}_{\mathrm{u}}^i = F_A(|\boldsymbol{U}^i|), \quad \boldsymbol{\phi}_{\mathrm{u}}^i = F_\phi(\arg(\boldsymbol{U}^i)), \tag{3.7}$$

where $\boldsymbol{U}^i = [U_1^i, U_2^i, \ldots, U_T^i]$ are updating information of the object functions calculated using Eq. 3.5. arg(.) and |.| indicate the argument and modulus of a complex number, respectively. Subsequently, the object functions are updated using mathematical constraints as follows:

$$\boldsymbol{O}^{i+1} = \boldsymbol{O}^i + \alpha(\boldsymbol{A}_{\mathrm{u}}^i * e^{\mathrm{i}\boldsymbol{\phi}_{\mathrm{u}}^i}), \tag{3.8}$$

where α is a learnable parameter of the update rate. Both F_A, F_ϕ, and α undergo adaptive training to regulate the amount of measurement constraint incorporated in $\boldsymbol{O}^{i+1}$.

After K iterations, we obtained the reconstructed object functions $\hat{\boldsymbol{O}}$ $(= \boldsymbol{O}^K)$ and derived the amplitude $\hat{\boldsymbol{A}} = |\hat{\boldsymbol{O}}|$ and phase $\hat{\phi} = \arg(\hat{\boldsymbol{O}})$ from the reconstructed object functions $\hat{\boldsymbol{O}}$.

3.3.5 Loss design and Model training

Within DL approaches, the image reconstruction problem can be solved by minimizing the difference between $\boldsymbol{I}$ and $\hat{\boldsymbol{I}} = |\mathcal{F}[P \times \hat{\boldsymbol{O}}]|^2$, which is quantified via the absolute error loss, expressed as

$$\mathcal{L}_{MAE}(\boldsymbol{I}, \hat{\boldsymbol{I}}) = \frac{1}{Tm^2} \sum_{t,i,j} |I_t[i,j] - \hat{I}_t[i,j]|. \tag{3.9}$$

However, in real-life experiments, diffraction image I_t are captured using photon-counting statistics, meaning the measured diffraction values are non-negative integers. Conversely, the mathematically estimated diffraction $\hat{I}_t$ consists of real values, which may cause numerical issues during loss calculation in optimization. This issue needs to be considered carefully when the signal in the diffraction image is weak due to short exposure time. To address this, we design a loss function that measures the difference in detected photon counts per pixel by employing independent Poisson distributions, which accurately reflect the photon-counting statistics of diffraction measurements (Bian et al., 2016; Chen and Candes, 2015). This loss function ensures a more robust and statistically grounded approach for modeling the experimental data, as follows:

$$\begin{aligned} \mathcal{L}_P(\boldsymbol{I}, \hat{\boldsymbol{I}}) &= -\sum_t Log f_{Poiss}(I_t; \lambda_t) \\ &= -\sum_t (\sum_{i,j} I_t[i,j] \log \lambda_t[i,j] - \lambda_t[i,j]), \end{aligned} \tag{3.10}$$

where λ_t indicates the number of photons detected in each pixel, which is determined by Poisson sampling with the rate $\hat{I}_t$ as shown in Fig. 3.3.

Additionally, to ensure a smoothness constraint in both spatial and temporal domains for the constructed images, we employed a 3D total variation loss, $\mathcal{L}_{TV}$Chambolle, 2004; Takayama et al., 2021a. Herein, $\mathcal{L}_{TV}$ is defined as

$$\begin{aligned} \mathcal{L}_{TV}(\hat{\boldsymbol{O}}) = \frac{1}{3Tm^2} \sum_{t=1}^{T} \sum_{i=1}^{m} \sum_{j=1}^{m} (||\hat{\boldsymbol{O}}_t[i,j] - \hat{\boldsymbol{O}}_t[i+1,j]||_2 + \\ ||\hat{\boldsymbol{O}}_t[i,j] - \hat{\boldsymbol{O}}_t[i,j+1]||_2 + ||\hat{\boldsymbol{O}}_t[i,j] - \hat{\boldsymbol{O}}_{t+1}[i,j]||_2), \end{aligned} \tag{3.11}$$

where $\hat{\boldsymbol{O}}$ is the complex reconstructed objects with the shape $T \times m \times m$.

In summary, our loss function $\mathcal{L}$ comprises two components: the diffraction loss ($\mathcal{L}_{diff}$), which is responsible for minimizing the disparity between the estimated and measured diffraction images, and the smoothness penalty loss ($\mathcal{L}_{TV}$). The total loss function $\mathcal{L}$ is formulated as

$$\mathcal{L} = \mathcal{L}_{diff}(\boldsymbol{I}, \hat{\boldsymbol{I}}) + \beta \mathcal{L}_{TV}(\boldsymbol{O}), \tag{3.12}$$

where the scalar weight β is learned and adjusted during the training for balancing the two loss components and $\mathcal{L}_{diff}$ is either $\mathcal{L}_{MAE}$ or $\mathcal{L}_{PO}$.

In the case of phase retrieval, PID3Net serves as a unsupervised learning model that learns to minimize the difference between the measured diffraction image and intensity image estimated by the network. The consistency of the reconstructions with the observed diffraction datasets during phase retrieval was monitored using

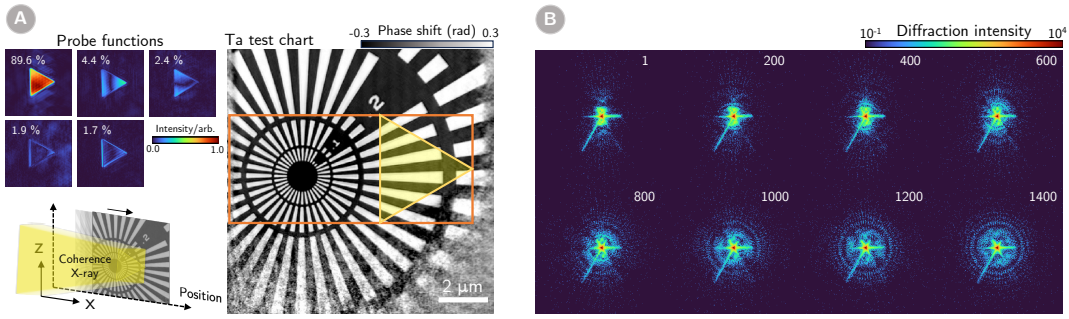


FIGURE 3.5: (a) Schematic of the evaluation experiment for imaging a moving Ta test chart. The chart is horizontally moved against the fixed X-ray, and the orange region indicates the illuminated area of the sample. The diffraction intensity images of the moving Ta test chart were captured on-the-fly using the experimental optical system with five modes of probe function. These five probe modes were reconstructed using the scanning CXDI. (b) Measured diffraction intensity images of the moving Ta test chart at eight different frames. The frame index of each image is indicated in the upper row. The color bar at the top represents the diffraction intensity.

the R_f factor, as indicated below

$$R_f = \frac{\sum_N \sum_{m^2} ||\Psi_t| - \sqrt{I_t}|}{\sum_N \sum_{m^2} |\sqrt{I_t}|}, \tag{3.13}$$

where N is the number of diffraction images in the datasets.

The model was trained with the Adam optimizer on an individual Tesla A100-PCIe graphics processing unit with a memory capacity of 40 GB. The learning rate was set to 0.001, and the batch size was set to eight. The network was trained for 20 and 50 epochs for the test chart and AuNP datasets, respectively. Remarkably, less than one million parameters were used in PID3Net. In this study, the number of sequence images for learning (T) was set to five for all the datasets. The number of iteration updates in the RB layer was set at five for the performance–speed trade-off. For the iterative reconstruction methods, the number of iterations was set such that the R_f score was saturated and did not decrease.

3.3.6 Experimental design

We conducted three evaluation experiments to demonstrate the efficacy of PID3Net for phase retrieval in single-shot CXDI for depicting dynamic process of the sample. These experiments involved capturing diffraction intensity images of the sample using an experimental optic system (Fig. 3.5a) or through numerical simulations under similar optical conditions. In the first evaluation experiment, we examined the efficacy of PID3Net in imaging a moving Ta test chart, commonly used as a proof-of-concept sample for evaluating optic systems or phase retrieval analysis. In this experiment, the Ta test chart was moving at a fixed velocity of 340 nm s^{-1} (Figure 3.5a) relative to the X-ray beam. The diffraction intensity images of the moving Ta test chart were captured on-the-fly using the experimental optic system. PID3Net was then used to reconstruct the movement of the sample from the measured diffraction intensity images.

The two following evaluation experiments examined the efficacy of PID3Net in imaging the dynamics of gold nanoparticles (AuNPs) dispersed in solution. Observing the motion of AuNPs over a broad spatiotemporal scale is essential, as they are widely used to probe the mechanical properties of materials and the rheological properties within living cells. In the first evaluation experiment of AuNPs, we evaluated the efficacy of PID3Net using simulations of moving AuNPs, providing both amplitude and phase data along with corresponding diffraction images. This allowed us to directly compare the phase information retrieved by PID3Net with the ground-truth data, facilitating an accurate assessment of our method's efficacy. In the second evaluation experiment of AuNPs, we employed a sample of AuNPs dispersed in aqueous polyvinyl alcohol (PVA) solution. We evaluated the efficacy of PID3Net in the real experimental scenario when the AuNPs were moving in the solution. We expected that the movement of the AuNPs would be captured using single-shot CXDI by measuring the diffraction images directly with the experimental optic system.

We assessed the efficacy of PID3Net by comparing it with a conventional method (Kang et al., 2021) that uses reciprocal-space constraints and gradient descent, specifically extended to a mixed-state model (Thibault and Menzel, 2013) where multiple probe functions are applied in optical systems. The multi-frame PIE-TV (mf-PIE) (Takayama et al., 2021a), which introduces virtual overlapping frames constraint for solving dynamic CXDI, was also applied for comparison. Hereinafter, we denote these two methods as mixed-state and mf-PIE. Besides the iterative methods, we compared our method with two state-of-the-art DL-based methods for phase retrieval, including the AutoPhaseNN (Yao et al., 2022a) and PtychoNN (Cherukara et al., 2020). Additionally, we investigated the effectiveness of the Poisson loss $\mathcal{L}_P$ and the MAE loss $\mathcal{L}_{MAE}$ when applied in PID3Net for phase retrieval, denoting the models as PID3Net-PO and PID3Net-MAE for the respective loss functions used. The efficacy of these methods was evaluated on the basis of three critical criteria: 1) the discrepancy between the diffraction images reproduced from the reconstructed sample images and measured diffraction images, 2) the spatial resolution of the retrieved phase information, and 3) the efficiency of the retrieved phase information for post analysis, in terms of applicability and accuracy.

To quantify the discrepancies between diffraction images, we used the R_f score (Miao et al., 2006), which provides a single scalar value indicative of the overall pixel-by-pixel difference between the measured diffraction images and those calculated from the reconstructed sample images. The R_f score is crucial for assessing how well the reconstructed sample images reproduce the measured diffraction images on average. To evaluate the spatial resolution of the retrieved phase information, we employed the phase retrieval transfer function (PRTF) (Chapman et al., 2006; Sekiguchi, Oroguchi, and Nakasako, 2016) or Fourier ring correlation (FRC) (Wakonig et al., 2019). This is crucial for determining the resolutions at which phase retrieval remains accurate and reliable. The PRTF assessed the fidelity of phase retrieval by comparing diffraction images calculated from reconstructed sample images with those measured across various spatial frequencies. In contrast, the FRC, applicable when ground-truth values are available, employed Fourier transforms of the reconstructed and actual images of the sample to evaluate discrepancies across spatial frequencies.

Additionally, the efficiency of the retrieved phase information was quantified using the structural similarity index measure (SSIM) (Takayama et al., 2021b; Hore and Ziou, 2010) when the ground truth of the sample was available in the simulation

scenario. The SSIM provided a single scalar value; however, it indicated the similarity between the retrieved phase information and those calculated from the ground truth of the sample. To quantitatively evaluate the efficiency of our method in real experimental scenarios, where images of the sample are reconstructed from diffraction images measured during actual X-ray diffraction experiments, we verified the fidelity of the reconstructed pattern on the sample by referencing ground-truth information about the samples and their moving behaviors. The ground-truth information included the shape and moving velocity of the sample in the experiment with the Ta test chart and the distinguishability between particles and the solution in the experiment with AuNPs dispersed in PVA solution (Takazawa et al., 2023).

3.4 Case study 1: Phase retrieval for movement of Ta test chart

In this section, we discuss the first proof-of-concept experiment that was conducted to evaluate the performance of PID3Net in phase retrieval for movement of the Ta test chart (Fig. 3.6a). Before performing the single-shot CXDI measurements, the phase image of the Ta test chart and probe function were reconstructed using the mixed-state method (Takazawa et al., 2021) via scanning CXDI with an exposure time of 10 s at each scan position. The diffraction images measured with a 7 ms exposure time per frame in the moving experiment are presented in Fig. 3.5b for some steps from the first to the 1400^{th} frames. Table B.1 presents the detailed settings of the three experiments.

Figure 3.6a presents the retrieved phase information from the diffraction images by using the mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO. At first glance, the patterns and line absorber shapes of the Ta test chart in the phase images reconstructed by PID3Net-PO and PID3Net-MAE are more precise and more stable over time than those retrieved by mixed-state or mf-PIE, as further illustrated in Appendix Movie C1. The phase information retrieved using AutoPhaseNN and PtychoNN is also shown in this movie and Appendix Figure B.1; however, the resolution of the reconstructed images is not as clear and stable as with PID3Net-PO and PID3Net-MAE.

Next, to evaluate the spatial resolution of the reconstructed image obtained by the phase retrieval methods, profiles of two circular arcs on the Ta test chart were analyzed for each frame. Figure 3.6b shows the enlarged view of the phase information retrieved from the measured diffraction images at the 400^{th} frame, highlighting the two circular arcs. The red circular arc crosses over 100 nm-width patterns, while the blue circular arc crosses over 200 nm-width patterns. These patterns include transmitted and absorbed regions, represented by areas with positive and negative phase shifts, respectively. Figure 3.6c shows plots of the retrieved phase information by the four applied methods along the two circular arcs. In the Ta test chart, the patterns are organized cyclically along the circular arcs, with equal widths along each arc but different widths between the two arcs for both the absorbed and transmitted regions. Both PID3Net-MAE and PID3Net-PO successfully reconstructed the patterns that the two circular arcs pass through, accurately reflecting the ground-truth shape of the Ta test chart. In contrast, the patterns are not clearly observed in the phase information retrieved by using the mixed-state and mf-PIE methods. Similar to the iterative methods, the DL methods, such as AutoPhaseNN and PtychoNN, did not reconstruct the patterns accurately, as shown in Appendix Figure B.2 b. Additionally, PID3Net-PO yielded a more stable and smoother cycle of phase transition than

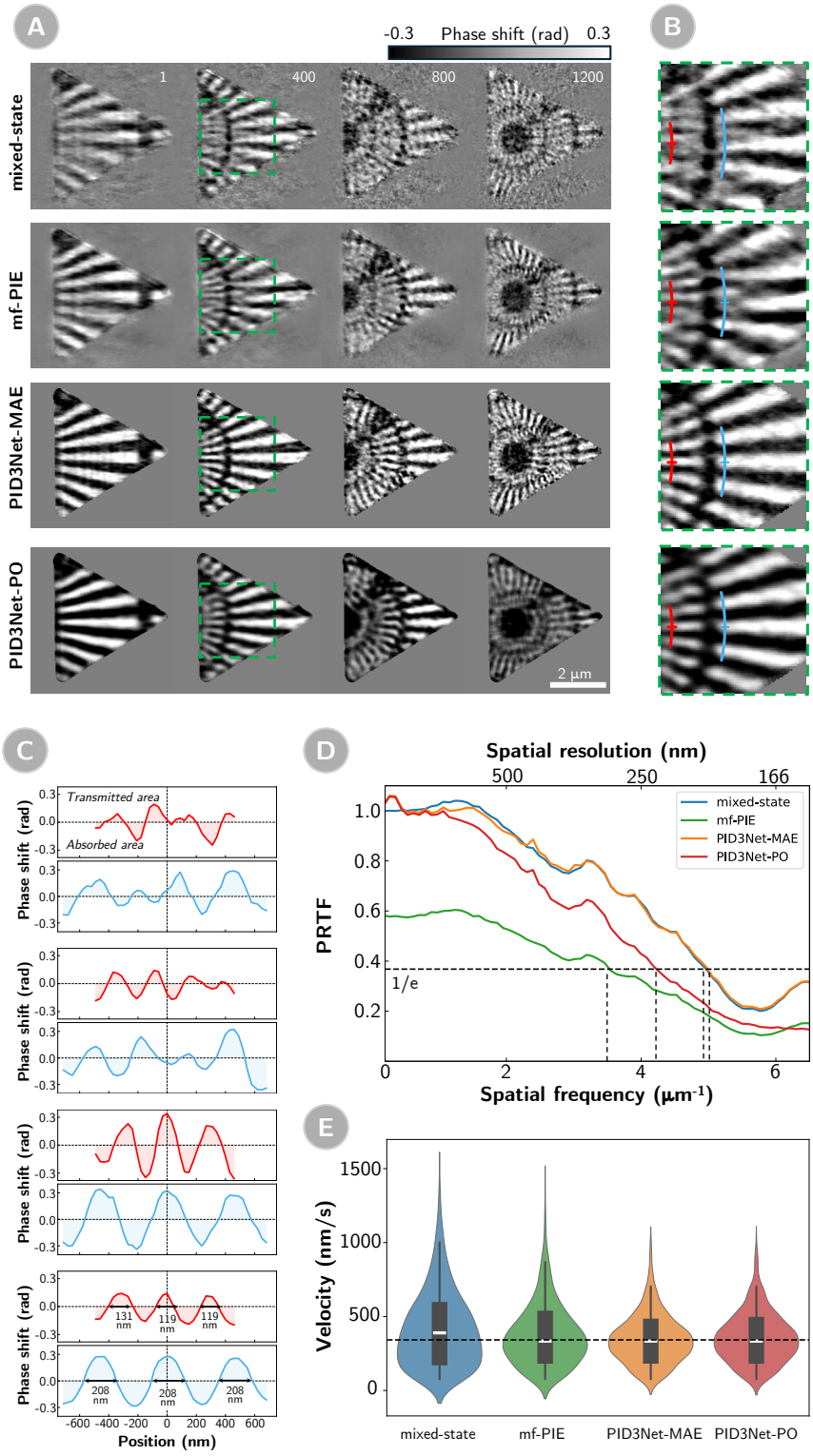


FIGURE 3.6: Phase information was retrieved from diffraction intensity images of a moving Ta test chart using four methods: mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO, with a 7 ms exposure time per frame. (a) The frame index for each image is shown at the top. (b) Magnified views of green square areas at frame 400, including profiles of two circular arcs with zero position markers. (c) Analysis of phase shifts along these arcs from the 400th frame diffraction image. (d) PRTF analysis of phase images from the four retrieval methods, with dashed lines indicating reliability thresholds. (e) Estimated velocity distribution from phase images over 400 frames, with a dashed line at 340 nm/s and a white bar indicating the median.

PID3Net-MAE. The estimated widths of the patterns reconstructed using PID3Net-PO are approximately 120 nm and 210 nm for the red and blue curved lines, respectively, aligning well with the widths of these patterns in the Ta test chart. The obtained results quantitatively demonstrate that PID3Net-PO and PID3Net-MAE can retrieve phase information with at least twice higher spatial resolution compared to the iterative methods such as mixed-state and mf-PIE methods, as well as DL methods like AutoPhaseNN and PtychoNN.

As an attempt to evaluate the reliability of the proposed methods in scenarios where ground-truth information of the sample is unavailable, unlike the scenario with the Ta test chart, we assessed how accurately the reconstructed sample images could reproduce the observed diffraction images. This assessment is crucial for determining the practical applicability of the proposed methods. The discrepancies on average between diffraction images reproduced from phase information retrieved using the mixed-state method and the measured diffraction images are the smallest, as evidenced by an R_f score of 0.84. The scores for PID3Net-MAE and PID3Net-PO are almost the same, at 0.84 and 0.85, respectively. Meanwhile, the results of other DL-based methods are slightly higher, at 0.87 for both the AutoPhaseNNand the PtychoNN. The diffraction images reproduced from phase information obtained via the mf-PIE method exhibit the largest discrepancies, with an R_f score of 0.91.

Figure 3.6d presents the PRTF indices, which assess the fidelity of phase retrieval across spatial frequencies, for the four methods, mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO. The PRTF indices of the mixed-state method and PID3Net-MAE are comparable and outperform the remaining methods at all spatial resolutions. In this context, the full-period spatial resolution of the mixed-state method and PID3Net-MAE is up to approximately 190 nm. The results of PID3Net-PO are at approximately 230 nm and are slightly worse than those of the mixed-state method and PID3Net-MAE. Conversely, the worst spatial resolution threshold is observed for mf-PIE, at approximately 270 nm, suggesting less reliable phase retrieval at these finer resolutions. Results of similar analyses for phase images reconstructed using the DL methods are presented in Appendix Figure B.2c. The threshold for PtychoNN is slightly worse than those of PID3Net-PO, while the number for AutoPhaseNN is slightly better and is up to approximately 200 nm.

The analysis of the R_f scores and PRTF indices indicates that the diffraction images obtained using the mixed-state method closely resemble the actual measured diffraction images compared to other phase retrieval methods we considered. Our method shows comparable results to the mixed-state method and performs slightly better than other DL-based methods. In addition to finding the solutions that match the measured diffraction images, the phase retrieval methods also need to reconstruct phase information that accurately depicts the dynamic behaviors of the sample. In our method, the temporal block (TB) is specifically designed to simultaneously retrieve phase information of adjacent frames from a sequence of diffraction images, ensuring the dynamic behaviors of the sample are reliably depicted. This approach contrasts with traditional methods that individually retrieve phase information from a single diffraction image.

Further investigations were performed to evaluate the efficiency of the phase information retrieved from our proposed methods in depicting the moving behaviors of the Ta test chart. We employed auto-correlation methods on pairs of adjacent phase images retrieved using the six methods, mixed-state, mf-PIE, AutoPhaseNN, PtychoNN, PID3Net-MAE, and PID3Net-PO, and monitored the number of shifted pixels to estimate the velocity of the movement of the Ta test chart. As shown in Fig. 3.6e, the averages of estimated velocities derived from these phase retrieval

methods are close to the actual velocity settings of the sample in the experiment. In contrast, the numbers derived from DL methods were lower than the actual velocity, as shown in Appendix Figure B.2d. Notably, the estimated instantaneous velocities by using PID3Net-MAE and PID3Net-PO during the experiment are more stable, with smaller variances, than those estimated by using the mixed-state and mf-PIE methods. The results indicate that the PID3Net method adeptly learns to reconstruct phase information that matches the dynamics behaviors of the sample, where the Ta test chart is moving horizontally relative to the X-ray beam from left to right at a fixed velocity.

TABLE 3.1: Time reconstruction of each methods for Test chart dataset 7ms with 1755 images.

Methods	Iterations (epochs)	Run time (s)	Inference time (s/image)
mixed-state (CPU)	100	7,830	4.45
mf-PIE (CPU)	200	55,970	31.90
PtychoNN (GPU)	20	194	0.001
AutoPhaseNN (GPU)	20	436	0.001
PID3Net-NR-P (GPU)	20	505	0.002
PID3Net-MAE (GPU)	20	1,745	0.003
PID3Net-PO (GPU)	20	1,760	0.003

The previous evaluation results have shown that our proposed method can reconstruct high-quality images for the moving Ta test chart by adding constraints to ensure smoothness in both spatial and temporal domains. These high-quality reconstructions may be attributed to the deep learning method's capacity to capture the underlying mechanisms of the sample's dynamic behaviors, thereby addressing both the issue of random noise in observed diffraction images and the issue of multiple solutions in phase retrieval. To test the capacity of the proposed methods, we simulated a hypothetical movement of the Ta test chart using the obtained experimental data by rearranging the order of the consecutive measured diffraction images. In this test, the Ta test chart repeatedly moved from left to right over 400 frames and then moved back to the left over the subsequent 200 frames instead of moving in only one direction as in the actual experiment. The PID3Net-PO model, which was trained on the experimental dataset featuring only one direction of movement, successfully retrieved high-resolution images for the simulated movement of the Ta test chart with changes in the direction of movement (Appendix Movie C2) without requiring retraining. This finding suggests that the DL method holds promise in capturing the dynamic behaviors of the sample moving horizontally relative to the X-ray beam, and thus producing high-quality image reconstructions. However, further detailed studies are required to fully assess this characteristic of the proposed method.

In summary, our method yielded high-quality images that accurately captured the movement of the Ta test chart, surpassing the iterative and DL methods. The iterative methods somewhat captured the dynamic behavior of the moving Ta test chart but resulted in lower-quality images, while the other DL methods failed to

TABLE 3.2: Comparative evaluation of PID3Net-MAE, PID3Net-PO and two other phase retrieval methods using R_f and SSIM values. A lower R_f score indicates that reconstructed sample images appropriately reproduce the measured diffraction images on average, whereas a higher SSIM score indicates that the reconstructed sample images closely resemble the actual samples. The bold numbers indicate the best values among the four methods.

	R_f score	SSIM score
mixed-state Kang et al., 2021	0.73	0.55
mf-PIETakayama et al., 2021a	0.82	0.81
PID3Net-MAE	**0.71**	0.92
PID3Net-PO	0.72	**0.95**

portray the sample's dynamic behavior accurately. Consequently, in the forthcoming evaluation experiments, we compared our method with only two iterative methods: mixed-state and mf-PIE. Besides accurately reconstructing high-quality images of the Ta test chart, our method is less time-intensive than iterative methods. As shown in Table 3.1, our method is four times faster than the mixed-state method and thirty-two times faster than the mf-PIE method in training, while the inference time of our method is thousands of times faster, making it highly suitable for practical deployment. More details about the comparison are presented in the section Appendix B.

3.5 Case Study 2: Phase retrieval for simulated movement of AuNP in solution

In the second evaluation experiment, we examined the efficacy of PID3Net in retrieving phase information for the motion of AuNPs in solution from simulated diffraction images. Starting from such numerical simulations, which provide a controlled setting as well as known ground truth, will facilitate an accurate assessment of our method's efficacy before applying it to the actual experimental scenario. Table B.2 presents the detailed settings of the second experiments.

Figure 3.7a shows the schematic of the simulation for the AuNPs dispersed in the solution and a simulated diffraction image. This simulation used four modes of the numerical probe function, which were reconstructed using scanning CXDITakazawa et al., 2023. Figure 3.7b presents the retrieved phase information from the simulated diffraction images using the mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO. The PID3Net method was superior to other methods, particularly in the contrast between the AuNPs and the background. This superior contrast led to a more accurate capture of the shape of particles from the retrieved phase information.

At first glance, the distinction between AuNPs and solutions in the phase images reconstructed by PID3Net-PO and PID3Net-MAE appears more precise and more stable over time compared with those retrieved by mixed-state or mf-PIE, as further illustrated in Appendix Movie C3. In particular, PID3Net-PO and PID3Net-MAE can retrieve finer details and more rounded edges, which match better with the ground truth of the simulation, and are crucial for interpreting and analyzing the

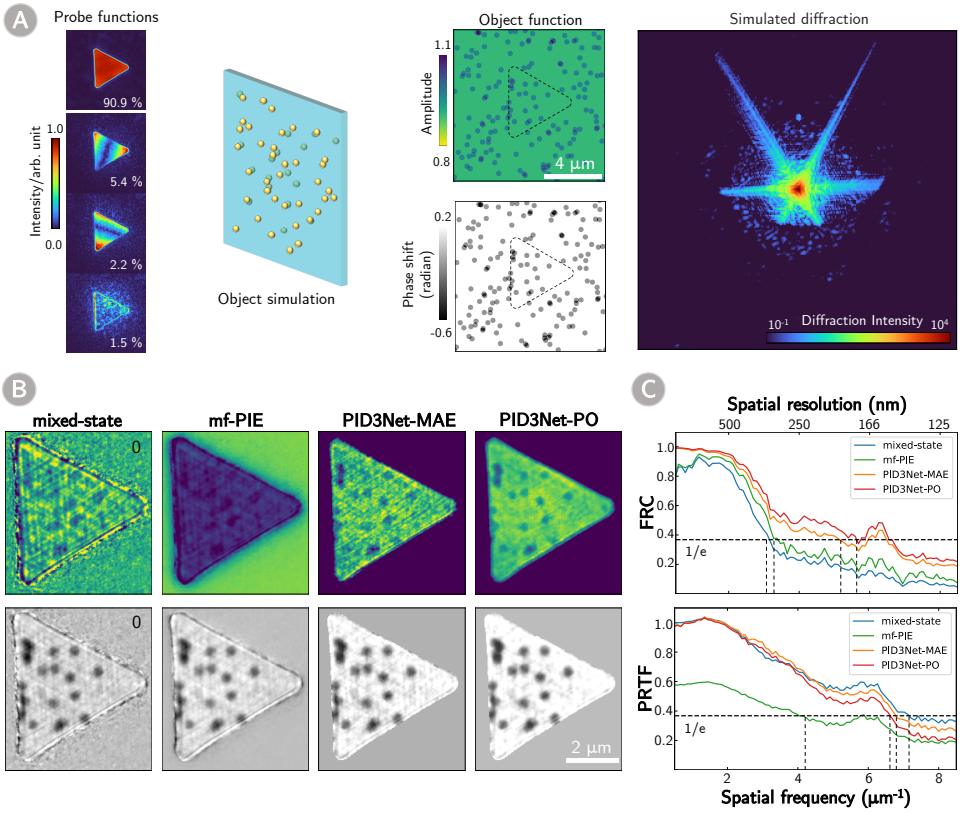


FIGURE 3.7: (a) Schematic of the simulation for the AuNPs dispersed in the solution and a simulated diffraction image. The simulated diffraction images are accumulated with a 100 ms exposure time and four-mode probe functions. The four-mode probe functions were reconstructed using scanning CXDI. (b) Amplitude (upper) and phase (below) information in the first frame reconstructed using the mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO methods. (c) The Fourier ring correlation (FRC) and the phase retrieval transfer function (PRTF) analysis of the phase images reconstructed using the four phase retrieval methods. The dashed line indicates a threshold value of $1/e$ and the corresponding spatial frequency.

moving behaviors of the AuNPs. Additionally, PID3Net-PO shows potential in accurately capturing amplitude information, which is essential for characterizing the intensity distribution of the sample. However, reconstructing amplitude for short exposure times in practice remains challenging for all the methods.

Table 3.2 presents a quantitative evaluation of phase retrieval accuracies for the four methods by considering the differences in both diffraction and phase images. The PID3Net-MAE method achieves the lowest R_f score of 0.71 among considered methods, indicating that the diffraction images it calculates are closest to the simulated diffraction images. However, the PID3Net-PO method, which achieved a slightly higher R_f score, reconstructed phase information most closely to the ground truth of the simulation, as evidenced by the highest SSIM score of 0.95, surpassing those of all other methods. The mixed-state exhibits a comparable R_f score with those of our methods, but the phase reconstructions achieved using this method lack accuracy compared with the ground truth of the simulation, as evidenced by its significantly lower SSIM score. Conversely, the worst R_f score is observed for mf-PIE, at 0.82, but its SSIM score is much better than the number of the mixed-state. The

obtained results again suggest that achieving better R_f does not guarantee better-quality phase information due to the multiple solutions issue.

Furthermore, the spatial resolution of retrieved phase images was statistically assessed using both the FRC and PRTF analysis methods (Fig. 3.7c). The FRC for phase information of PID3Net-MAE and PID3Net-PO outperformed those of other methods across all spatial resolutions. Notably, the spatial resolutions of the diffraction images reproduced from the retrieved phase information by PID3Net-MAE and PID3Net-PO were up to 185 nm and 150 nm. In comparison, the reliable spatial resolutions for mixed-state and mf-PIE were limited to 305 nm and 315 nm, respectively. However, regarding PRTF analysis, the mixed-state and PID3Net-MAE have slightly better spatial resolution at approximately 135nm. Meanwhile, the PID3Net-PO and mf-PIE results are up to 150 nm and 250 nm, respectively.

In summary, these statistical results underscore the efficacy of the PID3Net method in accurately retrieving phase information for the motion of AuNP in simulations with high spatial resolution, compared with the other two iterative phase retrieval methods. The promising outcomes suggest the potential applicability of the PID3Net method in real experimental scenarios, where diffraction images are measured using an experimental optic system instead of simulations. Additionally, this experiment shows that while the R_f and PRTF metrics provide insights into the reconstruction consistency on the basis of the observed diffraction images, the SSIM and FRC metrics offers a more nuanced perspective on the fidelity of the reconstructed image. This distinction underscores the importance of considering both metrics while evaluating phase retrieval methods.

3.6 Case study 3: Phase retrieval for experimental movement of AuNPs in the PVA solution

Previous evaluation experiment using numerical simulations of samples with nanoparticle motions have theoretically demonstrated the efficacy of the PID3Net method in phase retrieval for dynamic behaviors of AuNPs in solution. Statistical analyses on ground-truth images or prior knowledge about the AuNPs' dynamics, such as known velocities, showed that the PID3Net method enables high-resolution image reconstructions and aids in understanding the underlying mechanisms of motion of AuNPs in solution. In the third experiment, we used the proposed method to reconstruct the image of an actual sample with motions of colloidal AuNPs Takazawa et al., 2023 in a 4.5 wt% polyvinyl alcohol solution, with particle sizes of approximately 150 nm, from its measured diffraction images to assess the efficacy of the PID3Net method. Table B.3 presents the detailed settings of the third experiments.

Figure 3.8a shows reconstructed images from one-second-exposure diffraction images across five frames, using mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO. The images reconstructed by mixed-state and mf-PIE appear blurred and noisy, suggesting less effective particle position and contrast resolution. In contrast, PID3Net-PO clearly reproduces the particles' positions and contrasts, which results in smoother estimations and markedly reduced noise. However, phase recovery using PID3Net-MAE is difficult, yielding images with diminished contrast and increased blurriness. This issue is attributed to its reliance on the MAE calculation that averages pixel losses across the image, potentially smoothing out critical details. For a more detailed visualization of these results, refer to Appendix Movie C4. The evaluated R_f scores for mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO are 0.33, 0.62, 0.49, and 0.53, respectively. Moreover, the PRTF scores for mixed-state and PID3Net-MAE

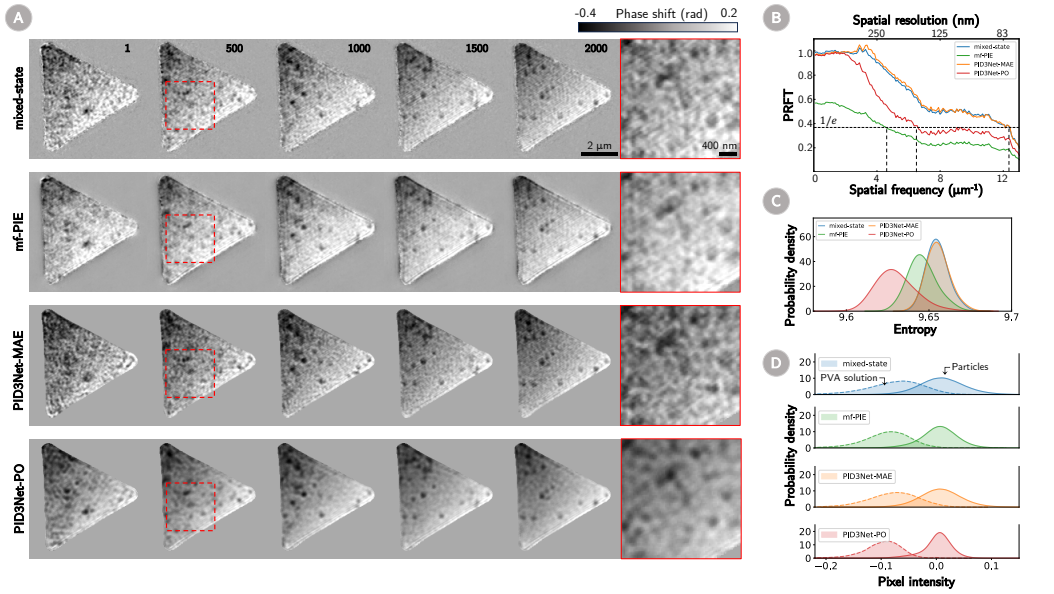


FIGURE 3.8: (a) Phase information were retrieved from measured diffraction images for the AuNPs dispersed in the PVA solution with a one-second exposure time per frame. These phase images are retrieved using mixed-state, mf-PIE, PID3Net-MAE, and PID3Net-PO. The rightmost images show zoomed-in views of the areas enclosed by the red squares in the 2000^{th} frame. (b) The phase retrieval transfer function (PRTF) analysis of the phase images reconstructed using the four phase retrieval methods. The dashed line indicates a threshold of $1/e$ and the corresponding spatial frequency. (c) Distributions of entropies of phase images reconstructed using the four methods. (d) Pixel intensity distributions of the particle and solution patterns in reconstructed phase images.

are comparable and outperform the other two methods at various spatial resolutions (Fig. 3.8b). The limits of full-period spatial resolution for mixed-state and PID3Net-MAE are approximately 80 nm, whereas for PID3Net-PO and mf-PIE, the limits extend to 160 and 220 nm, respectively.

Further investigations were conducted to quantitatively evaluate the efficiency of the phase information retrieved from our proposed methods for capturing the movement of the AuNPs in the PVA solution. Figure 3.8c shows the mean entropy of all the reconstructed images by using each method. Notably, PID3Net-PO demonstrates lower entropy values than PID3Net-MAE, and significantly lower entropy values than mixed-state and mf-PIE, indicating enhanced contrast in the images reconstructed by and superior noise-reduction capabilities of PID3Net-PO.

Additionally, an adaptive threshold filter (Takazawa et al., 2023) was employed to assess the performance of each method in supporting accurate gold particle detection. With consistent settings for differentiating between the signal from particles and those from the noise and the background (solution), the images reconstructed by PID3Net-PO distinctly show a significant enhancement in distinguishing particle and PVA solution distributions, as illustrated in Figure 3.8d. Moreover, Moreover, Figure 3.9 presents a comparison between PID3Net-PO and the mixed-State method in terms of particle detection and the effectiveness of tracking their movement over time. The results demonstrate that PID3Net-PO not only provides superior spatial

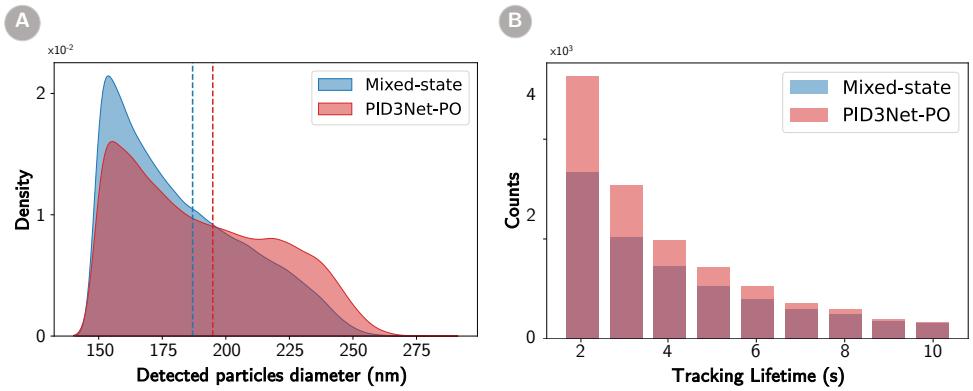


FIGURE 3.9: (a) Distribution diameter of single particles detected in both methods (b) The lifetime tracking of particles through dynamic CXDI.

resolution but also more effectively captures the dynamic behavior of the nanoparticles. Notably, the size of the detected particles in the PID3Net-PO reconstructions closely matches the sizes specified during fabrication and consistent with experimental settings. This agreement confirms the model's ability to accurately resolve particles at the expected scale, which is crucial for reliable analysis and validation of experimental results. The accurate sizing of particles ensures that quantitative measurements, such as particle diameter and volume, are trustworthy and can be used confidently in subsequent analyses.

Moreover, the number of detected particles is significantly higher when using PID3Net-PO compared to other reconstruction methods. This increase can be attributed to the model's improved capability to distinguish particles from the noise background and the PVA solution. This is particularly evident in the improved ability to track particle movements across consecutive frames, highlighting the model's capability in handling dynamic imaging scenarios. By leveraging the advanced reconstruction capabilities of PID3Net-PO, we achieved more accurate and reliable detection of gold nanoparticles, which is crucial for applications requiring precise characterization of particle distributions and dynamics. The integration of the adaptive threshold filter further enhanced the robustness of our analysis by minimizing the influence of noise and background signals. This ensured that the observed improvements in particle detection and tracking are attributable to the reconstruction method itself rather than variations in post-processing techniques.

Although further analyses are required to fully comprehend the nanoscale dynamics of each specific material, the results obtained from all three evaluation experiments underscore the efficacy of the PID3Net method. The temporal self-consistent learning approach employed in both Fourier and real spaces offers crucial support for improving phase-retrieval quality, providing valuable insights for further development of phase-retrieval methods in CXDI to depict dynamic process of the sample.

3.7 Contributions and limitations

PID3Net represents DL architectures specifically engineered to address the phase retrieval challenges in CXDI for depicting dynamic process of the sample. This

network leverages self-supervised learning strategy to reconstruct complex structures of the sample from sequential diffraction images, achieving exceptional performance. The strength of PID3Net lies in its application of 3D convolutional layers to the temporal sequences of diffraction images, which enables the model to learn diffraction images representations self-consistently. This capability enables PID3Net to effectively capture intricate features present in experimental data. Additionally, PID3Net integrates these learned representations to perform inverse reconstructions of sample images via deep dynamic diffraction imaging. This comprehensive approach facilitates accurate predictions and enables the generation of high-fidelity reconstructions, even in scenarios marked by dynamic variations and significant noise.

However, limitations associated with DL models compared with traditional iterative reconstruction methods should be acknowledged. Although iterative methods can effectively work with a single data point, DL models typically require substantial data for training and convergence. The availability and quality of training data, complexity of the imaging system, and computational resources required for training are crucial factors for consideration. Despite these challenges, the successful implementation of PID3Net has prompted advancements in the field of coherent X-ray imaging.

Future studies could focus on several directions to overcome the existing limitations of DL models. One promising direction is the integration of physical models and prior knowledge into the DL framework to reduce the dependency on extensive training datasets. Additionally, advancements in transfer learning could allow models trained on the data of one type material to be adapted for use with different types of materials, improving the versatility and applicability of DL methods in CXDI (Zhang et al., 2021b). Furthermore, integrating numerical simulation and self-supervised learning DL in a reinforcement learning manner could enhance the model's ability to learn and adapt to dynamic changes in real-time, significantly improving phase retrieval performance and robustness.

PID3Net is a valuable tool for various scientific and industrial applications owing to its ability to generate accurate reconstructions in real time and efficiently process dynamic variation positions. The development of PID3Net will enhance our understanding of material systems at the nanoscale, potentially transforming the landscape of nanomaterial research and development.

Chapter 4

Deep Learning Framework Design for Supervised Representation Learning in Material Property Prediction

4.1 Introduction

The main task of materials science consists of a combination of empirical knowledge and theoretical approaches to study the composition and structure of materials with specific properties. It is also necessary to confirm these materials experimentally, which is rather time-consuming and often depends on serendipity. To overcome such difficulties, it was the rapid growth of materials informatics (MI) as an interdisciplinary solution. MI extracts valuable information regarding materials and their physicochemical behaviors from experimental and computational data using data-driven techniques, hence accelerating the discovery and development process of superior materials. (Agrawal and Choudhary, 2016; Ramprasad et al., 2017; Butler et al., 2018; Siriwardane et al., 2022).

The majority of materials informatics (MI) approaches consist of three essential components Ward and Wolverton, 2017. The first component includes datasets that provide details about the structures of materials, measurement outcomes directly associated with these structures, and physical properties that align with material development objectives. The second component, known as representation, quantitatively describes the data from the first component, offering a fundamental characterization of materials for identification and analogical reasoning. The final component involves a system that utilizes machine learning or data mining algorithms—individually or in combination—to extract knowledge from the materials datasets for specific objectives, such as predicting properties or discovering new material compositions and structures.

Traditionally, materials have been defined by their elemental compositions and structural configurations. Researchers have primarily depended on their expertise and experience—often called tacit knowledge—to anticipate the properties of hypothetical materials with specific compositions and structures. Computational chemistry methods grounded in quantum mechanics and exceptionally dense functional theory (DFT) simulations enable the theoretical verification of these compositions and structures through *in-silico* experiments. However, despite their ability to provide accurate information on the physical properties of hypothetical materials, computational experiments have certain limitations. For instance, the vast number

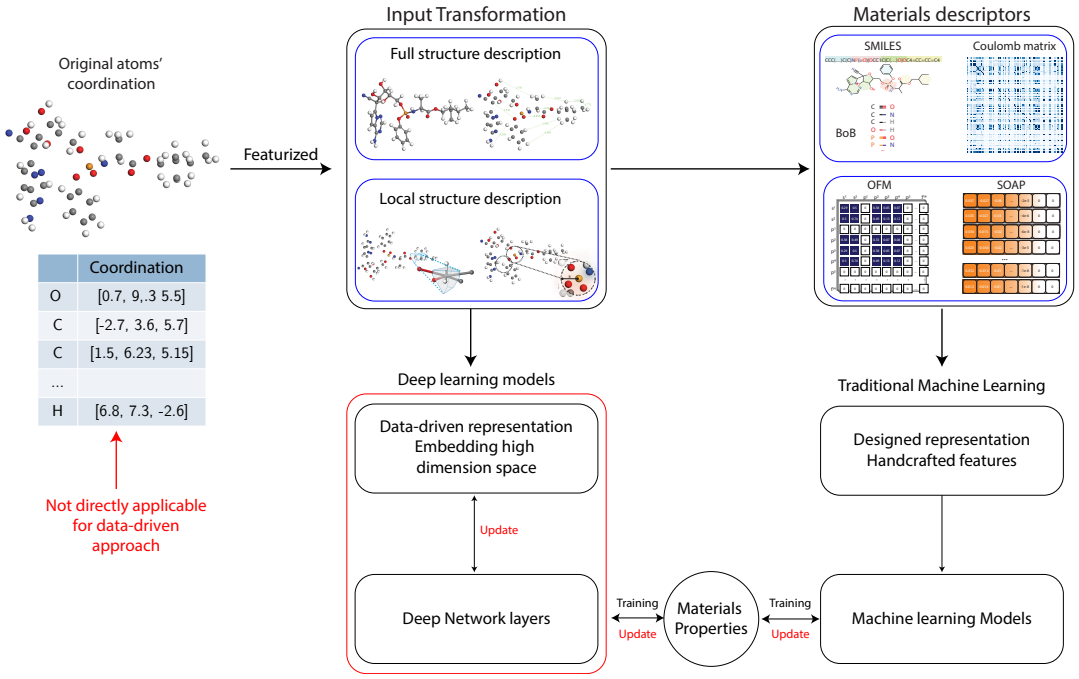


FIGURE 4.1: Illustrations of approaches for representing material structure and learning property.

of potential hypothetical materials makes designing materials with desired physical properties time-consuming and expensive due to the extensive calculations required. Furthermore, researchers need specialized and detailed knowledge to narrow down the potential compositions and material structures effectively.

Unlike traditional methods, materials informatics (MI) approaches begin by transforming basic data descriptions into suitable representations that facilitate mathematical reasoning and inference, as illustrated in Fig. 4.1. Specifically, MI systems are tasked with estimating both qualitative and quantitative relationships between materials based on these transformed representations, enabling the discovery of potential patterns within the material data Ramakrishnan et al., 2014; Himanen et al., 2019; Zhao et al., 2023. Developing material representations—such as designing material descriptors or developing methods to learn representations from data—is critical in MI approaches. The effectiveness of an MI algorithm significantly depends on the quality of the material representation, directly influencing the algorithm's performance and aiding in the explanation and interpretation of inference processes and prediction outcomes Rupp et al., 2012. Recent advancements in automated experimentation and high-performance computing have facilitated the acquisition of vast experimental and computational data. As a result, there is an increasing demand for developing explainable and interpretable MI methods to deepen our understanding of physical and chemical phenomena.

Recently, various deep learning (DL)-based materials informatics (MI) approaches have been developed to tackle challenges related to material representation and the prediction of physical properties Schütt et al., 2014; Karamad et al., 2020; Xie and Grossman, 2018; Rahaman and Gagliardi, 2020. A typical example is the DL architecture that incorporates a continuous-filter convolution layer with filter-generation networks, enabling the handling of atomistic systems and the accurate prediction of properties for both molecular and crystalline materials Schütt et al., 2014. Another noteworthy approach utilizes convolutional neural networks based on crystal

graphs, which can predict material properties with accuracy comparable to density functional theory (DFT) calculations while also providing atomic-level chemical insights Xie and Grossman, 2018. Beyond these methods, researchers have developed various other DL architectures designed to encode the local chemical environments of atoms and enhance prediction accuracy. These works have been achieved by integrating different material descriptors, applying graph neural networks (GNNs), and utilizing many-body tensor representations Karamad et al., 2020; Rahaman and Gagliardi, 2020. Additionally, several studies have incorporated prior knowledge into neural network models to ensure that the relationships between material structures and their properties are learned with high fidelity Anderson, Hy, and Kondor, 2019; Fuchs et al., 2020; Vaswani et al., 2017.

However, interpretability is a crucial challenge for traditional and DL-based machine learning methods. Most machine learning models try to include all the available information rather than selecting interpretable representations to enhance the accuracy of the predictions. The nature of the relationship between material representations and their properties is complex and nonlinear; therefore, the machine learning model works like a "black box," with no explicit correlations being manifested. Although the statistical evaluation based on existing data often reveals very high prediction accuracies, assessing their predictive performance for new materials is difficult. Moreover, a comprehensive understanding of the physicochemical phenomena by machine learning in order to shed light on these underlying processes remains a challenge.

Numerous studies have sought to improve model interpretability by integrating additional information or features. For example, graph convolutional networks utilize SMILES strings to represent molecules as inputs, which aids in identifying essential fingerprint fragments and facilitates interpretation Duvenaud et al., 2015; Wu et al., 2018. Despite these advancements, these networks still need help in accurately predicting the properties of molecular and crystalline materials due to the lack of 3D structural information. Message-passing neural network-based models (MPNNs) Fung et al., 2021; Gilmer et al., 2017; Yang et al., 2019 incorporate heuristic bonding information to capture atomic interactions but encounter several issues, including handling long-range interactions, ensuring feature interpretability, representing global information, and maintaining scalability when processing large molecule or crystal datasets. To address these limitations, recent research has shifted towards transformer-based networks Fuchs et al., 2020; Chen et al., 2022b; Cao et al., 2023; Kang et al., 2023; Korolev and Protsenko, 2023; Das et al., 2023; Gunning et al., 2019; Moran et al., 2023, which leverage attention mechanisms Vaswani et al., 2017. These networks present a promising approach by modeling interatomic interactions through attention scores, which reflect the importance of each atom in learning the representations of other atoms. Subsequently, various pooling methods, such as max or average pooling Pham et al., 2019; Schütt et al., 2018; Moran et al., 2023; Fuchs et al., 2020; Anderson, Hy, and Kondor, 2019; Wu et al., 2021b; Schweidtmann et al., 2023; Xu et al., 2019, are employed to generate a comprehensive representation of the entire structure. However, extracting meaningful structure-property relationships from these transformer-based networks remains a challenging and complex task.

4.2 Integrating Domain Knowledge into Model Design Strategies

Developing advanced machine learning models for materials science necessitates metic harmonizing domain-specific knowledge with computational techniques. This integration is pivotal because materials science encompasses complex phenomena governed by the laws of chemistry and physics. By embedding chemical and physical insights into the model design, we can significantly enhance predictive accuracy and interpretability. Such embedding ensures that the models are mathematically robust and deeply aligned with the fundamental principles governing material behavior. This alignment facilitates models that can generalize better, offer meaningful insights, and ultimately contribute to accelerated material discovery and innovation.

4.2.1 Designing Output Targets for Material Property Prediction

The ultimate objective in materials modeling is to accurately predict material properties and gain insights into the underlying structure-property relationships. Achieving this dual goal requires the development of highly predictive and inherently interpretable models. To realize this, we first define our desired outputs:

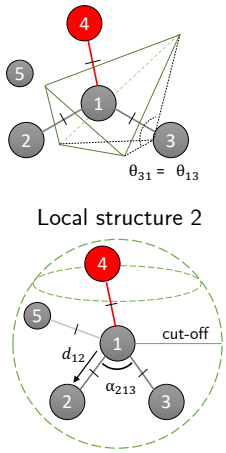
- Precise Property Predictions: The model should reliably and accurately predict material properties, enabling the identification and characterization of materials with desired functionalities.
- Meaningful Interpretations of Structure-Property Relationships: The model should provide insights into how specific structural features contribute to material properties, highlighting the regions or components within a material that are most influential.

While structure-property relationships in materials are inherently complex and span a wide range of aspects—including electronic structure, bonding, defects, and microstructure—we narrow down our focus to make the problem tractable within a computational and deep learning (DL) framework. By setting up the second output to qualitatively estimate the contribution of different regions within the material's structure to its properties, we create a pathway for the model to identify and emphasize these critical areas.

4.2.2 Designing material structures transformations based on output

Understanding that material properties are intrinsically linked to atomic and molecular structures, we prioritize input representations that capture essential chemical and physical characteristics. This involves selecting features that reflect the local atomic environments and their interactions, which are critical determinants of material properties. To address this, we propose using representations built upon local structures. By focusing on local atomic environments, we can more accurately model the interactions that govern material properties. Local structures encapsulate information about an atom's immediate neighbors, bond lengths, bond angles, and coordination geometry, which are fundamental to understanding chemical bonding and physical interactions.

Building input representations from local structures involves constructing descriptors that capture the geometry and chemistry of an atom's neighborhood. This approach aligns with the chemical intuition that an atom's properties are influenced by its local environment.



Features	Distance-based	Voronoi-based	Angle-based
	d_{ij}	d_{ij} , θ_{ij}	d_{ij} , α_{jik}
N-body interaction	2-body	2-body	3-body
Computational cost with N neighbor	$\mathcal{O}(N)$	$\mathcal{O}(N)$	$\mathcal{O}(N^2)$
Local structure 1 description	$d_{12}, d_{13}, d_{14}, d_{15}$	d_{12}, d_{13}, d_{14} θ_{12} , θ_{13}, θ_{14}	$d_{12}, d_{13}, d_{14}, d_{15}$ $\alpha_{213}, \alpha_{214}, \alpha_{215},$ $\alpha_{314}, \alpha_{315}, \alpha_{415}$
Local structure 2 description	$d_{12}, d_{13}, d_{14}, d_{15}$	d_{12}, d_{13}, d_{14} $\theta_{12}, \theta_{13}, \theta_{14}$	$d_{12}, d_{13}, d_{14}, d_{15}$ $\alpha_{213}, \alpha_{214}, \alpha_{215},$ $\alpha_{314}, \alpha_{315}, \alpha_{415}$
Identity between local structures	✗	✓	✓

FIGURE 4.2: Comparison of the descriptive ability of local structure representation methods: (1) Distance-based descriptors, emphasizing pairwise atomic distances; (2) Voronoi-based descriptors, focusing on spatial partitioning and local atomic environments; and (3) Angle-based descriptors, capturing angular relationships to highlight geometric orientation.

- Local Atomic Environments: For each atom in the material, we consider its neighboring atoms within certain conditions. This neighborhood defines the local atomic environment, characterized by interatomic distances, angles, and the types of neighboring atoms.
- Voronoi Tessellation: We utilize Voronoi tessellation to partition space and define local environments more naturally. This method divides space into regions based on the proximity to atoms, resulting in Voronoi polyhedra representing each atom's spatial influence. Voronoi-based descriptors effectively capture atoms' spatial arrangement and connectivity, which are crucial for modeling material properties as shown in Fig 4.2.
- Distance-Based and Angle-Based Descriptors: We incorporate descriptors that quantify interatomic distances and Voronoi solid angles. These geometric features are essential for characterizing the local geometry and understanding how atomic arrangements influence material properties.
- Chemical Feature Embeddings: We embed chemical information like atomic numbers into the input representations. This embedding allows the model to consider geometric and chemical factors when learning structure-property relationships.

By constructing input representations emphasizing local structures with domain specific knowledge, we provide the model with rich, detailed information about the material's atomic-scale configuration. This approach enhances the model's ability to learn meaningful patterns and relationships grounded in chemical and physical reality.

4.2.3 Model architecture design incorporating Materials science principles

We began designing our model architecture by grounding our approach in fundamental materials science principles. Recognizing that multiple-scale interactions determine material properties—from atomic-level interactions to collective behaviors—we aimed to create a model that mirrored this hierarchical nature. By integrating domain knowledge into the architectural design, we ensured that the model predicted material properties accurately and provided interpretable insights into the underlying structure-property relationships.

- Local structure representation: The model design should first focus on capturing the interactions within local atomic environments. This involves learning representations that encapsulate the essential features of an atom and its immediate neighbors, such as bonding patterns and geometric configurations. By doing so, the model aligns with the chemical intuition that local environments are foundational to material behavior.
- Global structure repsentation: Simultaneously, the model must incorporate a mechanism to understand how these local structures interact and contribute to the material as a whole. Drawing from domain knowledge, we recognize that not all local structures equally impact a material's properties; certain regions or configurations may play a more pivotal role. By designing the model to weigh the importance of different local structures when forming a global representation, we ensure that significant features are emphasized, mirroring the material's true physical and chemical priorities.

Essentially, aligning the model architecture with domain knowledge involves creating a system reflecting the materials' hierarchical and interconnected nature. By capturing detailed local interactions and effectively integrating them into a coherent global understanding, the model can more accurately predict properties and provide meaningful insights within the context of materials science. This design strategy leverages fundamental principles to guide architectural choices, ensuring that the model is not just a computational tool but also a representation of the material's intrinsic characteristics.

4.2.4 Designing loss functions and evaluation metrics for accurate property prediction

In developing a supervised learning model for materials science, designing the loss function and learning strategy in alignment with domain knowledge is essential to ensure both accurate predictions and meaningful interpretations. The choice of target properties, the formulation of the loss function, and the evaluation metrics all play pivotal roles in this process.

The selection of target properties for prediction is guided by understanding which material properties can be meaningfully interpreted within the model's framework. Not all properties are equally suitable for interpretation based on local and global structural features. For instance, properties like electronic band gaps, formation energies, or mechanical strengths are directly influenced by atomic configurations and are thus appropriate targets. These properties have well-established correlations with structural features, making them conducive to models that aim to interpret

structure-property relationships. We ensure the model's predictions and interpretations are meaningful and relevant by focusing on properties where domain knowledge indicates a structure-property solid relationship.

The loss function is a critical component that directs learning by quantifying the discrepancy between the model's predictions and the actual values. The Mean Absolute Error (MAE) is often the primary loss function for the supervised learning model. MAE measures the average magnitude of errors without considering their direction, making it a straightforward and interpretable metric.

While the loss function addresses prediction accuracy, evaluating the model's interpretability requires additional metrics that involve domain expertise and physical validation. Interpretation in materials science often necessitates confirming that the model's internal representations and outputs align with known physical laws and chemical principles.

- Expert Knowledge Evaluation: Scientists and engineers assess the model's interpretative outputs—such as feature importances—to determine if they correspond with an established understanding of material behavior.
- Physical and Mathematical Confirmation: To validate the found knowledge, the model's interpretations are compared against theoretical calculations, simulations, or experimental data. For example, if the model highlights specific atomic configurations as critical, this should be consistent with known structure-property relationships.

By involving domain experts and leveraging physical and mathematical analyses, we ensure that the interpretability metrics are grounded in reality and provide meaningful insights.

Integrating domain knowledge into model design strategies is crucial for advancing machine learning applications in materials science. By starting with the desired outputs—accurate property predictions and meaningful interpretations of structure-property relationships—we inform our choices for input representations, model architecture, and learning strategies. We create robust and interpretable models by embedding domain knowledge consistently at every stage of the model development. These models serve as powerful tools for property prediction and contribute to a deeper understanding of structure-property relationships, ultimately accelerating innovation and discovery in materials science. The specific details and implementation of the model—including input representations, architectural components, and learning strategies—are presented in the following methodology section.

Based on all these desin strategies, this study presents an interpretable DL architecture that integrates attention mechanisms to predict material structural properties and elucidate the relationships between structure and properties. The proposed architecture begins by learning representations of local atomic structures within a material through the recursive application of attention mechanisms to the surrounding atoms. The overall material structure representation is then derived from these localized atomic representations. This architecture employs attention mechanisms to incorporate information about the geometric configurations of neighboring atoms into the local structure representations. Furthermore, it quantitatively assesses the degree of attention each local structure receives from a global perspective when forming the material structure representation. By training the model with a specific target property, our approach facilitates the interpretation of the structure-property relationships in materials.

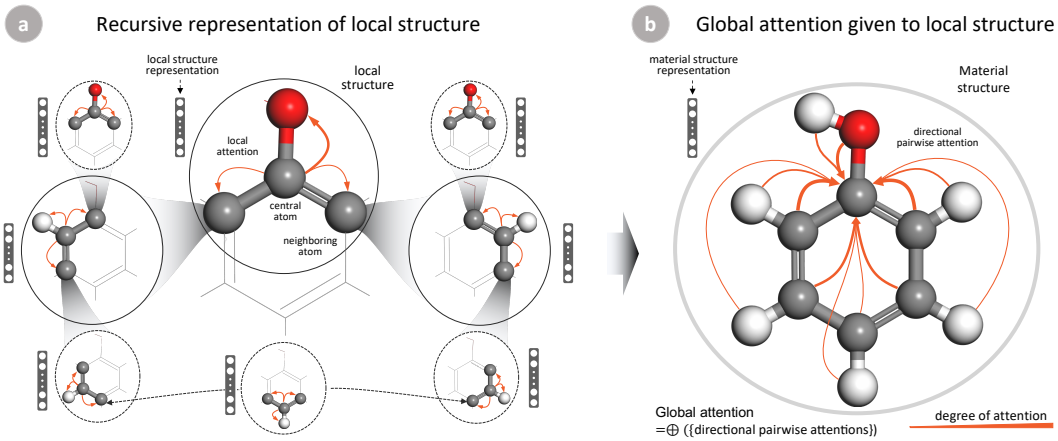


FIGURE 4.3: Illustrations of representations for local structure and material structure. Schematics of (a) the learning recursive representation of a local structure (central atom and its neighboring atoms) within the molecular structure of phenol (C_6H_5OH), and (b) measurement of the global attention given to a local structure when determining representation of the molecular structure.

4.3 Methodology

We introduce a DL architecture named the self-consistent attention neural network (SCANN). SCANN is designed to represent material structures by focusing on the local atomic structures and assigning learned weights to them, thereby enabling both the prediction and interpretation of material properties. The primary objective of SCANN is to recursively learn consistent representations of these local atomic structures within a material, as illustrated in Fig. 1a. These local representations are then appropriately combined to form an overall representation of the material's structure. A detailed overview of the proposed SCANN architecture is provided in Figure 4.3.

4.3.1 Characterization of material structure

In this study, each material structure S in a dataset $\mathcal{D}$ is represented using the coordinates of M atoms ($\mathcal{A}_S = \{a_1, a_2, \cdots, a_M\}$) and the corresponding of its atomic numbers Z. The atomic number Z of an atom a_i ($1 \leq i \leq M$) is considered and represented using a v-dimensional embedding vector $\mathbf{e}_i$ ($\mathbf{e}_i \in \mathbb{R}^v$). Next, a linear function $F_e : \mathbb{R}^v \rightarrow \mathbb{R}^h$ is learned to project this information to provide a better representation of atom a_i. Consequently, atom a_i is represented by an h-dimensional vector $\mathbf{c}_i^0 = F_e(\mathbf{e}_i)$. The vector $\mathbf{c}_i^0 \in \mathbb{R}^h$ is used as the initialization of local structure for the local attention layers computation in the next steps. Hereinafter, we denote the matrix $\mathbf{C}^0 = [\mathbf{c}_i^0]_{1 \leq i \leq M}$ as $[\mathbf{c}_i^0]_{1 \leq i \leq M} = [\mathbf{c}_1^0, \mathbf{c}_2^0, ..., \mathbf{c}_M^0]$.

Each local structure consists of a central atom, its neighboring atoms, and their arrangement around the central atom. To determine the neighboring atoms and segment each material structure into local structures, we employ the definition of O'Keeffe (O'Keeffe, 1979; Pham et al., 2017) instead of the assumption about chemical bonds between the atoms in the structure. According to O'Keeffe's definition, all atoms at these atomic sites share Voronoi polyhedron faces with the atomic site of an atom under consideration (the central atom of the local structure) and are regarded as neighboring atoms. Subsequently, the local structures of the neighboring

atoms are referred to as the neighboring local structures. By incorporating the information from the Voronoi polyhedron faces, we assess the geometrical influences of neighboring atoms on the central atoms for conveying the structural information of structure S to SCANN for learning the appropriate representation of S. The Voronoi method is employed to accurately identify neighboring atoms within a local structure, leveraging material domain knowledge and aligning with the logical framework of our approach, as illustrated in Fig. 4.2.

For each atom a_i in the structure S, the Voronoi tessellation is utilized to determine $\mathcal{N}_i \subset \mathsf{A}_S$, which contains N atoms whose atomic sites share Voronoi polyhedron faces with an atomic site of a_i. Subsequently, the geometrical influence of a neighboring atom $a_j \in \mathcal{N}_i$ on atom a_i is represented by a vector $\mathbf{g}_{ij} \in \mathbb{R}^h$. This vector is defined by combining the Euclidean distance d_{ij} (Å) and Voronoi solid angle $\theta_{ij} \in [0, 4\pi]$ information between the atoms (Pham et al., 2017). The Euclidean distance d_{ij} is expanded by using k Gaussian basis functions $\phi_i(x) = \exp(-(x-\mu_k^d)/2\sigma^2)$ located at each centers 0 Å $< \mu_k^d < d_t$ Å for every $\sigma = 0.5$ Å (Chen et al., 2019; Schütt et al., 2018). Next, the distance embedding layers is defined as shown below:

$$\mathbf{DE}(d_{ij}) = F_d([\phi_1(d_{ij}), \phi_2(d_{ij}), \ldots, \phi_k(d_{ij})]) \tag{4.1}$$

$$\tag{4.2}$$

where $F_d : \mathbb{R}^k \to \mathbb{R}^h$ is the fully-connected layer with a Swish activation function (Ramachandran, Zoph, and Le, 2017). Subsequently, the geometrical influence $\mathbf{g}_{ij}^0 \in \mathbb{R}^h$ is defined as follows:

$$\mathbf{g}_{ij}^0 = \mathbf{DE}(d_{ij}) \times \frac{\theta_{ij}}{\max(\theta_{ik})} \tag{4.3}$$

A comprehensive depiction of the proposed SCANN architecture is presented in Figure 4.4.

4.3.2 Local attention layers

The SCANN architecture comprises a series of L local attention layers, each utilizing attention mechanisms (Vaswani et al., 2017) to represent the local structures within a material structure. The SCANN, with the design of multiple layers of local attention, could iteratively learn and enhance the consistency of local structure representations, thereby providing insights regarding long-range interactions between these local structures. At the $(l+1)^{th}$ local attention layer, the representation $\mathbf{c}_i^{l+1} \in \mathbb{R}^h$ of local structure $\{a_i, \mathcal{N}_i\}$ is derived from the representation vectors in the preceding layer of itself ($\mathbf{c}_i^l$), its N neighboring local structures ($\mathbf{C}_{\mathcal{N}_i}^l = [\mathbf{c}_j^l]_{a_j \in \mathcal{N}_i}$), and the geometrical influence ($\mathbf{G}_{\mathcal{N}_i}^l = [\mathbf{g}_{ij}^l]_{a_j \in \mathcal{N}_i}$) as follows:

$$\begin{aligned}
\mathbf{c}_i^{l+1} &= \text{LocalAttention}^{l+1}(\mathbf{c}_i^l, \mathbf{C}_{\mathcal{N}_i}^l \times \mathbf{G}_{\mathcal{N}_i}^l) \\
&= Attention(\mathbf{q}_i^l, \mathbf{K}_{\mathcal{N}_i}^l) + \mathbf{q}_i^l \\
&= softmax({\mathbf{q}_i^l}^\top \mathbf{K}_{\mathcal{N}_i}^l)\mathbf{K}_{\mathcal{N}_i}^l + \mathbf{q}_i^l \\
&= \sum_{a_j \in \mathcal{N}_i} \alpha_{ij}^l \mathbf{k}_j^l + \mathbf{q}_i^l,
\end{aligned} \tag{4.4}$$

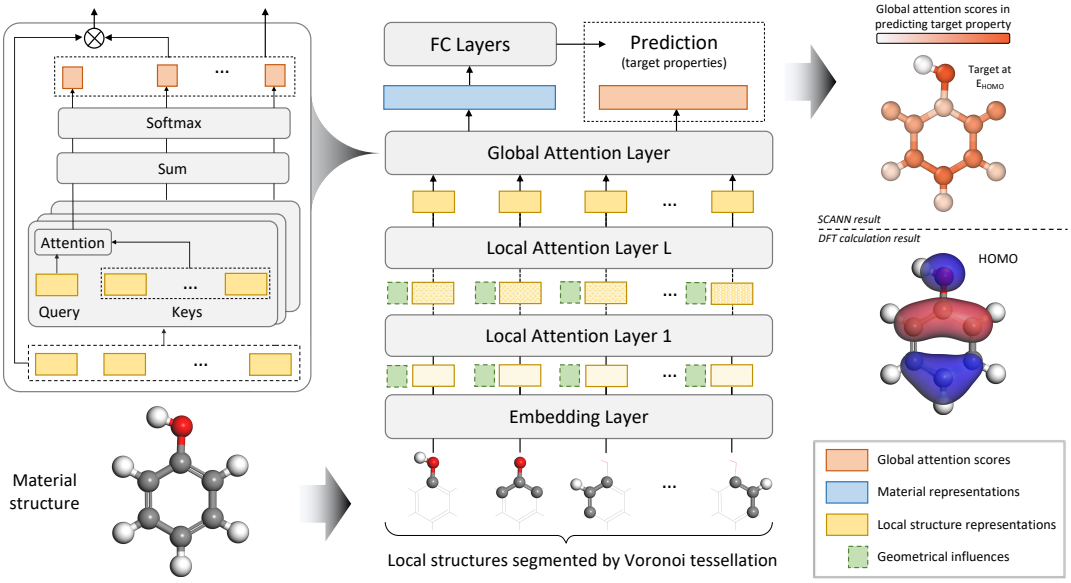


FIGURE 4.4: Overview of the proposed SCANN architecture. SCANN combines an embedding layer and local attention layers to learn representations of local structures. A global attention layer assigns attention scores to these structures, guiding their contribution to the material's representation. Fully connected layers (FC) use this representation to predict material properties.

where $\mathbf{K}^l_{\mathcal{N}_i} = [\mathbf{C}^l_{\mathcal{N}_i} \times \mathbf{G}^l_{\mathcal{N}_i}]\mathbf{W}^l_k = [\mathbf{k}^l_1, \mathbf{k}^l_2, ..., \mathbf{k}^l_N]$ and $\mathbf{q}^l_i = \mathbf{c}^l_i \mathbf{W}^l_q$ ($\mathbf{k}^l_j, \mathbf{q}^l_i \in \mathbb{R}^h$) are the degree each atom a_j influencing the atom a_i and degree atom a_i accepting the influence from the neighboring atoms, respectively. In addition, $\mathbf{W}^l_q$ and $\mathbf{W}^l_k \in \mathbb{R}^{h \times h}$ are the trainable weight parameters. The local attention score α^l_{ij} can be interpreted as the degree of attention to which the information of the local structure centered at atom a_j should be referred to appropriately represent the local structure centered at atom a_i:

$$\alpha^l_{ij} = \frac{e^{s_{ij}}}{\sum_{a_k \in \mathcal{N}_i} e^{s_{ik}}}, \quad s_{ij} = {\mathbf{q}^l_i}^\top \mathbf{k}^l_j. \tag{4.5}$$

The local attention layers are learned to ensure that the neighboring atoms a_j exerting a more significant impact on the central atom a_i have higher s_{ij} scores. Herein, a softmax function is used to normalize these scores to the interval $[0, 1]$ to obtain the attention scores α^l_{ij}. The sum of the attention scores of the neighboring atoms a_j is 1 ($\sum_{a_j \in \mathcal{N}_i} \alpha^l_{ij} = 1$). Consequently, we employed the standard attention mechanism Vaswani et al., 2017 to create the representation for the local structure as follows.

$$\mathbf{c}^{l+1}_i = \text{LayerNorm}(F_n(\mathbf{c}^{l+1}_i) + \mathbf{c}^{l+1}_i), \tag{4.6}$$

where $F_n : \mathbb{R}^h \to \mathbb{R}^h$ is a fully-connected layer with a Swish activation function, and LayerNorm is the layer normalization (**Ba2016layer**).

By employing multiple local attention layers, the attention information related to a target property within a material structure S can through the attention interactions among neighboring local structures. In our evaluation experiments, SCANN models utilize L local attention layers, with the number of layers L being adjusted to optimize performance for each specific dataset. Consequently, we preserve the structural

information of S from the representations of all its local structures obtained from the final local attention layer, to produce $\mathbf{C}^L$, where $\mathbf{C}^L = [\mathbf{c}_i^L]_{a_i \in \mathcal{A}_S}$.

4.3.3 Material structure representation

Previous research has typically represented a material structure by aggregating its local structures using operations such as summation or pooling. However, these methods either assume that all local structures contribute equally (as in sum and average pooling) Pham et al., 2019; Schütt et al., 2018; Moran et al., 2023; Fuchs et al., 2020; Anderson, Hy, and Kondor, 2019 or concentrate on a single specific local structure (as in max and min pooling) Wu et al., 2021b; Schweidtmann et al., 2023; Xu et al., 2019. These approaches can hinder the clear understanding of structure–property relationships. To overcome this limitation, SCANN represents a material structure as a linear combination of the representation vectors of its local structures, where the global attention (GA) scores of each local structure act as the coefficients.

SCANN again utilizes the dot-product key-query attention (Vaswani et al., 2017) to coherently learn the representation of local structures and integrate them into the representation of material structure in a target-dependent manner. We define $\mathbf{Q}^g = \mathbf{C}^L\mathbf{W}_\mathbf{q}^g = [\mathbf{q}_1^g, \mathbf{q}_2^g, ..., \mathbf{q}_M^g]$ and $\mathbf{K}^g = \mathbf{C}^L\mathbf{W}_\mathbf{k}^g = [\mathbf{k}_1^g, \mathbf{k}_2^g, ..., \mathbf{k}_M^g]$ as query and key matrix, where $\mathbf{W}_\mathbf{k}^g, \mathbf{W}_\mathbf{q}^g \in \mathbb{R}^{h \times h}$ are learnable parameters of the global-attention layer. Then, we compute the attention matrix $\mathbf{A} = \mathbf{Q}^g\mathbf{K}^{g\top} \in \mathbb{R}^{M \times M}$ in the same manner as previous local attention layers.

When representing the material structure S, we propose that local structures receiving higher cumulative attention scores from other local structures should be given priority. Consequently, the degree of attention to a local structure $\{a_i, \mathcal{N}_i\}$ in S is quantitatively modeled by summing all the attention it receives from other atoms. We show here the details for calculating the Global attention (GA) score of local structure $\{a_i, \mathcal{N}_i\}$ as the sum of the attention from all query vectors $\mathbf{q}_j^g$ $(j \neq i)$ to the key vector $\mathbf{k}_i^g$ as shown below:

$$\mathrm{s}_i = \sum_{j=1}^{M} [\mathbf{A}(1 - \mathbf{I})]_{j,i} = \sum_{j,j \neq i}^{M} \mathbf{q}_j^{g\top} \mathbf{k}_i^g, \tag{4.7}$$

where I is denoted for the identity matrix. In practice, these scores s_i are normalized by using the ℓ_2 normalization to prevent the sum of them from becoming extremely high in structures that contain a significant number of atoms, as follows:

$$\hat{\mathrm{s}}_i = \frac{\mathrm{s}_i}{||[\mathrm{s}_1, \cdots, \mathrm{s}_M]||_2} \tag{4.8}$$

The function $\rho(.)$ is designed based on the hypothesis that significant attention should be given to a local structure if its representation is crucial for accurately representing other local structures to interpret the structure-property relationship in S effectively, as follows:

$$\rho(\mathbf{A}) = \boldsymbol{\alpha}^g = \mathrm{softmax}([\hat{\mathrm{s}}_1, \hat{\mathrm{s}}_2, ..., \hat{\mathrm{s}}_M]) \tag{4.9}$$

The representation vector $\mathbf{x}_S$ of the material structure S is then formulated by aggregating the representations of M local structures according to the obtained global

attention (GA) scores, as follows:

$$\begin{aligned} \mathbf{x}_S &= \text{GlobalAttention}(\mathbf{C}^L) \\ &= \mathit{SAttention}(\mathbf{Q}^g, \mathbf{K}^g) \\ &= \rho(\mathbf{Q}^{g\top}\mathbf{K}^g)\mathbf{K}^g = \rho(\mathbf{A})\mathbf{K}^g \\ &= \boldsymbol{\alpha}^g\mathbf{K}^g = \sum_{i=1}^{M} \alpha_i^g \mathbf{k}_i^g, \end{aligned} \tag{4.10}$$

The GA score $\boldsymbol{\alpha}^g = [\alpha_1^g, \alpha_2^g, ..., \alpha_M^g]$, which describe the degrees of attention paid to each local structure in S, are used to reveal critical aspects to help interpret the structure-property relationship of S.

Consequently, the physical property y_S of the material structure S can be predicted from the learned representation $\mathbf{x}_S$ with fully connected layers F_S, as follows:

$$\widehat{y}_S = F_S(\mathbf{x}_S) \tag{4.11}$$

The design of the SCANN architecture, especially the inclusion of a fully connected layer, is specifically crafted to capture the intricate and nonlinear relationships between representations and their corresponding properties. Additionally, the GA scores $\boldsymbol{\alpha}^g$ of the local structures, derived from the global attention layer, aid in identifying key factors that enhance the understanding of the material's structure–property relationships.

4.3.4 Refining Model Design: From SCANN to SCANN+ through Iterative Development

The design of SCANN was iteratively refined to address limitations and incorporate improvements inspired by observed trends during preliminary evaluations and theoretical analysis. The updated model, SCANN$^+$, integrates enhanced mechanisms to better capture complex local structures and their interactions within materials. Key updates include modifications to the geometry influence for improved representation learning and adjustments to the architecture to optimize performance in predictive tasks. These enhancements were guided by a continuous feedback loop, ensuring alignment with the model's objectives and the dynamic nature of material data representation. The SCANN$^+$ introduced the embedding vector for the Voronoi solid angle θ_{ij} as follows:

$$\mathbf{g}_{ij}^0 = \mathbf{DE}(d_{ij}) \times \mathbf{AE}(\theta_{ij}), \tag{4.12}$$

where $\mathbf{DE}(d_{ij})$ and $\mathbf{AE}(\theta_{ij})$ are the distance and angle embedding layers corresponding to the distance d_{ij} and angle θ_{ij} of an h-dimensional vector. Similar to the d_{ij} , the Voronoi solid angle θ_{ij} is also expanded by k Gaussian basis functions $\phi_i(x) = \exp(-(x-\mu_k^a)/2\sigma^2)$ at each center 0 rad $< \mu_k^a < 4\pi$ rad for every $\sigma = 0.5$ rad and applied to the angle embedding layer. Next, the embedding layers is defined as shown below:

$$\mathbf{AE}(d_{ij}) = F_a([\phi_1(\theta_{ij}), \phi_2(\theta_{ij}), \ldots, \phi_k(\theta_{ij})]) \tag{4.13}$$

where $F_a : \mathbb{R}^k \to \mathbb{R}^h$ is the fully-connected layer with a Swish activation function.

In addition, the geometry influences between the neighbor $\mathbf{c}_j^l$ and the center $\mathbf{c}_i^l$ are updated based on the following formulation:

$$\mathbf{g}_{ij}^{l+1} = F_g^l([\mathbf{c}_i^l \oplus \mathbf{g}_{ij}^l \oplus \mathbf{c}_j^l]) + \mathbf{g}_{ij}^l, \tag{4.14}$$

where $\oplus$ denotes the concatenating vectors and $F_g : \mathbb{R}^h \rightarrow \mathbb{R}^h$ is a fully connected (FC) layer.

These updates collectively improve the iterative information propagation through layers, enabling a more effective and nuanced representation of material structures. By refining the modeling of geometric influences, SCANN$^+$ enhances its ability to capture the complex spatial relationships and local structural interactions that are critical for accurately predicting material properties. This advancement ensures a more comprehensive integration of distance and angular dependencies, aligning the model with the dynamic and multidimensional nature of material data.

4.3.5 Loss design and Model training

The training of the DL model using the proposed architecture begins with initializing all learnable parameters. Weighting matrices such as $\mathbf{W}_\mathbf{q}^l$, $\mathbf{W}_\mathbf{k}^l$, $\mathbf{W}_\mathbf{q}^g$, and $\mathbf{W}_\mathbf{k}^g$ are initialized as random matrices using the Glorot Uniform method Glorot and Bengio, 2010, while all bias vector entries are set to zero. To enhance regularization, dropout layers, and attention dropout Vaswani et al., 2017 are applied within the local attention layers at a rate of 0.1.

During the training process, all parameters of the proposed DL model are updated by minimizing a loss function using Adam optimization Kingma and Ba, 2017, with a scheduled learning rate decay ranging from 5×10^{-4} to 10^{-4}. To predict the physical property y_S of a material structure S in the training dataset D, the loss function is defined as follows:

$$\mathcal{L} = \frac{1}{|\mathsf{D}|} \sum_{S \in \mathsf{D}} (y_S - \hat{y}_S)^2 \tag{4.15}$$

Remarkably, SCANN consists of fewer than one million parameters, primarily influenced by the configuration settings of the number of LocalAttention layers (L). Appendix Table II presents the epoch-wise time cost for the QM9 dataset with a batch size 128. SCANN excels in training times per epoch and is notable for its commendable memory efficiency, making it highly suitable for practical deployment.

4.3.6 Experimental design

In this study, we develop two variants of deep learning models based on the proposed SCANN architecture and its enhanced version, SCANN$^+$. Each variant is trained independently on different datasets with distinct target properties to assess the architecture's effectiveness in predicting these properties and its capability to elucidate structure-property relationships (interpretability) across five molecular and crystal structure datasets (see Table 4.1). The properties within these datasets are derived from quantum mechanical calculations performed using density functional theory (DFT). The data is divided into training, validation, and test sets to evaluate predictive performance. The models are trained on the training set and optimized to minimize the mean absolute error (MAE) on the validation set. The MAEs for the predicted target properties on the test sets are then reported and compared with

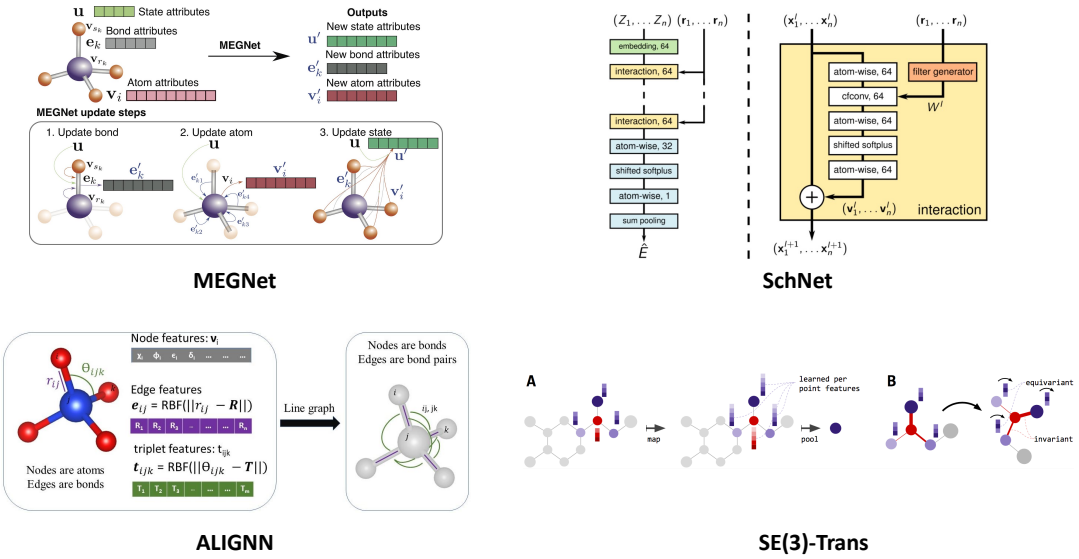


FIGURE 4.5: Illustration of four neural networks designed for materials with their key innovations. MEGNet (Chen et al., 2019): inclusion of state attributes. SchNet (Schütt et al., 2018): convolution filter for atomic interaction. ALIGNN (Choudhary and DeCost, 2021): updating bond angle representations by line graph. SE(3)-Trans (Fuchs et al., 2020): equivalence network for rotations and translations. Reprinted and adapted with permission from Refs. Chen et al., 2019; Schütt et al., 2018; Choudhary and DeCost, 2021; Fuchs et al., 2020.

other models documented in the literature. From this point forward, the models implemented using the SCANN architecture and trained on their respective datasets will be referred to as SCANN models.

We assessed the predictive capabilities of the SCANN models by comparing them with seven DL models using the QM9 dataset (Ramakrishnan et al., 2014). The models compared include SchNet (Schütt et al., 2018), and MEGNet (Chen et al., 2019), all of which employ graph neural networks to represent molecules or crystals as atomistic graphs. Additionally, Cormorant (Anderson, Hy, and Kondor, 2019) and SE(3)-Trans (Fuchs et al., 2020) are variants of graph neural networks that integrate physical constraints, such as covariant or equivalence principles, on the three-dimensional coordinates of atoms. On the other hand, ALIGNN (Choudhary and DeCost, 2021), the leading network in this area, utilizes an extra line graph where bonds act as nodes, and edges represent the angular relationships between bonds in addition to the atomistic graph. These added information allows ALIGNN to effectively capture the geometric arrangement of triplets of atoms in a molecule or crystal.

Furthermore, the interpretability of the SCANN models is assessed by examining the relationship between the learned GA scores of the local structures and the corresponding results from first-principles calculations. The results demonstrate the ability of the SCANN models to provide valuable information regarding the structure–property relationships of materials in four scenarios: the local structures and HOMO/LUMO molecular orbitals (QM9 Ramakrishnan et al., 2014 and Fullerene-MD Vu and Chi, 2023), the deformation energy ΔU and the deformation of the Pt/graphene structures (Pt/graphene-MD Vu and Chi, 2023).

TABLE 4.1: Summary of datasets used in evaluation experiments. The table shows information of five datasets regarding eight properties analyzed with the SCANN models, including dataset size (number of molecules/crystals - #Size), number of atoms present in structures (#Atoms), and the specific physical properties examined.

Dataset	#Size	#Atoms	Properties
QM9 (Ramakrishnan et al., 2014)	130,831	4 to 29	E_{HOMO} , E_{LUMO}, E_{gap}, α, C_v
Fullerence-MD (Vu and Chi, 2023)	3000	60, 70, 72	E_{HOMO} , E_{LUMO}
Pt/Graphene-MD (Vu and Chi, 2023)	21,666	103	ΔU

E_{HOMO} (meV): Energy of the highest occupied molecular orbital; E_{LUMO} (meV): Energy of the lowest unoccupied molecular orbital; E_{gap} (meV): Energy HOMO-LUMO gap; α (Bohr^3): Isotropic polarizability; C_v (cal mol^{-1} K^{-1}): Heat capacity at 298 K; ΔU (eV): Deformation energy; ΔE (meV atom^{-1}): Formation energy per atom; E_g (meV): Band gap.

4.4 Case Study 1: Material property prediction for small molecules

The QM9 dataset, as described in Ramakrishnan et al., 2014, is an extensive repository containing data on 133,885 drug-like organic molecules, predominantly composed of five elements: carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and fluorine (F). During the refinement process for our analysis, 3,054 entries were excluded due to concerns about their geometric stability, as noted in Anderson, Hy, and Kondor, 2019. Consequently, the final dataset consists of 130,831 well-defined molecules, which are the foundation for our subsequent experiments.

To evaluate the predictive performance of the SCANN models, we focus on five key physical properties derived from the QM9 dataset. These properties are essential for understanding molecular behavior and include:

1. The energy of the highest occupied molecular orbital (E_{HOMO}), which provides insights into a molecule's electron-donating capabilities.

2. The energy of the lowest unoccupied molecular orbital (E_{LUMO}) helps determine a molecule's electron-accepting potential.

3. The energy gap (E_g), calculated as the difference between E_{LUMO} and E_{HOMO}, indicating the molecule's stability and reactivity.

4. The isotropic polarizability (α) measures how easily an external electric field can distort the electron cloud of a molecule.

5. The heat capacity at 298 K (C_v) reflects how a molecule absorbs and stores thermal energy at room temperature.

In the context of dynamic phenomena, the QM9 dataset provides valuable insights into the foundational quantum interactions that govern molecular behavior. Although the dataset primarily consists of static molecular structures, its quantum mechanical properties directly influence the dynamic processes in materials, such as:

- Electronic Dynamics: Properties like the HOMO-LUMO gap and polarizability provide a basis for understanding electronic transitions and responses, which are crucial for dynamic phenomena like charge transport and exciton diffusion in materials.
- Structural Transformations: The dataset's inclusion of molecular geometries allows exploration of how atomic arrangements evolve under external stimuli, laying the groundwork for understanding structural dynamics at larger scales.
- Transfer Learning: Pre-trained models on QM9 capture fundamental chemical relationships and patterns that can be transferred to other datasets or related tasks. This is especially beneficial when working with limited data on dynamic processes, as the model has already learned basic chemical principles.

In our experiments, we compare the predictive capabilities of the SCANN models against those of several recent state-of-the-art DL models, aiming to advance our understanding of molecular property predictions.

4.4.1 Evaluation of the predictive performance

To evaluate the predictive capability of SCANN in forecasting five physical material properties within the QM9 dataset, we perform train–validation–test splits in an 80:10:10 ratio. Additionally, six DL methods are employed for comparison, with their prediction accuracies assessed using mean absolute error (MAE). The evaluation process is repeated five times to calculate an average MAE for the test set, thereby providing a robust assessment of the model's predictive performance Fuchs et al., 2020; Anderson, Hy, and Kondor, 2019.

Table 4.2 presents the average Mean Absolute Error (MAE) scores obtained from five training runs of the SCANN models and scores from competing models on the QM9 dataset. The ALIGNN model outperforms all other competing models across the four properties evaluated. In comparison, the MAEs of the SCANN models are 2 to 2.5 times higher than those of ALIGNN. Despite this disparity, SCANN demonstrates competitive performance relative to other remaining models, particularly in predicting E_{HOMO}, E_{LUMO}, and E_g.

Incorporating traditional prior knowledge—such as numerous atomic features and bonding information between atoms Choudhary and DeCost, 2021; Chen et al., 2019; Xie and Grossman, 2018—or integrating physical constraints like equivalencies, covariates, and equations Anderson, Hy, and Kondor, 2019; Fuchs et al., 2020; Hirn, Mallat, and Poilvert, 2017 into the learning process for structural representations can enhance prediction accuracy. For instance, the ALIGNN model outperformed all competing models by introducing additional angular information among triplets of atoms. In contrast, other models only considered two-body interactions (such as distances and bond valences). To better capture the geometrical structures of molecules or crystals, we developed an enhanced version of SCANN, termed SCANN$^+$, which includes minor modifications to the original architecture by incorporating a Voronoi solid angle embedding layer and refining the geometrical information through multiple LocalAttention layers. These enhancements significantly improve the method's predictive capabilities; the SCANN$^+$ models outperform all other competitors, except for the ALIGNN model, in predicting electronic properties (E_{HOMO}, E_{LUMO}, and E_g in the QM9 dataset; that are sensitive to the geometrical structure of molecules or crystals (see Tables 4.2). The performance of SCANN$^+$ on the training data for each property is shown in the Fig. **qm9_reg**

TABLE 4.2: Comparative evaluation of SCANN, SCANN$^+$, and six other DL models predicting five physical properties using the QM9 dataset.

	E_{HOMO} (meV)	E_{LUMO} (meV)	E_{gap} (meV)	α (Bohr^3)	C_v (cal mol^{-1} K^{-1})
WaveScatt	85	76	118	0.160	0.049
SchNet	41	34	63	0.235	0.033
MEGNet	38	31	61	0.081	0.030
Cormorant	34	38	61	0.085	**0.026**
SE(3)-Trans	35	33	53	0.142	0.054
ALIGNN	**21**	**19**	**38**	**0.056**	–
SCANN	41	37	61	0.141	0.05
SCANN$^+$	32	31	52	0.115	0.041

E_{HOMO}: Energy of the highest occupied molecular orbital; E_{LUMO}: Energy of the lowest unoccupied molecular orbital; E_{gap}: Energy gap; α: Isotropic polarizability; C_v: Heat capacity at 298 K. The dash symbol (–) indicates the result has not been reported yet. The bold numbers denote the lowest mean absolute errors (MAEs) among the eight models.

Implementing these strategies increases dimensionality, which can introduce biases into the model. This bias may favor certain materials while overlooking others, or it may lead to an oversimplification of complex phenomena that arise from the constraints or inaccuracies present in the heuristic information utilized during the training phase. Consequently, these challenges can hinder a detailed understanding of the critical structure–property relationships that are central to this study's objectives. In the context of the QM9 dataset, the field recognizes specific benchmarks for "chemical accuracy". These thresholds are notably set at 43 meV for the three energy-related properties: t E_{HOMO}, E_{LUMO}, and E_g; 0.1 Bohr^3 for the isotropic polarizability α; and 0.05 cal mol^{-1} K^{-1} for the heat capacity C_v at 298 K (Faber et al., 2017). Notably, the SCANN models demonstrated impressive performance by achieving a prediction error of 41 meV for E_{HOMO}, 34 meV for E_{LUMO} and 0.05 cal mol^{-1} K^{-1} for C_v. These results indicate that the models have successfully fulfilled the criteria for chemical accuracy regarding these specific properties.

In practical applications, it is often unnecessary to surpass the threshold for chemical accuracy by simply increasing the complexity of the models used. This is particularly true when the data originates from DFT calculations. Adopting a more complicated model can risk overfitting the data or introducing biases, which in turn can compromise the model's interpretability and obscure the underlying chemical principles. Therefore, we focus on examining the relationship between the molecules structures within the QM9 dataset and their associated properties (E_{HOMO} and E_{LUMO}) by using the GA scores obtained from the SCANN models instead of those from the more complex SCANN$^+$ model, which possesses higher dimensionality and more parameters.

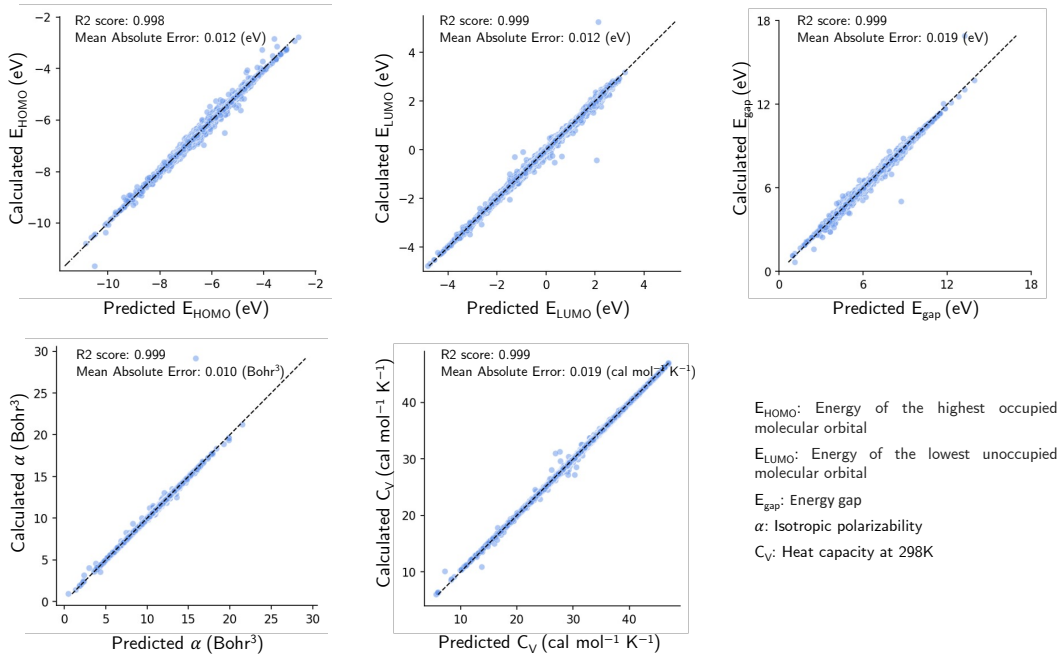


FIGURE 4.6: Illustration of regression performance of SCANN$^+$ on five properties in QM9 dataset.

4.4.2 Correspondence between the learned attentions of local structures and the molecular orbitals of small molecules:

For the small molecules in the QM9 dataset, the SCANN models demonstrate a remarkable correspondence between the obtained GA scores of the local structures and molecular orbitals results obtained via DFT calculations. As a representative example, Figure 4.7 shows the comparison between the GA scores of the local structures and the HOMO/LUMO orbitals obtained from DFT calculations for four molecules. Notably, an apparent correspondence between the relative GA scores of the local structures and the HOMO orbitals of the dimethyl butadiene molecule (*cis*-2,3-dimethyl-1,3-butadiene) is evident (Fig. 3a). Furthermore, the GA scores of the local structures can be easily linked to the interpretation that dimethyl butadiene readily undergoes the Diels–Alder reaction. Similarly, the correspondence between the HOMO orbital and the GA scores of the local structures is apparent for the thymine molecule (5-methyl pyrimidine-2,4 (1*H*,3*H*)-dione), which is a nucleobase in DNA (Fig. 4.7b).

Moreover, similar correspondences are confirmed between the GA scores of the local structures and the LUMO orbitals obtained from the DFT calculations for methyl acrylate (methyl prop-2-enoate) and dimethyl fumarate (dimethyl(2E)-but-2-enedioate). Methyl acrylate is a reagent that is commonly used in the synthesis of various pharmaceutical intermediates (Ohara et al., 2020), whereas dimethyl fumarate has been proposed to exhibit immunomodulatory properties without causing significant immunosuppression (Schulze-Topphoff et al., 2016); thus, it has been evaluated as a potential treatment for COVID-19 (Mantero et al., 2021). The apparent correspondence between the LUMO orbitals and the GA scores of the local structures of these two molecules (Fig. 4.7c and d) further highlight that the attention scores of the SCANN model provide valuable insights to interpret the structure–property relationships of molecules. Further investigations show that the obtained GA scores from the SCANN$^+$ models are almost consistent with those of the SCANN models for these molecules.

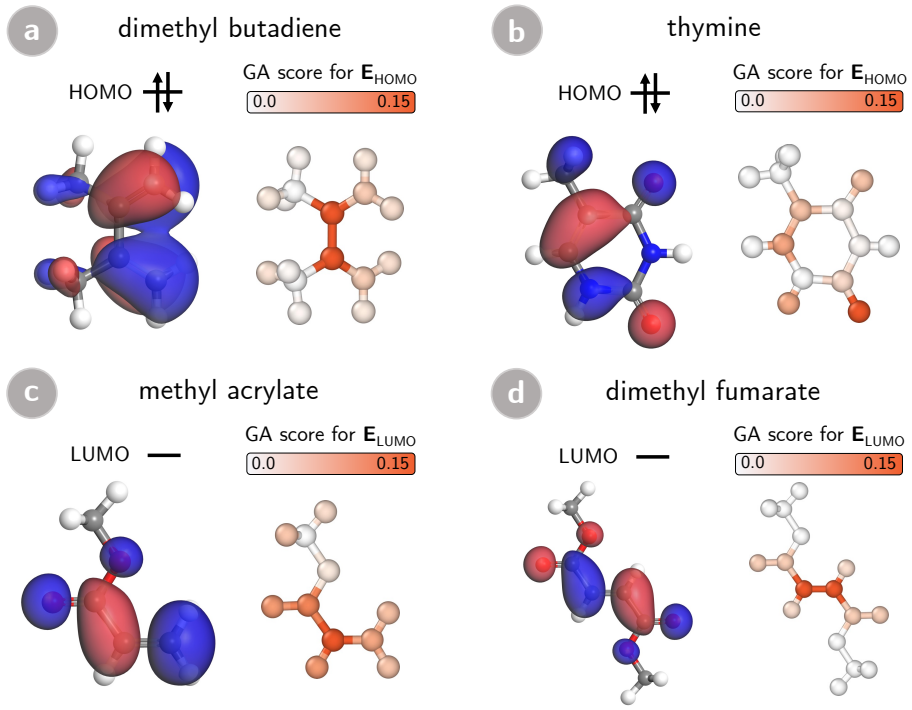


FIGURE 4.7: Visualization of structure–property relationships in the QM9 dataset, showing the correspondence between GA scores and molecular orbitals for four molecules: (a) dimethyl butadiene, (b) thymine, (c) methyl acrylate, and (d) dimethyl fumarate. For each molecule, the left side of the figure illustrates the wave function of the HOMO (a), (b), or the LUMO (c), (d), as calculated via DFT. The isosurfaces with positive and negative values of the wave functions are represented by blue and red lobes, respectively. The right-side figures display the GA scores of the local structures derived from the SCANN models, where atom colors indicate estimated GA scores; link colors do not signify the sign or nodes of the molecular orbital wave functions.

All carbon, nitrogen, and oxygen atomic sites in the QM9 dataset were statistically analyzed for a systematic evaluation of the GA scores obtained by the SCANN models. Since the GA scores of atomic sites were normalized to 1, the relative GA scores were calculated based on the average GA score of the sp^3-hybridized carbon atoms in each molecule. Molecules without any sp3-hybridized carbon atoms were excluded (Fig. 4.8). The analysis of the GA scores for the HOMO energy reveals that the influence on HOMO follows the order of oxygen, nitrogen, and carbon. Specifically, sp^3-hybridized carbon sites have a lower influence compared to sp^2-hybridized or sp-hybridized carbon sites (Fig. 4a). These findings align with the electronegativity and bonding characteristics of the elements. Oxygen and nitrogen exhibit strong electronegativity and electron-rich regions in π-bonds, leading to a more significant electron density shift and higher HOMO energy localized around oxygen, nitrogen, and carbon sites with double or triple bonds.

In contrast, the GA scores for the LUMO energy show no significant difference among the three elements. This observation is consistent with the understanding that unoccupied orbitals primarily influence the LUMO energy, resulting in a less pronounced difference in electronegativity compared to its impact on the HOMO energy (Fig. 4.8b).

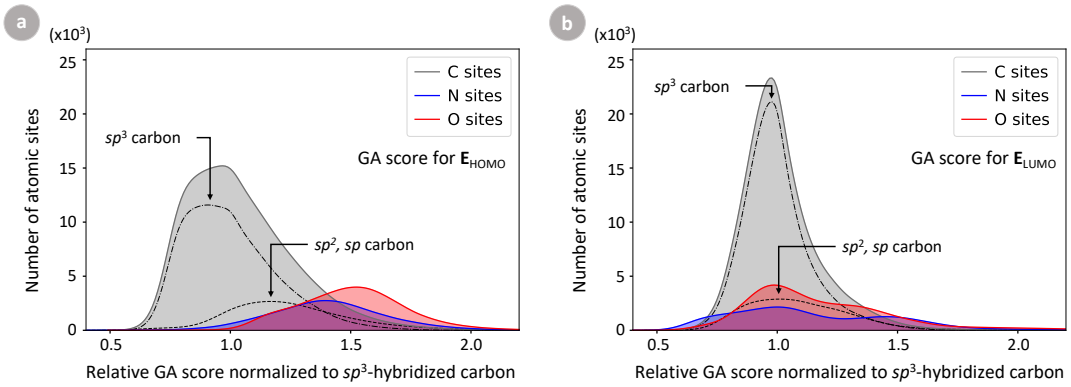


FIGURE 4.8: Correspondence between obtained GA scores of carbon, nitrogen, and oxygen atomic sites and molecular orbitals of molecular structures in QM9 dataset. Statistics of the relative GA scores for E_{HOMO} (a) and E_{LUMO} (b) for all carbon, nitrogen, and oxygen atomic sites in the molecular structures of the QM9 dataset, calculated based on the average GA score of sp^3-hybridized carbon atoms in each molecule. Gray, blue, and red lines and filled regions represent the statistics for carbon, nitrogen, and oxygen sites, respectively.

4.5 Case Study 2: Material property prediction for molecular dynamics of fullerene molecules

Fullerene-MD (Vu and Chi, 2023) is an in-house developed computational material dataset comprising data on three well-known fullerene molecules: C_{60} (I_h), C_{70} (D_{5h}), and C_{72} (D_{6h}). Fullerenes are a class of carbon allotropes characterized by their closed-cage structures, which exhibit unique electronic, optical, and mechanical properties. These molecules have garnered significant attention due to their potential applications in various fields, including materials science, electronics, nanotechnology, and medicine. For instance, fullerenes are utilized in developing organic photovoltaics, as drug delivery agents, and in creating advanced composite materials.

The dataset includes optimized structures and 3,000 deformed structures obtained from molecular dynamics simulations, with 1,000 structures for each molecule. The HOMO (E_{HOMO}) and LUMO (E_{LUMO}) energies of these structures are determined using density functional theory (DFT) calculations, following the methodology employed in the QM9 dataset. These energy levels are crucial for understanding the electronic properties and reactivity of fullerenes, which influence their suitability for applications such as electronic devices and photovoltaic cells.

Experiments conducted on this dataset aim to evaluate the predictive capability of the SCANN models for HOMO and LUMO energies and to assess the interpretability of the model's predictions within dynamic scenarios. A distinctive feature of all structures in this dataset is that they contain only carbon atoms, simplifying the analysis while highlighting the intrinsic properties of carbon-based nanomaterials. Furthermore, due to the symmetric nature of fullerene molecules, the local structures within each molecule are highly similar, with only minor differences arising from deformations. This uniformity allows for a precise evaluation of the interpretability of the SCANN model, as it can effectively discern subtle variations in structure-property relationships.

In the evaluation experiment using the Fullerene-MD dataset, SCANN models

pre-trained on the QM9 dataset are applied to train prediction models for the HOMO and LUMO energies of the fullerene molecules. This approach leverages extensive training on a diverse set of molecules to enhance the accuracy and generalizability of predictions for highly symmetric and structurally similar fullerene compounds. By doing so, the study not only validates the performance of the SCANN architecture in predicting electronic properties but also underscores its capability to provide meaningful insights into the structure-property relationships of carbon-based nanomaterials.

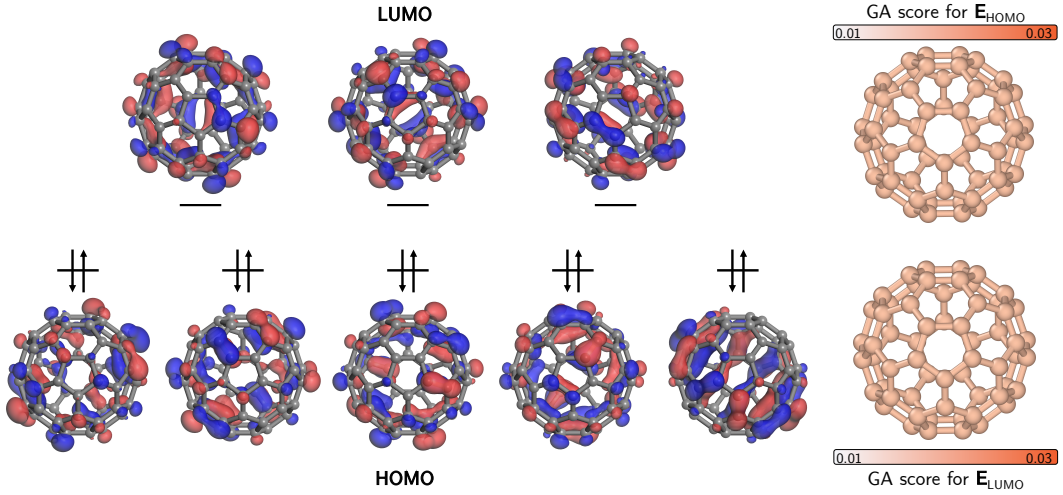


FIGURE 4.9: Visualizations of structure–property relationships in fullerene molecules. Correspondence between the obtained GA scores and the molecular orbitals of C_{60}. The left panel illustrates the wave functions of the degenerate HOMO (bottom) and LUMO (top) orbitals calculated via DFT, where blue and red lobes represent positive and negative isosurfaces, respectively. The right panel displays the GA scores of local structures derived from the SCANN model for the corresponding property.

4.5.1 Evaluation of the predictive performance

Herein, a similar number of train–validation–test splits are applied as those used in the QM9 dataset experiments. In addition, to predict E_{HOMO} and E_{LUMO} for the fullerene structures in Fullerene-MD, the weights from the QM9 dataset are applied as the pre-train model to initialize weights of the SCANN model in the learning process on the Fullerene-MD dataset. A detailed explanation regarding the optimal SCANN hyperparameters for these datasets is presented in section V.

In the test for predicting the HOMO and LUMO energies with the Fullerene-MD dataset using pre-trained weights, the SCANN models yield MAEs of 23 meV and 27 meV, respectively, which are less than two-third of the "chemical accuracy" threshold. The remarkable prediction accuracy of SCANN confirms its practical applicability and suggests that the interpretation derived from the attention scores provides valuable insights into key structure–property relationships for the investigated material properties. In the following sections, we examine the correspondence between the obtained GA scores of the local structures and the corresponding results from first-principles calculations to assess the interpretability of the SCANN models.

4.5.2 Correspondence between the learned attentions of local structures and molecular orbitals of fullerene molecules

To further evaluate the interpretability of the proposed method, the correspondence between the obtained GA scores of the local structures and the molecular orbitals obtained from DFT calculations for fullerene molecules is examined. Supplementary Figure 4.9 shows the GA scores of the local structures for the HOMO and LUMO energies of the C_{60} molecule (I_h symmetry). In this case, the target molecule has a truncated icosahedral structure composed of 20 hexagons and 12 pentagons, with all carbon atoms exhibiting equivalent local structures. The SCANN model estimates identical GA scores for all local structures of the C_{60} molecule, thus indicating its ability to handle large and symmetric molecules.

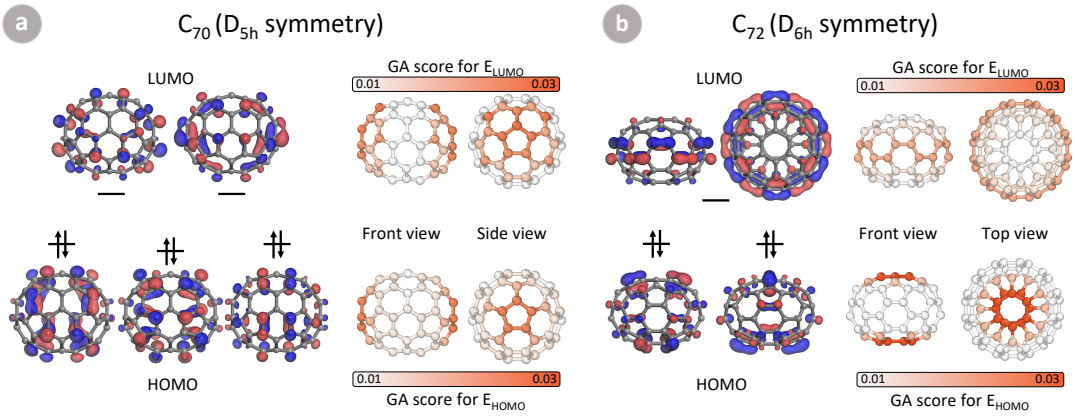


FIGURE 4.10: Visualizations of structure–property relationships for fullerene molecules. Correspondence between obtained GA scores and the molecular orbitals of (a) C_{70} and (b) C_{72}. For each molecule, the left side shows wave functions of the degenerate HOMO (bottom) and LUMO (top) orbitals calculated via DFT, with blue and red lobes representing positive and negative isosurfaces, respectively. he blue and red lobes in the illustration represent positive and negative isosurfaces, respectively. The right panel shows the GA scores for local structures derived from the SCANN model, corresponding to the specific property being analyzed.

As the number of carbon atoms in the fullerene molecule increases, the symmetry of the C_{70} (D_{5h} symmetry) and C_{72} (D_{6h} symmetry) molecules becomes slightly broken, and the local structures of the carbon atoms in these molecules are no longer equivalent. Figure 4.10 demonstrates the significant correspondence between the GA scores of the local structures and the HOMO and LUMO results obtained from DFT calculations for the C_{70} and C_{72} molecules. The GA scores of the local structures in the C_{70} and C_{72} molecules exhibit a five-fold (top view) and six-fold (top view) symmetry upon the prediction of the HOMO energy, respectively. These results align with the structural symmetry and degenerate HOMO orbitals of the two fullerene molecules. Notably, the C_{70} molecule possesses an additional 10-carbon ring, forming a plane symmetry, resulting in a planar symmetry of its HOMO with the node situated on that ring's plane. The SCANN model reveals a clear correspondence between the HOMO of the C_{70} molecule and the GA scores of the local structures (Fig. 4.10a), along with the LUMO and their corresponding GA scores. Furthermore, the shapes of LUMO and HOMO of the C_{72} molecule exhibit a perfect correspondence with the GA scores of the local structures obtained using the

SCANN models (Fig. 4.10b). Compared to C_{60}, the C_{72} molecule has an additional ring of 24 carbon atoms with six-fold symmetry, consisting of 12 pairs of carbon–carbon bonds in five-membered carbon rings. The high GA scores of the local structures in the ring indicate the localization of the LUMO of the C_{72} molecule on the ring. In contrast, the HOMO orbitals are located on two opposite sides of the ring and are also captured by the local structures with high GA scores. This evaluation experiment provides further confirmation that SCANN-derived GA scores offer valuable insights for understanding the structure–property relationship, even for large molecules.

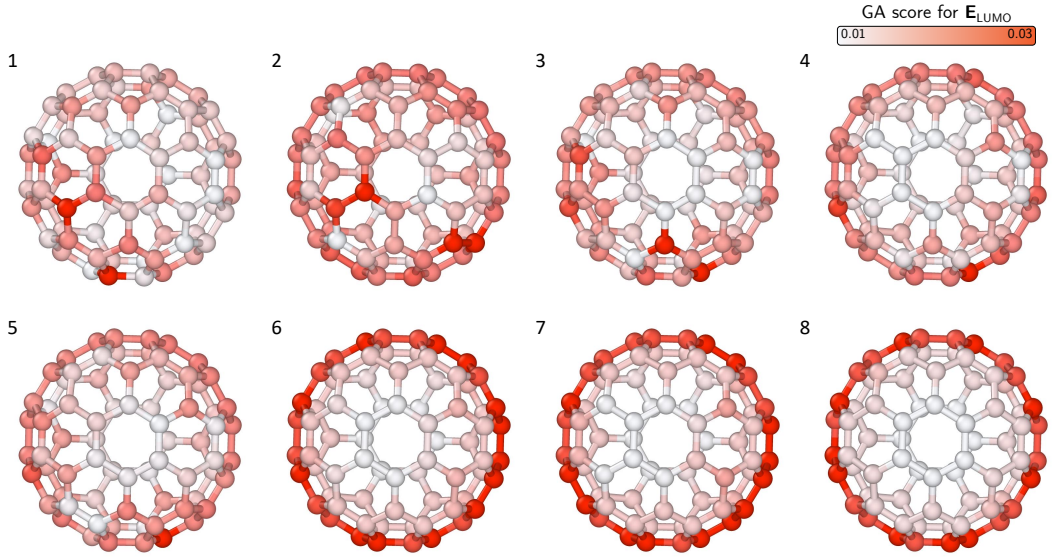


FIGURE 4.11: Visualizations of structure–property relationships for fullerene molecules at 8 consecutive molecular dynamics steps of LUMO property. The number in the top left corner indicate the index of the molecular dynamics steps.

Figures 4.11 and 4.12 further illustrate this by displaying eight consecutive molecular dynamics structures of the C_{72} molecule as it converges towards its stable configuration. Throughout the MD simulation, the molecule undergoes structural changes that influence its electronic properties. The GA scores for the local structures continuously adjust in response to these changes, particularly for the LUMO and HOMO. As the structure evolves, the GA scores progressively converge to those of the stable configuration, reflecting the gradual localization of the LUMO on the additional carbon ring. This dynamic adjustment is reasonable and demonstrates how the SCANN model captures the interplay between structural dynamics and electronic properties. More details about the visualization of attention for the molecular dynamics are shown in the videos at Appendix A.

The ability of the GA scores to adapt and accurately represent changes in electronic localization during molecular dynamics simulations underscores the effectiveness of the SCANN model in analyzing dynamic phenomena. It highlights the model's potential in providing deeper understanding of how local structural variations impact the electronic properties of molecules over time. This enhanced descriptive capability is crucial for studying large molecular systems where structural and electronic complexities are significant.

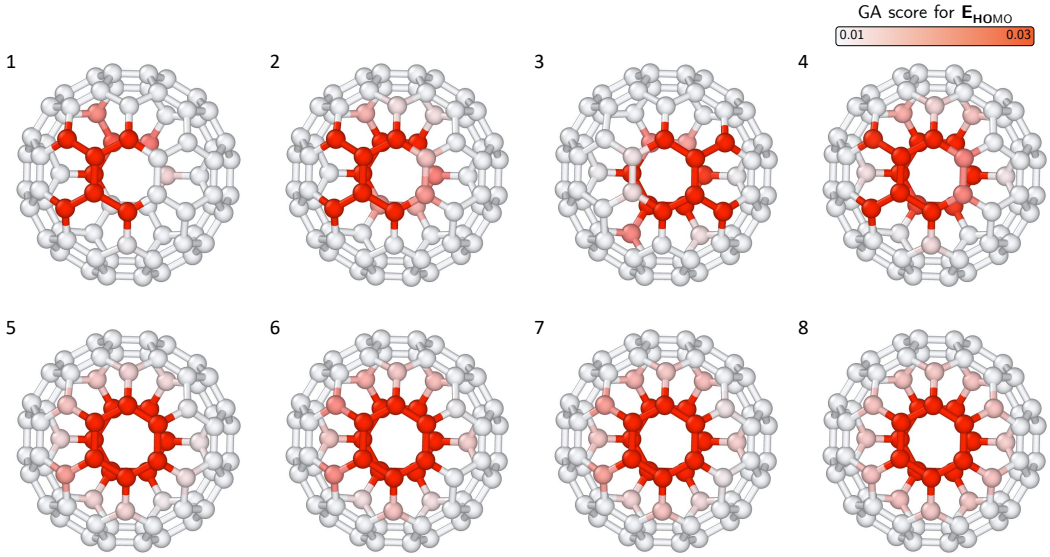


FIGURE 4.12: Visualizations of structure–property relationships for fullerene molecules for 8 consecutive molecular dynamics steps of HOMO property. The number in the top left corner indicate the index of the molecular dynamics steps.

4.6 Case Study 3: Material property prediction for structural deformation in Pt/graphene

Pt/graphene-MD Vu and Chi, 2023 is an in-house developed computational material dataset that represents a system composed of a platinum (Pt) atom adsorbed on a graphene flake terminated by hydrogen atoms Chi et al., 2006; Dam et al., 2009. This dataset includes approximately 21,000 optimized and deformed structures generated through molecular dynamics simulations, providing a comprehensive overview of the Pt-graphene interactions under various conditions. The adsorption energies of these structures are determined using density functional theory (DFT) calculations, following the methodology employed in the QM9 dataset.

The primary objectives of the experiments conducted on this dataset are twofold: first, to evaluate the predictive performance of the SCANN models in forecasting the deformation energies (ΔU) of the structures, and second, to assess the interpretability of the model's predictions concerning these deformation energies. A distinctive feature of this dataset is the presence of a two-dimensional honeycomb network of carbon atoms forming the graphene flake. Although the local structures of each carbon atom in the system exhibit slight distortions from the ideal sp^2 hybridization structure Dam et al., 2009, this dataset facilitates a quantitative evaluation of the interpretability of the SCANN models in terms of the distortion of the honeycomb network on the graphene surface.

Platinum-graphene systems are highly interested in materials science and nanotechnology due to their exceptional catalytic properties. Pt atoms supported on graphene surfaces serve as highly efficient catalysts for various chemical reactions, including hydrogenation, oxygen reduction, and carbon dioxide reduction Dam et al., 2009. The unique interaction between Pt atoms and the graphene substrate enhances catalytic activity and stability, making these materials pivotal in fuel cells, chemical synthesis, and environmental remediation applications.

In terms of properties, the Pt/graphene-MD dataset captures the intricate balance between adsorption energy and structural deformation, providing insights into the stability and reactivity of Pt-graphene interfaces. The ΔU are critical for understanding how Pt atoms influence the mechanical and electronic properties of graphene, which in turn affect the material's overall performance in practical applications. By leveraging the SCANN models trained on this dataset, researchers can better understand the structure-property relationships, enabling the design of more efficient and durable Pt/graphene-based materials.

4.6.1 Evaluation of the predictive performance

Herein, a similar number of train–validation–test splits are applied as those used in the QM9 dataset experiments. For the Pt/Graphene-MD dataset, the SCANN model achieves an MAE of 0.16 eV in predicting the ΔU of the system. Figure 4.13 shows the ΔU and the SCANN model's predictions over the first 500 steps of the molecular dynamics simulation. The SCANN model closely tracks the actual deformation energy throughout these steps, demonstrating its ability to accurately predict energy fluctuations associated with the dynamic structural changes in the Pt/Graphene system. This alignment indicates that the model effectively captures the complex interactions and deformations occurring during the simulation.

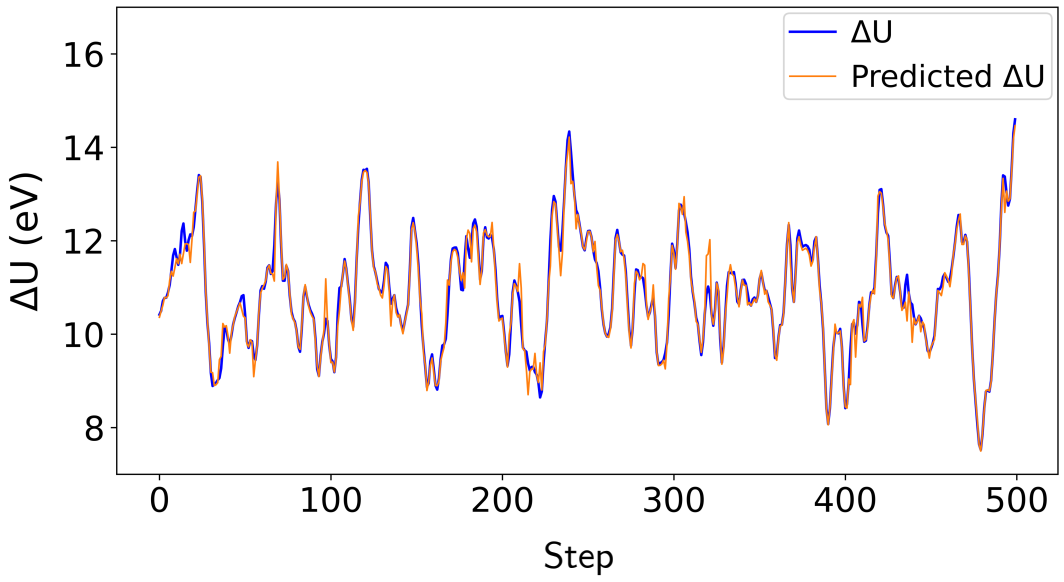


FIGURE 4.13: Visualization of the ΔU over the first 500 steps of the molecular dynamics simulation, alongside the predictions made by the SCANN model.

The ability of the SCANN model to accurately predict deformation energy underscores its effectiveness in delivering insightful interpretations through GA scores. These scores highlight the contributions of specific local structures to the overall deformation energy, providing a deeper understanding of the structural dynamics within the Pt/Graphene system during molecular dynamics simulations.

4.6.2 Correspondence between the learned attentions of local structures and structural deformation in Pt/graphene:

Figure 4.14 a presents the GA scores of the local structures obtained by the SCANN model for predicting the deformation energy of a system comprising a platinum atom adsorbed on a graphene flake. The deformation energy is defined as the difference between the total energy of the deformed and optimized structures. A detailed examination of the obtained GA scores reveals that local structures with high GA scores possess relatively elongated carbon–carbon bonds (Fig. 4.14 b). Additionally, the carbon atoms that form high local curvatures upon the formation of a convex from the planar structure of the sp^2 hybridization bonding network received high GA scores (Fig. 4.14c).

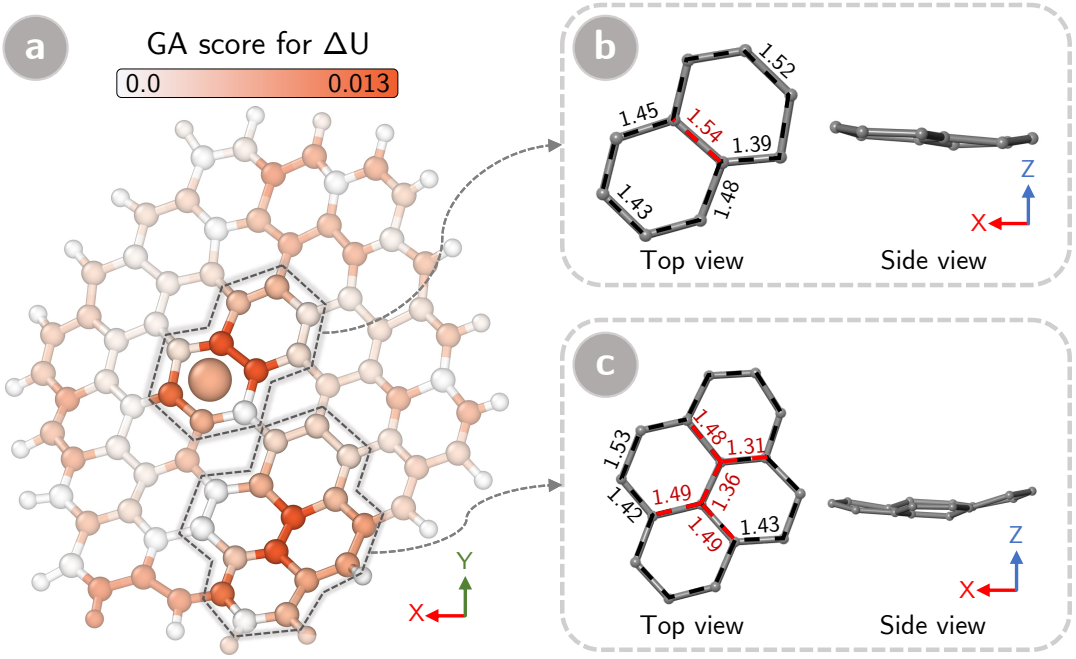


FIGURE 4.14: Visualization of the relationship between adsorption energy and deformation of a graphene flake with an adsorbed platinum atom. (a) GA scores from the SCANN model for the deformed Pt/Graphene system; atom colors indicate estimated GA scores, while link colors do not represent molecular orbital wave functions. Structural visualizations of the high attention local structures during the deformation: (b) elongated carbon–carbon bond, and (c) convexed carbon–carbon configuration. Distances between adjacent carbon atoms (in Å) highlight the distortion caused by the deformation.

The results obtained from the experiment on the system where a platinum atom was adsorbed on a graphene flake reveal that the GA scores obtained by the SCANN model exhibit a high correspondence with the observed structural deformations. In particular, the high GA scores for the increased carbon–carbon bond lengths and the convexed carbon atoms align well with the contribution to the deformation energy, as determined by DFT calculations. This finding indicates that the GA scores generated by SCANN are reliable indicators of structural deformations in such systems, demonstrating the model's capability to capture and interpret the underlying material instability.

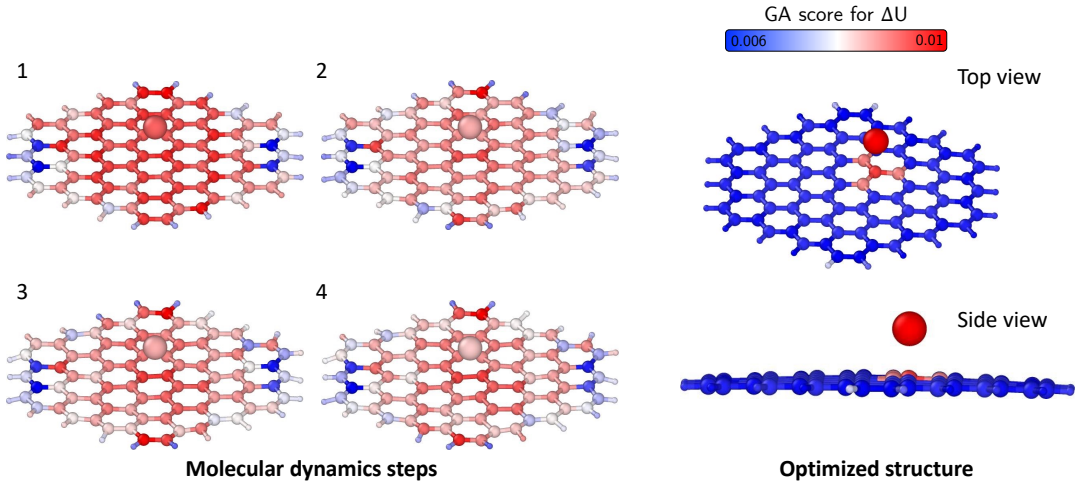


FIGURE 4.15: Visualizations of structure–property relationships in Pt/graphene structures at 4 consecutive molecular dynamics steps and optimized structure for ΔU property. The number in the top left corner indicate the index of the molecular dynamics steps.

Furthermore, Figure 4.15 presents four consecutive frames from the molecular dynamics simulation as from the flatten configuration of graphene structure. Throughout these frames, the GA scores dynamically adjust in response to the evolving structural features. Notably, the GA scores highlight that the regions of the graphene flake remaining relatively flat exhibit higher attention scores, indicating a significant contribution to the deformation energy. This observation is insightful because the actual optimized structure of the Pt/Graphene system features a slight curvature in the graphene sheet due to the adsorption of the platinum atom, rather than being perfectly flat as seen in the first simulation steps 1. More details about the visualization of attention for the molecular dynamics are shown in the videos at Appendix A.

The SCANN model effectively captures this subtle structural nuance by assigning higher GA scores to the flatter regions, which are under strain as they transition toward the curved optimized configuration. This alignment between the GA scores and the physical deformation highlights the model's capability to interpret and predict material instabilities accurately. The GA scores not only reflect the immediate structural distortions, such as elongated carbon–carbon bonds and convexed carbon configurations, but also provide a dynamic understanding of how these deformations contribute to the overall energy of the system during the simulation. These results validate the usefulness of SCANN in understanding and predicting structural deformations in materials, particularly in cases involving the interaction of different elements or adsorption onto surfaces.

4.7 Contributions and limitations

This study proposes SCANN, an attention-based DL architecture designed for material dataset analysis. SCANN leverages attention mechanisms to learn from material datasets, predict material properties, and interpret the underlying characteristics of material structures. By applying attention recursively to neighboring local structures, SCANN learns representations of atomic local structures in a self-consistent

manner. The architecture then combines these local structure representations to create a comprehensive representation of the entire material structure, enabling precise property predictions. During the learning process, global attention scores are estimated, indicating the importance of each local structure in representing the overall material structure. Experimental results based on five molecular and crystalline material structure datasets demonstrated the excellent predictive capability of SCANN for different material properties. Furthermore, an in-depth qualitative analysis of the global attention scores of local structures revealed that the trained models can extract essential information from material datasets, facilitating a deeper understanding of the structure–property relationships in both molecular and crystalline materials. The ability of the proposed architecture to interpret the attention scores can aid in identifying critical features and accelerating the material design process.

However, there are limitations to consider with the SCANN architecture. First, the performance of SCANN is highly dependent on the quality and diversity of the training datasets. If the datasets are limited in size or lack representation of certain material classes, the model may not generalize well to unseen materials. Second, while SCANN provides valuable interpretability through global attention scores, these scores can sometimes be challenging to correlate directly with specific physical or chemical properties without additional analysis (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Grimsley, Mayfield, and R.S. Bursten, 2020). Third, the computational complexity of attention mechanisms can lead to increased training times and resource requirements, especially for large-scale material datasets or when modeling materials with very complex structures. Despite these challenges, the findings of this study demonstrate the potential of attention mechanisms in uncovering valuable information that can provide a the better understanding of structure–property relationships in materials.

Chapter 5

Conclusion and Limitation

5.1 Conclusion

This thesis has discussed how deep learning frameworks can integrate with domain knowledge to solve some of the significant challenges in materials science. By systematically designing and tailoring deep learning models for specific scientific applications, we showcased how representation learning can be optimized to improve predictive accuracy and interpretability. The broad goal was to connect the dots between sophisticated computational methods and complex demands in material property prediction and diffraction image reconstruction for deeper insights and new understandings in research.

The focus of the first part of this work was on material property prediction using supervised deep learning methodologies. We designed deep learning frameworks that incorporate domain-specific knowledge, resulting in models that can not only make accurate predictions of various properties of materials but also can be used to provide meaningful interpretations of the underlying mechanisms. By incorporating domain-enriched representations, these models could capture far more sophisticated relationships in the data, which enhanced their predictive performance and provided further scientific insight. This may constitute a route to using supervised deep learning to advance materials informatics by providing tools that assist in the rational design and discovery of new materials possessing target properties.

The second part of this thesis focused on unsupervised deep learning approaches toward diffraction image reconstruction. Here, we designed models that could reconstruct high-fidelity patterns with no dependency on labeled data but, instead, exploited the intrinsic structures of the data diffraction itself. Intrinsic domain knowledge was inculcated into model architecture and training strategies such that the unsupervised models captured the rich dynamics of diffraction processes and principles of Fourier transformation, leading to accurate and reliable reconstructions of images. This work put into perspective the capability of unsupervised deep learning in handling high-dimensional and complex scientific data, therefore opening further exploratory and diagnostic tasks in materials science.

Finally, embedding domain knowledge into the deep learning framework presented a critical theme recurring throughout this work, strongly influencing model performance and interpretability. We developed means for embedding scientific principles and domain-specific constraints into the design of inputs, architectures, loss functions, and metrics for evaluation, thereby enhancing the generalization capabilities of the models, along with their robustness to data variability. The synergistic way this was done improved the accuracy and reliability of the predictions and reconstructions and ensured that the learned representations were scientifically coherent and meaningful. The successful fusion of deep learning with domain knowledge underlines the importance of an interdisciplinary approach, opening the future

toward advancements powered by computational intelligence and scientific expertise to make this innovation possible in materials science.

This thesis has shown how combining deep learning with domain-specific knowledge could transform some of the intractable problems in materials science. We have realized remarkable improvement in material property prediction and diffraction image reconstruction through appropriate supervised and unsupervised learning methods tailored by scientific insights. First, the methodologies and frameworks developed herein contribute to improving accuracy and interpretability in deep learning models and to broader scientific understanding and the discovery of new materials. In the future, this work's general principles and strategies lay a strong foundation for further exploration and application of deep learning in various scientific fields, fostering further innovation and progress.

5.2 Limitation

While this thesis has made significant strides in integrating deep learning frameworks with domain knowledge to enhance material property prediction and diffraction image reconstruction, several limitations must be considered. First, the models currently possess different levels of uncertainty in predictions and reconstructions. The supervised models, while achieving high accuracy in the prediction of material properties, usually do not have appropriate mechanisms for uncertainty quantification. This may further limit the model's predictability, mainly when new or unusual materials are analyzed where data sparsity is considerable. In addition, unsupervised models for diffraction image reconstruction may contain partial or incorrect reconstructions in the presence of increasingly complex material patterns and require the development of advanced uncertainty estimation techniques for evaluating the confidence levels of the obtained generated output.

Another limitation involves transforming and integrating domain knowledge into deep learning architectures. Although this work has successfully embedded material science principles into model design, the process is somewhat manual and heuristic. Choosing the most adequate domain-specific transformations and constraints is challenging since it usually requires thorough scientific domain knowledge and a deep understanding of the intricacies of deep learning methodologies. This can only be manually integrated, which may often lead to suboptimal representations and not capture material behaviors' complexities.

While promising, the reconstruction capabilities of unsupervised models are similarly bound by the quality and diversity of their training data. High-dimensional and complex diffraction images demand large and varied datasets so that models can generalize well. These datasets are often tricky and time-consuming to acquire, which can be a significant barrier to using the models. Moreover, current reconstruction models need more robustness and may easily break if applied to highly noisy or incomplete data, an inherent feature of most experimental setups. Overcoming these issues involves designing more robust methods for training and investigating advanced augmentation strategies that could further improve the model's robustness.

Future research efforts should improve uncertainty prediction and quantification in deep learning, supervised and unsupervised. Applying Bayesian deep learning techniques or introducing ensemble methods will make the uncertainty estimates more realistic and, therefore, gain more trust in model predictions or reconstructions. Another critical issue is the further development of methods that incorporate

domain knowledge. The research would also be considerably enhanced in applicability and impact by being done on an increased range of material properties and reconstruction tasks. Some potential future research might consider using transfer learning to adapt existing models to new material systems when only limited labeled data are available, improving the versatility of the models. Also, this would enable more successful and more varied datasets through collaborative work with experimentalists, letting the models learn from various material behaviors and diffraction patterns. Finally, the investigation of the integration of real-time data acquisition and model inference may enable dynamic and interactive material discovery platforms that will further accelerate the innovation cycle in materials science.

Chapter 6

Publication list

List of publications

- **Tien-Sinh Vu**, Minh-Quyet Ha, Duong-Nguyen Nguyen, Viet-Cuong Nguyen, Yukihiro Abe, Truyen Tran, Huan Tran, Hiori Kino, Takashi Miyake, Koji Tsuda and Hieu-Chi Dam. "Towards understanding structure–property relations in materials with interpretable deep learning", npj Computational Materials, volume 9, Article number: 215 (2023)
- **Tien-Sinh Vu**, Minh-Quyet Ha, Adam Mukharil Bachtiar, Duc-Anh Dao, Truyen Tran, Hiori Kino, Shuntaro Takazawa, Nozomu Ishiguro, Yuhei Sasaki, Masaki Abe, Hideshi Uematsu, Naru Okawa, Kyosuke Ozaki, Kazuo Kobayashi, Yoshiaki Honjo, Haruki Nishino, Yasumasa Joti, Takaki Hatsui, Yukio Takahashi and Hieu-Chi Dam. "PID3Net: a deep learning approach for single-shot coherent X-ray diffraction imaging of dynamic phenomena", npj Computational Materials, volume 11, Article number: 66 (2025)
- Adam Mukharil Bachtiar, **Tien-Sinh Vu**, Minh-Quyet Ha, Shuntaro Takazawa, Yukio Takahashi and Hieu-Chi Dam. "Impact of corner radius of the aperture on the accuracy of phase retrieval analysis in single-frame CDI for nanomaterials", 2024 Sanibel Proceedings, Molecular Physics, Article: e2388303. (2024)

List of published software

- The Python software for training the SCANN and other customized models have been deposited to a GitHub repository - `https://github.com/sinhvt3421/scann-material`.
- The Python software for training the PID3Net and other customized models have been deposited to a GitHub repository - `https://github.com/sinhvt3421/PID3Net`.

List of presentations

- **Tien-Sinh Vu**, Minh-Quyet Ha, Nguyen-Duong Nguyen, Hieu-Chi Dam, "Deep Learning Reveals Where To Pay Attention To For Specific Materials' Properties", "International Conference on Materials for Advanced Technologies 11th " (ICMAT 2023)
- **Tien-Sinh Vu**, Hieu-Chi Dam. "Deep learning for phase retrieval from virtual experiments in coherent X-ray diffraction imaging", Materials Innovation for Sustainable Development Goals, (MRM2023/IUMRS-ICA2023)

- **Tien-Sinh Vu**, Duy-Tai Dinh, Duong-Nguyen Nguyen, Hieu-Chi Dam, "An interpretable attention-based deep learning model for extracting material structure-property relationships", "International Chemical Congress of Pacific Basin Societies" (Pacifichem 2021)
- Duc-Anh Dao, **Tien-Sinh Vu**, Duong-Nguyen Nguyen, Ishizuka Keisuke, Oshima Yoshifumi, Tomitori Masahiko, Hieu-Chi DAM, "Elucidating atomic-scale phenomena with transmission electron microscopy and property-based sequential matching: a study of gold nanocontact", "International Chemical Congress of Pacific Basin Societies" (Pacifichem 2021)
- **Tien-Sinh Vu**, Duy-Tai Dinh, Duong-Nguyen Nguyen, Hieu-Chi Dam, "Deep attention model for extracting material structure-property relationships", "Conference on Computational Physics" (CCP 2021)
- Duc-Anh Dao, **Tien-Sinh Vu**, Duong-Nguyen Nguyen, Keisuke Ishizuka, Yoshifumi Oshima, Masahiko Tomitori, Hieu-Chi Dam, "Elucidating atomic-scale phenomena with transmission electron microscopy: a study of gold nanocontact", "Conference on Computational

Appendix A

Description of additional appendix movies for attention visualization

The supplementary materials include additional movies that provide dynamic visualizations of the GA score over time. These movies demonstrate the capabilities of our proposed reconstruction method SCANN in capturing the dynamic phenomena and structure-property relationship overtime

All the supplementary movies referenced in this study have been uploaded and are accessible at the following link: https://jstorage.box.com/v/VsinhThesisSuppl

- File Name: Appendix fullerene op homo

  Description: This is a time-series of Fullerene dynamics molecules with the visualization of GA score for HOMO property over time. The molecules are optimized to the stable structure over 20 steps.

- File Name: Appendix fullerene op lumo

  Description: This is a time-series of Fullerene dynamics molecules with the visualization of GA score for LUMO property over time. The molecules are optimized to the stable structure over 20 steps.

- File Name: Appendix ptgp op u

  Description: This is a time-series of Pt/graphene dynamics with the visualization of GA score for ΔU property over time. The molecules are optimized to the stable structure over 20 steps.

Appendix B

Setups of CXDI for experiments

B.1 CXDI experiments settings

Experimental CXDI: The experimental setup included a source that delivered an incident 5 keV monochromatic X-ray beam, which was shaped by a triangular aperture (side length: 10 μm). The triangular apertures were fabricated using focused ion beam processing on a 15-μm-thick platinum foil polished on both sides. A Fresnel zone plate was installed to reduce the triangular aperture size by a factor of two, resulting in a triangular X-ray beam of approximately 5 μm per side, which illuminated the sample. To inhibit X-ray scattering in air, all the optical elements, the sample, and the detector were enclosed in a vacuum chamber maintained at a pressure less than 1 Pa. The temperature increase owing to X-ray absorption in the solution was estimated to be less than 0.2 K s^{-1}. Given that the solution conducted heat via convection, the temperature increase due to X-ray irradiation was considered negligible.

In the first experiment, the probe functions were reconstructed using the mixed-state method (Takazawa et al., 2021; Li et al., 2016) via scanning CXDI with an exposure time of 10 s at each scan position. The sample was exposed to 15×15 overlapping fields of view, separated by 500 nm in the horizontal and vertical directions. As shown in Figure 3a, the probe functions are divided into five orthogonal modes. All probes feature a half-sized triangular aperture imaged using the FZP, with the first mode capturing 89.6 % of all photons. The intensities of the five modes probe were distributed as follows: 89.6, 4.4, 2.4, and 1.7 %. Subsequently, the model was applied to image the Ta test chart that was continuously translated against the same X-ray beam for single-shot CXDI. During these translations, the diffraction images generated from the illuminated area were continuously recorded at intervals of 7 ms for 15 s at 340 nm s^{-1}. The incident photon flux on the sample surface was maintained at 3 $\times$ 10^7 photons s^{-1}. The diffraction intensity images were recorded using an in-vacuum pixelated detector (CITIUS detector)(Takahashi et al., 2023; Ozaki et al., 2023) with a pixel size of 72.6 μm. The detector was positioned 3.30 m downstream from the sample.

In the third experiment, we reused four modes of the probe function, which were introduced in our previous study to image the dynamics of the gold nanoparticles in the solution (Takazawa et al., 2023). These probe functions are reconstructed through scanning CXDI. The distribution of these four modes accounted for 90.9, 5.4, 2.2, and 1.5 % of all the photons, which were orthogonal to each other in a single exposure. In the experiment, we used the gold to fabricate probe particles due to its chemical inertness, biocompatibility, and resistance to deformation, making it safe for biological systems and suitable for studying mechanical stress and strain in materials and cells. The AuNPs are fabricated with a diameter of 150 nm. A constant incident photon flux of approximately 3 $\times$ 10^6 photons s^{-1} was maintained on the sample surface.

TABLE B.1: The detailed settings of the simulation and experimental optical systems were used in first evaluation experiments. FZP stands for Fresnel zone plate.

		Experimental Ta chart
Optical	X-ray energy (keV)	5.0
	Focal length of FZP (mm)	60.49
	Aperture-FZP offset (μm)	100
	Aperture-FZP distance (mm)	181.47
	FZP-sample distance (mm)	90.74
	Sample-detector distance (m)	3.30
	Detector resolution (μm)	72.6
	Detector	CITIUS
	FZP manufacturer	XRnanotech
Measurement	Number of frames	1755
	Exposure time (ms)	7
	Velocity (nm s^{-1})	340
	Pixel resolution (nm $pixel^{-1}$)	29.8

The diffraction intensity images were recorded using an in-vacuum pixelated detector (EIGER 1M, Dectris) with a pixel size of 75 μm. The detector was positioned 3.14 m downstream from the sample.

Simulation of CXDI: A wave-optical simulation of the illumination optics was conducted in the second experiment, using an off-axis configuration FZP under the specified experimental conditions. This simulation was employed to evaluate the efficacy of our proposed phase retrieval method in imaging the motion of AuNPs before applying it in a real scenario. Thus, all simulation settings were aligned with those of the actual experiment described previously. AuNPs were simulated with a diameter of 300 nm, and their quantity was determined based on a 0.1 ratio for the entire simulated area. The simulation used an optical configuration featuring a photon energy of 5 keV and a photon flux of 3×10^6 photons s^{-1}, similar to the actual experiment. The diffraction intensity images were captured using consecutive images with an exposure time of 100 ms. Photon-counting noise with the Poisson statistics was also added to the diffraction images. In our previous study (Takazawa et al., 2023), the average velocity of AuNPs was reported as approximately 200 nm s^{-1} in the actual experiment, estimated using X-ray photon correlation spectroscopy (XPCS). Consistent with the aforementioned experimental observations, a controlled particle velocity of approximately 200 nm s^{-1} was applied in the simulation.

TABLE B.2: The detailed settings of the simulation and experimental optical systems were used in the second evaluation experiments. FZP stands for Fresnel zone plate.

		Simulation AuNPs
Optical	X-ray energy (keV)	5.0
	Focal length of FZP (mm)	48.38
	Aperture-FZP offset (μm)	100
	Aperture-FZP distance (mm)	145.1
	FZP-sample distance (mm)	72.6
	Sample-detector distance (m)	3.14
	Detector resolution (μm)	75
	Detector	–
	FZP manufacturer	–
Measurement	Number of frames	1000
	Exposure time (ms)	100
	Velocity (nm s^{-1})	200
	Pixel resolution (nm $pixel^{-1}$)	27.03

B.2 Impact of temporal block and measurement-informed refinement block

In this section, we studied the impact of the Temporal Block (TB) and the Refinement Block (RB) on phase retrieval in the CXDI experiment. Firstly, we analyzed the convergence of the phase retrieval process using our method PID3Net-P by deactivating the RB. The modified version is referred to as PID3Net-NR-P (No Refinement Block). We compared the performance of PID3Net-NR-P with AutoPhaseNN and PtychoNN. The PtychoNN model used a supervised learning approach with a traditional 2D CNN architecture to reconstruct the ground truth amplitude and phase from a single diffraction image without temporal information. Meanwhile, AutoPhaseNN employed X-ray Bragg coherent diffraction imaging with 3×3×3 CNN layers to reconstruct 3D gold crystals. On the other hand, PID3Net-NR-P learned the temporal information on multilevel through different kernels in temporal block (1×3×3, 3×3×3 and 5×3×3 CNN). Since the original designs of PtychoNN and AutoPhaseNN did not include the probe function with forward Fourier transformation, we customized these models to align the amplitude and phase decoder output with the self-supervised learning strategy and optics systems in the examination. The decoder outputs were utilized to calculate the numerical diffraction intensity and loss function.

Appendix Figure 1 showcases the reconstructed phase information from the measured diffraction of the Ta test chart using PtychoNN, AutoPhaseNN, PID3Net-NR-P, and PID3Net-P. Despite learning only from data without prior knowledge or mathematical constraints, these models captured the main patterns of the test chart.

TABLE B.3: The detailed settings of the simulation and experimental optical systems were used in the third evaluation experiments. FZP stands for Fresnel zone plate.

		Experimental AuNPs
Optical	X-ray energy (keV)	5.0
	Focal length of FZP (mm)	48.38
	Aperture-FZP offset (μm)	100
	Aperture-FZP distance (mm)	145.1
	FZP-sample distance (mm)	72.6
	Sample-detector distance (m)	3.14
	Detector resolution (μm)	75
	Detector	EIGER 1M
	FZP manufacturer	NTT-AT
Measurement	Number of frames	2000
	Exposure time (ms)	1000
	Velocity (nm s^{-1})	–
	Pixel resolution (nm $pixel^{-1}$)	27.03

However, the PtychoNN method exhibited relatively poor performance, resulting in distorted and twisted line patterns. In contrast, the AutoPhaseNN method, which employed $3\times3\times3$ CNN layers to capture temporal relationships between three time-evolving sequence of images, produced more stable reconstructions of the large fabricated lines. However, it still struggled with the tiny lines and central area due to complex symmetry patterns and multiple solutions of the phase retrieval problem.

PID3Net-NR-P, incorporating multilevel temporal learning, yielded more reasonable phase images with smooth transitions but also exhibited tiny twisted errors, particularly at frames 1000^{th} and 1200^{th} as shown in Appendix Figure 1 and Appendix Movie C5. This subpar performance can be attributed to the intrinsic ambiguity of the phase origin and the remarkably short exposure time in the phase problem. Notably, the quality of these reconstruction results was surpassed by those obtained using the full model PID3Net-P, underscoring the advantage of employing the RB layer with a measurement-informed process.

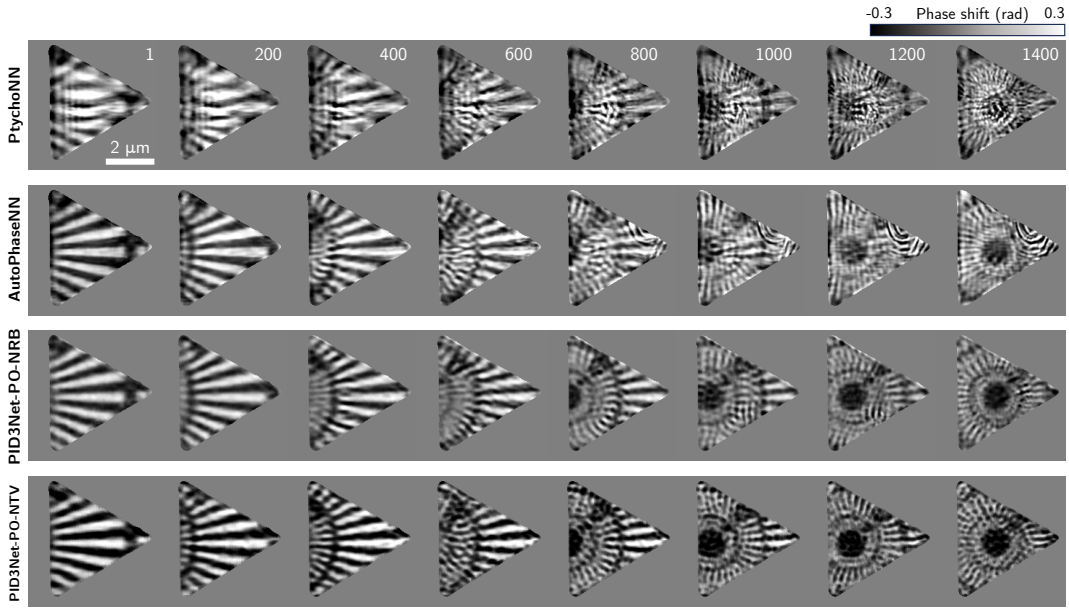


FIGURE B.1: Reconstructed phase information from the measured diffraction images of the moving Ta test chart by using the PtychoNN, AutoPhaseNN, PID3Net-NR-P and PID3Net-P methods

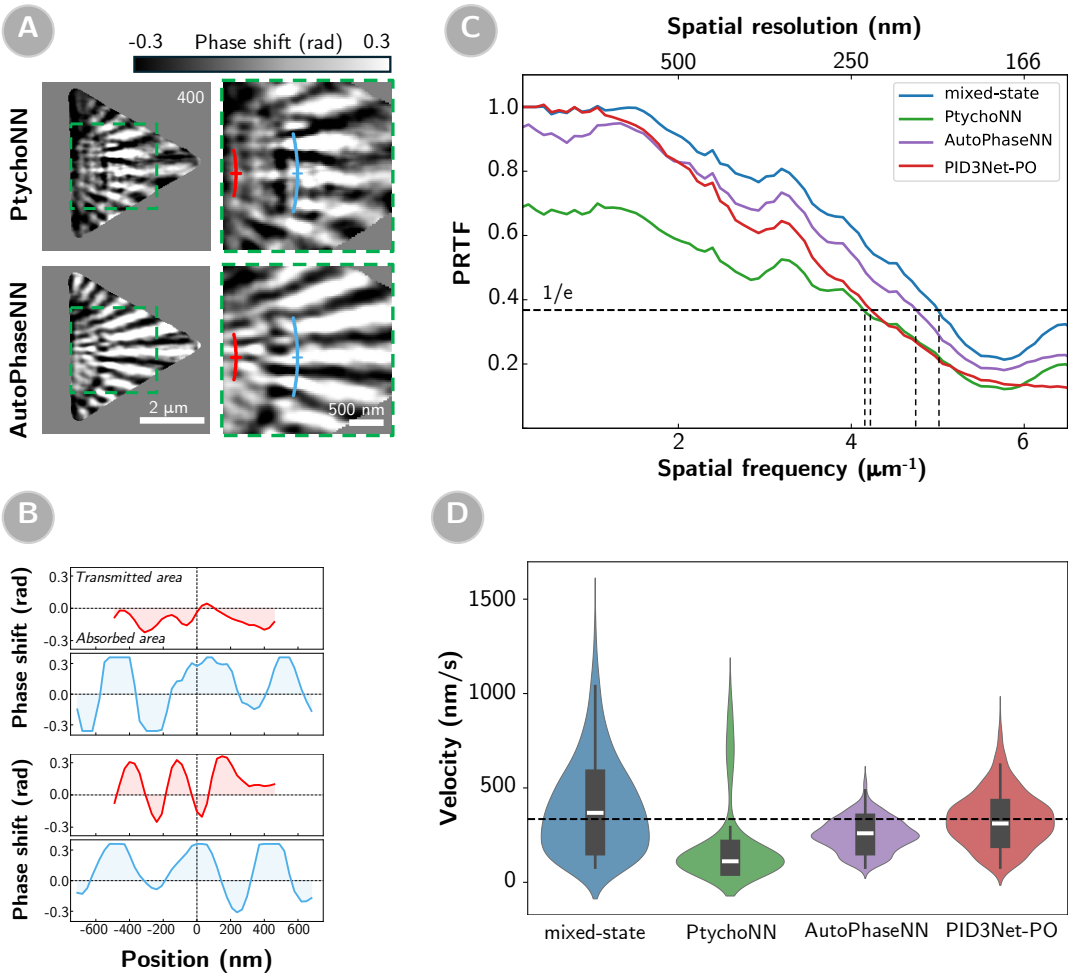


FIGURE B.2: (a) Right plots show phase information retrieved from diffraction intensity images measuring the moving of Ta test chart at the 400^{th} frame with a 7 ms exposure time using the PtychoNN and AutoPhaseNN methods. The frame index of each image is indicated in the top row. Left plots show the magnified views of the areas enclosed by the green squares at the 400^{th} frame along with profiles of two circular arcs. The horizontal mark at the middle of each arc indicates its zero position. (b) Analysis of the profiles of these two circular arcs in phase information retrieved from the measured diffraction intensity image at the 400^{th} frame. The phase shifts at different positions in these curved lines are monitored. (b) The phase retrieval transfer function (PRTF) analysis of the phase images reconstructed using the four phase retrieval methods. Dashed horizontal lines indicate the spatial resolutions at which the PRTF value falls below $1/e$, marking the threshold below which phase retrieval is less reliable. (c) The distribution of estimated velocity from the acquired phase images for the first 400 frames. The dashed line indicates the velocity is set in measurement at 340 nm/s, and the white bar represents the median of the distribution.

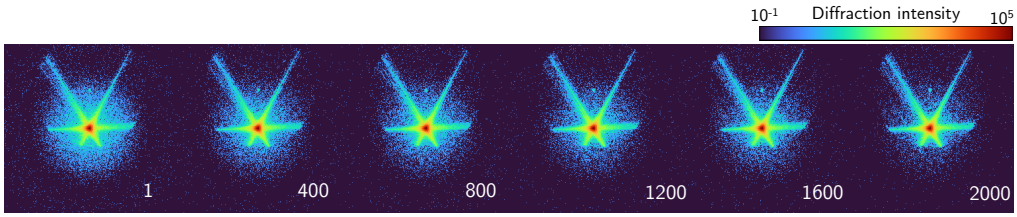


FIGURE B.3: Measured diffraction images of the AuNPs dispersed in the PVA solution. The frame index of each image is indicated in the right bottom. The color bar at the top indicates the diffraction intensity.

Appendix C

Description of additional appendix movies for phase retrieval

The supplementary materials include additional movies that provide dynamic visualizations of the reconstructed over time. These movies demonstrate the capabilities of our proposed reconstruction method, PID3Net-PO, in capturing the temporal evolution of the sample with enhanced accuracy and clarity.

All the supplementary movies referenced in this study have been uploaded and are accessible at the following link: https://www.nature.com/articles/s41524-025-01549-x

- File Name: Appendix Movie 1

  Description: This is a time-series of reconstructed frames of the Ta test chart observed using single-shot coherent diffraction imaging. The frames have been enlarged without interpolation for easier viewing. The white bar indicates the scale of the image. The movie has a frame rate of 30 Hz and consists of 1755 frames for a total 15 s. The read-out time when collecting frames is 15.677 μs, which is negligible compared to the effective exposure time of 7.2 ms.

- File Name: Appendix Movie 2

  Description: This is a time-series of diffraction intensity and reconstructed frames of the modified Ta test chart dataset. Following each set of 400 frames, the diffraction data was reversed for total 200 frames before being input into the model for reconstruction. The model subsequently proceeds to reconstruct the motion of the test chart in both forward and backward directions. The movie has a frame rate of 30 Hz.

- File Name: Appendix Movie 3

  Description: Time-series of reconstructed frames for numerical demonstration of single-shot coherent diffraction imaging. Details of the ground truth data are provided in the main text. The number of frames and exposure time were 1,000 and 100 ms, respectively. The movie has a frame rate of 30 Hz with 27.03 nm pixel^{-1} spatial resolution.

- File Name: Appendix Movie 4

  Description: Movie of gold colloidal particles in an aqueous polyvinyl alcohol solution reconstructed from diffraction patterns. The images are magnified for easier viewing. The scale bar size in the image is 2 μm. The recording consists of 2,000 frames with an exposure time of 1000 ms. The frame rate of the movie is 30 Hz and the resolution is 27.03 nm pixel^{-1}.

- File Name: Appendix Movie 5

Description: Reconstructed frames by PID3Net-NR-PO models for the Ta test chart observed using single-shot coherent diffraction imaging. The frames have been enlarged without interpolation for easier viewing. The white bar indicates the scale of the image. The movie has a frame rate of 30 Hz and consists of 1755 frames for a total 15 s.

Bibliography

Abbey, Brian (2013). "From Grain Boundaries to Single Defects: A Review of Coherent Methods for Materials Imaging in the X-ray Sciences". In: *JOM* 65.9, pp. 1183–1201. DOI: 10.1007/s11837-013-0702-4. URL: https://doi.org/10.1007/s11837-013-0702-4.

Abbey, Brian et al. (2008). "Keyhole coherent diffractive imaging". In: *Nature Physics* 4.5, pp. 394–398. DOI: 10.1038/nphys896. URL: https://doi.org/10.1038/nphys896.

Agrawal, Ankit and Alok Choudhary (Apr. 2016). "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science". In: *APL Materials* 4.5. 053208. ISSN: 2166-532X. DOI: 10.1063/1.4946894. URL: https://doi.org/10.1063/1.4946894.

Anderson, Brandon, Truong Son Hy, and Risi Kondor (2019). "Cormorant: Covariant Molecular Neural Networks". In: *NeurIPS* 32, pp. 14537–14546.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473. URL: https://api.semanticscholar.org/CorpusID:11212020.

Bengio, Yoshua (2009). "Learning Deep Architectures for AI". In: *Foundations and Trends® in Machine Learning* 2.1, pp. 1–127. ISSN: 1935-8237. DOI: 10.1561/2200000006. URL: http://dx.doi.org/10.1561/2200000006.

Bertasius, Gedas, Heng Wang, and Lorenzo Torresani (2021). "Is Space-Time Attention All You Need for Video Understanding?" In: *ArXiv* abs/2102.05095. URL: https://api.semanticscholar.org/CorpusID:231861462.

Bian, Liheng et al. (2016). "Fourier ptychographic reconstruction using Poisson maximum likelihood and truncated Wirtinger gradient". In: *Scientific Reports* 6, Art. No. 27384. ISSN: 2045-2322. DOI: 10.1038/srep27384.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387310738.

Bohra, Pakshal et al. (2023). "Dynamic Fourier ptychography with deep spatiotemporal priors". In: *Inverse Problems* 39.6, p. 064005.

Bridle, John S. (1990). "Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition". In: *Neurocomputing*. Ed. by Françoise Fogelman Soulié and Jeanny Hérault. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 227–236. ISBN: 978-3-642-76153-9.

Butler, Keith T et al. (2018). "Machine learning for molecular and materials science". In: *Nature* 559.7715, pp. 547–555. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0337-2.

Cao, Zhonglin et al. (2023). "MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction". In: *J. Am. Chem. Soc.* 145.5, pp. 2958–2967. ISSN: 0002-7863. DOI: 10.1021/jacs.2c11420.

Cha, Eunju et al. (2021). "Deepphasecut: Deep relaxation in phase for unsupervised fourier phase retrieval". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12, pp. 9931–9943.

Chambolle, Antonin (2004). "An algorithm for total variation minimization and applications". In: *Journal of Mathematical imaging and vision* 20, pp. 89–97.

Chapman, Henry N. and Keith A. Nugent (2010). "Coherent lensless X-ray imaging". In: *Nature Photonics* 4.12, pp. 833–839. DOI: 10.1038/nphoton.2010.240. URL: https://doi.org/10.1038/nphoton.2010.240.

Chapman, Henry N et al. (2006). "High-resolution ab initio three-dimensional x-ray diffraction microscopy". In: *JOSA A* 23.5, pp. 1179–1200.

Chen, Chi et al. (2019). "Graph networks as a universal machine learning framework for molecules and crystals". In: *Chem. Mater.* 31.9, pp. 3564–3572.

Chen, Mingqin et al. (2022a). "Unsupervised Phase Retrieval Using Deep Approximate MMSE Estimation". In: *IEEE Transactions on Signal Processing* 70, pp. 2239–2252. DOI: 10.1109/TSP.2022.3170710.

Chen, Pin et al. (2022b). "Interpretable Graph Transformer Network for Predicting Adsorption Isotherms of Metal–Organic Frameworks". In: *J. Chem. Inf. Model.* 62.22, pp. 5446–5456. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.2c00876.

Chen, Yuxin and Emmanuel Candes (2015). "Solving Random Quadratic Systems of Equations Is Nearly as Easy as Solving Linear Systems". In: 28. Ed. by C. Cortes et al. URL: https://proceedings.neurips.cc/paper_files/paper/2015/file/7380ad8a673226ae47fce7bff88e9c33-Paper.pdf.

Cherukara, Mathew J, Youssef SG Nashed, and Ross J Harder (2018). "Real-time coherent diffraction inversion using deep generative networks". In: *Scientific reports* 8.1, p. 16520.

Cherukara, Mathew J. et al. (July 2020). "AI-enabled high-resolution scanning coherent diffraction imaging". In: *Applied Physics Letters* 117.4, p. 044103. ISSN: 0003-6951. DOI: 10.1063/5.0013065. URL: https://doi.org/10.1063/5.0013065.

Chi, Dam Hieu et al. (2006). "Electronic structures of Pt clusters adsorbed on (5,5) single wall carbon nanotube". In: *Chem. Phys. Lett.* 432.1, pp. 213–217. ISSN: 0009-2614. DOI: https://doi.org/10.1016/j.cplett.2006.10.063.

Choudhary, Kamal and Brian DeCost (2021). "Atomistic Line Graph Neural Network for improved materials property predictions". In: *Npj Comput. Mater.* 7.1, p. 185. ISSN: 2057-3960. DOI: 10.1038/s41524-021-00650-1.

Chung, Junyoung et al. (2014). *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. arXiv: 1412.3555 [cs.NE]. URL: https://arxiv.org/abs/1412.3555.

Dam, Hieu Chi et al. (2009). "Substrate-mediated interactions of Pt atoms adsorbed on single-wall carbon nanotubes: Density functional calculations". In: *Phys. Rev. B* 79 (11), p. 115426. DOI: 10.1103/PhysRevB.79.115426.

Dan, Yabo et al. (2020). "Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials". In: *npj Computational Materials* 6.1, p. 84. ISSN: 2057-3960. DOI: 10.1038/s41524-020-00352-0. URL: https://doi.org/10.1038/s41524-020-00352-0.

Das, Kishalay et al. (2023). "CrysMMNet: Multimodal Representation for Crystal Property Prediction." In: *PMLR* 216, pp. 507–517.

Datta, Kaushik et al. (2019). "Computational requirements for real-time ptychographic image reconstruction". In: *Appl. Opt.* 58.7, B19–B27. DOI: 10.1364/AO.58.000B19. URL: https://opg.optica.org/ao/abstract.cfm?URI=ao-58-7-B19.

Duvenaud, David K et al. (2015). "Convolutional Networks on Graphs for Learning Molecular Fingerprints". In: *NeurIPS* 28, pp. 2224–2232.

Faber, Felix A et al. (2017). "Prediction errors of molecular machine learning models lower than hybrid DFT error". In: *J. Chem. Theory Comput.* 13.11, pp. 5255–5264.

Fienup, James R (1982). "Phase retrieval algorithms: a comparison". In: *Applied optics* 21.15, pp. 2758–2769.

Fuchs, Fabian B. et al. (2020). "SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks". In: *NeurIPS* 33, pp. 1970–1981.

Fung, Victor et al. (2021). "Benchmarking graph neural networks for materials chemistry". In: *Npj Comput. Mater.* 7.1, p. 84. ISSN: 2057-3960. DOI: 10.1038/s41524-021-00554-0.

Gilmer, Justin et al. (2017). "Neural message passing for quantum chemistry". In: *ICML* 70, 1263–1272.

Glorot, Xavier and Yoshua Bengio (2010). "Understanding the difficulty of training deep feedforward neural networks". In: *PMLR* 9, pp. 249–256.

Goodfellow, Ian J. et al. (2014). *Generative Adversarial Networks*. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

Grimsley, Christopher, Elijah Mayfield, and Julia R.S. Bursten (May 2020). "Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models". In: *LREC* 12, pp. 1780–1790.

Gugel, Leon and Shai Dekel (2022). "Pr-dad: phase retrieval using deep auto-decoders". In: *arXiv preprint arXiv:2204.09051*.

Guizar-Sicairos, Manuel and James R Fienup (2012). "Understanding the twin-image problem in phase retrieval". In: *JOSA A* 29.11, pp. 2367–2375.

Gunning, David et al. (2019). "XAI–Explainable artificial intelligence". In: *Sci. Robot.* 4.37, eaay7120. DOI: 10.1126/scirobotics.aay7120.

Haan, Kevin de et al. (2019). "Resolution enhancement in scanning electron microscopy using deep learning". In: *Scientific Reports* 9.1, p. 12050. ISSN: 2045-2322. DOI: 10.1038/s41598-019-48444-2. URL: https://doi.org/10.1038/s41598-019-48444-2.

Han, Song, Huizi Mao, and William J. Dally (2016). *Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding*. arXiv: 1510.00149 [cs.CV]. URL: https://arxiv.org/abs/1510.00149.

He, Kaiming et al. (2015). *Deep Residual Learning for Image Recognition*. arXiv: 1512.03385 [cs.CV]. URL: https://arxiv.org/abs/1512.03385.

Himanen, Lauri et al. (2019). "Data-Driven Materials Science: Status, Challenges, and Perspectives". In: *Adv. Sci.* 6.21, p. 1900808. DOI: https://doi.org/10.1002/advs.201900808.

Hirn, Matthew, Stéphane Mallat, and Nicolas Poilvert (2017). "Wavelet scattering regression of quantum chemical energies". In: *Multiscale Model. Simul.* 15.2, pp. 827–863.

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hore, Alain and Djemel Ziou (2010). "Image quality metrics: PSNR vs. SSIM". In: *2010 20th international conference on pattern recognition*. IEEE, pp. 2366–2369.

Jain, Sarthak and Byron C. Wallace (2019). "Attention is not Explanation". In: *arXiv* 1902, 10186, Preprint at http://arxiv.org/abs/1902.10186.

Jha, Dipendra et al. (2018). "ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition". In: *Scientific Reports* 8.1, p. 17593. ISSN: 2045-2322. DOI: 10.1038/s41598-018-35934-y. URL: https://doi.org/10.1038/s41598-018-35934-y.

Johnson, Allan S. et al. (2023). "Ultrafast X-ray imaging of the light-induced phase transition in VO2". In: *Nature Physics* 19.2, pp. 215–220. DOI: 10.1038/s41567-022-01848-w. URL: https://doi.org/10.1038/s41567-022-01848-w.

Kang, Jungmin et al. (2021). "Single-frame coherent diffraction imaging of extended objects using triangular aperture". In: *Optics Express* 29.2, pp. 1441–1453.

Kang, Yeonghun et al. (2023). "A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks". In: *Nat. Mach. Intell.* 5.3, pp. 309–318. DOI: 10.1038/s42256-023-00628-2.

Kappeler, Armin et al. (2017). "Ptychnet: CNN based Fourier ptychography". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 1712–1716.

Karamad, Mohammadreza et al. (2020). "Orbital graph convolutional neural network for material property prediction". In: *Phys. Rev. Mater.* 4.9, p. 093801.

Keller, Joseph B (1957). "Diffraction by an aperture". In: *Journal of Applied Physics* 28.4, pp. 426–444.

Khakurel, Krishna P. et al. (2017). "Generation of apodized X-ray illumination and its application to scanning and diffraction microscopy". In: *Journal of Synchrotron Radiation* 24.1, pp. 142–149. DOI: 10.1107/S1600577516017677. URL: https://doi.org/10.1107/S1600577516017677.

Kimanius, Dari et al. (2024). "Data-driven regularization lowers the size barrier of cryo-EM structure determination". In: *Nature Methods* 21.7, pp. 1216–1221. ISSN: 1548-7105. DOI: 10.1038/s41592-024-02304-8. URL: https://doi.org/10.1038/s41592-024-02304-8.

Kingma, Diederik and Jimmy Ba (2017). "Adam: A Method for Stochastic Optimization". In: *arXiv* 1412, 6980, Preprint at https://arxiv.org/abs/1412.6980.

Kliuiev, Pavel et al. (2016). "Application of iterative phase-retrieval algorithms to ARPES orbital tomography". In: *New Journal of Physics* 18.9, p. 093041.

Korolev, Vadim and Pavel Protsenko (2023). "Accurate, interpretable predictions of materials properties within transformer language models". In: *Patterns* 4, p. 100803. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2023.100803.

Kourousias, George et al. (2018). "Monitoring dynamic electrochemical processes with in situ ptychography". In: *Applied Nanoscience* 8, pp. 627–636.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (May 2017). "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6, 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

Latychevskaia, Tatiana (2018). "Iterative phase retrieval in coherent diffractive imaging: practical issues". In: *Applied optics* 57.25, pp. 7187–7197.

Lecun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Lee, Sung Yun et al. (2021). "Denoising low-intensity diffraction signals using k-space deep learning: Applications to phase recovery". In: *Physical Review Research* 3.4, p. 043066.

Levitan, Abraham L. et al. (2020). "Single-frame far-field diffractive imaging with randomized illumination". In: *Opt. Express* 28.25, pp. 37103–37117. DOI: 10.1364/OE.397421. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-28-25-37103.

Li, Peng et al. (2016). "Breaking ambiguities in mixed state ptychography". In: *Optics express* 24.8, pp. 9038–9052.

Lo, Yuan Hung et al. (2018). "In situ coherent diffractive imaging". In: *Nature Communications* 9.1, p. 1826. DOI: 10.1038/s41467-018-04259-9. URL: https://doi.org/10.1038/s41467-018-04259-9.

Lucas, Alice et al. (2018). "Using Deep Neural Networks for Inverse Problems in Imaging: Beyond Analytical Methods". In: *IEEE Signal Processing Magazine* 35.1, pp. 20–36. DOI: 10.1109/MSP.2017.2760358.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning (2015). "Effective Approaches to Attention-based Neural Machine Translation". In: *CoRR* abs/1508.04025. arXiv: 1508.04025. URL: http://arxiv.org/abs/1508.04025.

Maiden, Andrew M and John M Rodenburg (2009). "An improved ptychographical phase retrieval algorithm for diffractive imaging". In: *Ultramicroscopy* 109.10, pp. 1256–1262.

Mantero, Vittorio et al. (2021). "COVID-19 in dimethyl fumarate-treated patients with multiple sclerosis". In: *J. Neurol.* 268.6, pp. 2023–2025. DOI: 10.1007/s00415-020-10015-1.

Marchesini, S (2004). "Benchmarking iterative projection algorithms for phase retrieval". In: *arXiv preprint physics/0404091*.

Marchesini, Stefano et al. (2003). "X-ray image reconstruction from a diffraction pattern alone". In: *Physical Review B* 68.14, p. 140101.

Metzler, Christopher et al. (2018). "prDeep: Robust phase retrieval with a flexible deep network". In: *International Conference on Machine Learning*. PMLR, pp. 3501–3510.

Miao, Jianwei et al. (2006). "Three-dimensional GaN- Ga 2 O 3 core shell structure revealed by X-ray diffraction microscopy". In: *Physical review letters* 97.21, p. 215503.

Miao, Jianwei et al. (2015a). "Beyond crystallography: Diffractive imaging using coherent x-ray light sources". In: *Science* 348.6234, pp. 530–535. DOI: 10.1126/science.aaa1394. eprint: https://www.science.org/doi/pdf/10.1126/science.aaa1394. URL: https://www.science.org/doi/abs/10.1126/science.aaa1394.

Miao, Jianwei et al. (2015b). "Beyond crystallography: Diffractive imaging using coherent x-ray light sources". In: *Science* 348.6234, pp. 530–535.

Moran, Michael et al. (2023). "Site-Net: using global self-attention and real-space supercells to capture long-range interactions in crystal structures". In: *Digit. Discov.* 2, pp. 1297–1310. DOI: 10.1039/D3DD00005B.

Nair, Vinod and Geoffrey E. Hinton (2010). "Rectified linear units improve restricted boltzmann machines". In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 807–814. ISBN: 9781605589077.

Nashed, Youssef SG et al. (2014). "Parallel ptychographic reconstruction". In: *Optics express* 22.26, pp. 32082–32097.

Nguyen, Duong-Nguyen et al. (2023). "Explainable active learning in investigating structure–stability of SmFe12-α-βXαYβ structures X, Y {Mo, Zn, Co, Cu, Ti, Al, Ga}". In: *MRS Bulletin* 48.1, pp. 31–44. ISSN: 1938-1425. DOI: 10.1557/s43577-022-00372-9. URL: https://doi.org/10.1557/s43577-022-00372-9.

Ohara, Takashi et al. (2020). "Acrylic Acid and Derivatives". In: *Ullmann's Encyclopedia of Industrial Chemistry*, pp. 1–21. DOI: https://doi.org/10.1002/14356007.a01_161.pub4.

O'Keeffe, M. (1979). "A proposed rigorous definition of coordination number". In: *Acta Crystallogr. A: Found. Adv.* 35.5, pp. 772–775. DOI: 10.1107/S0567739479001765.

Ozaki, K. et al. (2023). "A 17400 frame/s X-ray imaging detector CITIUS with a linear response up to 945 Mcps/pixel (18 Tcps/cm^2): evaluation in ptychography". In: *Acta Crystallographica Section A* 79.a2, p. C1240. DOI: 10.1107/S205327332308381X. URL: https://doi.org/10.1107/S205327332308381X.

Pedersen, A. F. et al. (2020). "X-ray coherent diffraction imaging with an objective lens: Towards three-dimensional mapping of thick polycrystals". In: *Phys. Rev. Res.* 2 (3), p. 033031. DOI: 10.1103/PhysRevResearch.2.033031. URL: https://link.aps.org/doi/10.1103/PhysRevResearch.2.033031.

Peng, Xi et al. (2022). "DeepSENSE: Learning coil sensitivity functions for SENSE reconstruction using deep learning". In: *Magnetic Resonance in Medicine* 87.4, pp. 1894–1902. DOI: https://doi.org/10.1002/mrm.29085. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.29085. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.29085.

Pfeiffer, Franz (2018). "X-ray ptychography". In: *Nature Photonics* 12.1, pp. 9–17.

Pham, Tien Lam et al. (2017). "Machine learning reveals orbital interaction in materials". In: *Sci. Technol. Adv. Mater.* 18.1, p. 756.

Pham, Tien-Lam et al. (2019). "Learning Materials Properties from Orbital Interactions". In: *J. Phys.: Conf. Ser.* 1290.1, p. 012012.

Rahaman, Obaidur and Alessio Gagliardi (2020). "Deep Learning Total Energies and Orbital Energies of Large Organic Molecules Using Hybridization of Molecular Fingerprints". In: *J. Chem. Inf. Model.* 60.12, pp. 5971–5983. DOI: 10.1021/acs.jcim.0c00687.

Raissi, M., P. Perdikaris, and G.E. Karniadakis (2019). "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378, pp. 686–707. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.10.045. URL: https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Ramachandran, Prajit, Barret Zoph, and Quoc V. Le (2017). "Swish: a Self-Gated Activation Function". In: *arXiv: Neural and Evolutionary Computing*.

Ramakrishnan, Raghunathan et al. (2014). "Quantum chemistry structures and properties of 134 kilo molecules". In: *Sci. data* 1.1, pp. 1–7.

Ramprasad, Rampi et al. (2017). "Machine learning in materials informatics: recent applications and prospects". In: *Npj Comput. Mater.* 3.1, p. 54. ISSN: 2057-3960. DOI: 10.1038/s41524-017-0056-5.

Robinson, Ian and Ross Harder (2009). "Coherent X-ray diffraction imaging of strain at the nanoscale". In: *Nature Materials* 8.4, pp. 291–298. DOI: 10.1038/nmat2400. URL: https://doi.org/10.1038/nmat2400.

Rodenburg, John and Andrew Maiden (2019). "Ptychography". In: *Springer Handbook of Microscopy*, pp. 819–904.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Ed. by Nassir Navab et al. Cham: Springer International Publishing, pp. 234–241. ISBN: 978-3-319-24574-4.

Rosenblatt, F (1958). *The perceptron: A probabilistic model for information storage and organization in the brain*. US. DOI: 10.1037/h0042519.

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747*.

Rupp, Matthias et al. (2012). "Fast and accurate modeling of molecular atomization energies with machine learning". In: *Phys. Rev. Lett.* 108.5, p. 058301.

Schulze-Topphoff, Ulf et al. (2016). "Dimethyl fumarate treatment induces adaptive and innate immune modulation independent of Nrf2". In: *PNAS* 113.17, pp. 4777–4782. DOI: 10.1073/pnas.1603907113.

Schütt, Kristof T et al. (2014). "How to represent crystal structures for machine learning: Towards fast prediction of electronic properties". In: *Phys. Rev. B* 89.20, p. 205118.

Schütt, Kristof T et al. (2018). "SchNet–A deep learning architecture for molecules and materials". In: *J. Chem. Phys* 148.24, p. 241722.

Schweidtmann, Artur M. et al. (2023). "Physical pooling functions in graph neural networks for molecular property prediction". In: *Comput. Chem. Eng.* 172, p. 108202. ISSN: 0098-1354.

Sekiguchi, Yuki, Tomotaka Oroguchi, and Masayoshi Nakasako (2016). "Classification and assessment of retrieved electron density maps in coherent X-ray diffraction imaging using multivariate analysis". In: *Journal of Synchrotron Radiation* 23.1, pp. 312–323.

Shamshad, Fahad, Farwa Abbas, and Ali Ahmed (2019). "Deep ptych: Subsampled fourier ptychography using generative priors". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7720–7724.

Simonyan, Karen and Andrew Zisserman (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv: 1409.1556 [cs.CV]. URL: https://arxiv.org/abs/1409.1556.

Siriwardane, Edirisuriya M Dilanga et al. (2022). "Generative design of stable semiconductor materials using deep learning and density functional theory". In: *Npj Comput. Mater.* 8.1, p. 164. ISSN: 2057-3960. DOI: 10.1038/s41524-022-00850-3.

Sun, Tao et al. (2012). "Three-dimensional coherent X-ray surface scattering imaging near total external reflection". In: *Nature Photonics* 6.9, pp. 586–590. DOI: 10.1038/nphoton.2012.178. URL: https://doi.org/10.1038/nphoton.2012.178.

Takahashi, Yukio et al. (2023). "High-resolution and high-sensitivity X-ray ptychographic coherent diffraction imaging using the CITIUS detector". In: *Journal of Synchrotron Radiation* 30.5, pp. 989–994. DOI: 10.1107/S1600577523004897. URL: https://doi.org/10.1107/S1600577523004897.

Takayama, Yuki et al. (2021a). "Dynamic nanoimaging of extended objects via hard X-ray multiple-shot coherent diffraction with projection illumination optics". In: *Communications Physics* 4.1, p. 48.

— (2021b). "Dynamic nanoimaging of extended objects via hard X-ray multiple-shot coherent diffraction with projection illumination optics". In: *Communications Physics* 4.1, p. 48.

Takazawa, Shuntaro et al. (2021). "Demonstration of single-frame coherent X-ray diffraction imaging using triangular aperture: Towards dynamic nanoimaging of extended objects". In: *Optics Express* 29.10, pp. 14394–14402.

Takazawa, Shuntaro et al. (2023). "Coupling x-ray photon correlation spectroscopy and dynamic coherent x-ray diffraction imaging: Particle motion analysis from nano-to-micrometer scale". In: *Physical Review Research* 5.4, p. L042019.

Thibault, Pierre and Andreas Menzel (2013). "Reconstructing state mixtures from diffraction measurements". In: *Nature* 494.7435, pp. 68–71.

Tran, Du et al. (2018). "A Closer Look at Spatiotemporal Convolutions for Action Recognition". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459. DOI: 10.1109/CVPR.2018.00675.

Tripathi, Ashish, Ian McNulty, and Oleg G Shpyrko (2014). "Ptychographic overlap constraint errors and the limits of their numerical recovery using conjugate gradient descent methods". In: *Optics Express* 22.2, pp. 1452–1466.

Ulvestad, A. et al. (2015). "Topological defect dynamics in operando battery nanoparticles". In: *Science* 348.6241, pp. 1344–1347. DOI: 10.1126/science.aaa1313. eprint: https://www.science.org/doi/pdf/10.1126/science.aaa1313. URL: https://www.science.org/doi/abs/10.1126/science.aaa1313.

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *NeurIPS* 30, pp. 6000–6010.

Vu, Tien-Sinh and Dam Hieu Chi (Apr. 2023). "Fullerene structures and Pt absorbed on Graphene structures with HOMO, LUMO and Total energy properties". Version 1.0. In: *Zenodo*, https://zenodo.org/record/7792716.

Wakonig, Klaus et al. (2019). "X-ray Fourier ptychography". In: *Science advances* 5.2, eaav0282.

Wang, Ge, Jong Chul Ye, and Bruno De Man (2020). "Deep learning for tomographic image reconstruction". In: *Nature Machine Intelligence* 2.12, pp. 737–748. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00273-z. URL: https://doi.org/10.1038/s42256-020-00273-z.

Wang, Luzhi et al. (Jan. 2024). "Contrastive Graph Similarity Networks". In: *ACM Trans. Web* 18.2. ISSN: 1559-1131. DOI: 10.1145/3580511. URL: https://doi.org/10.1145/3580511.

Wang, Yujie et al. (2008). "Ultrafast X-ray study of dense-liquid-jet flow dynamics using structure-tracking velocimetry". In: *Nature Physics* 4.4, pp. 305–309. DOI: 10.1038/nphys840. URL: https://doi.org/10.1038/nphys840.

Ward, Logan and Chris Wolverton (2017). "Atomistic calculations and materials informatics: A review". In: *Curr. Opin. Solid State Mater. Sci.* 21.3, pp. 167–176.

Welker, Simon et al. (2022). "Deep iterative phase retrieval for ptychography". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1591–1595.

Wengrowicz, Omri et al. (2020). "Deep neural networks in single-shot ptychography". In: *Optics Express* 28.12, pp. 17511–17520.

Wiegreffe, Sarah and Yuval Pinter (2019). "Attention is not not Explanation". In: *arXiv* 1908, 04626, Preprint at https://arxiv.org/abs/1908.04626.

Withers, Philip J. et al. (2021). "X-ray computed tomography". In: *Nature Reviews Methods Primers* 1.1, p. 18. DOI: 10.1038/s43586-021-00015-4. URL: https://doi.org/10.1038/s43586-021-00015-4.

Wu, Longlong et al. (2021a). "Three-dimensional coherent x-ray diffraction imaging via deep convolutional neural networks". In: *npj Computational Materials* 7.1, p. 175.

Wu, Zhenqin et al. (2018). "MoleculeNet: a benchmark for molecular machine learning". In: *Chem. Sci.* 9.2, pp. 513–530.

Wu, Zonghan et al. (2021b). "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Trans. Neural Netw. Learn. Syst.* 32.1, pp. 4–24. DOI: 10.1109/TNNLS.2020.2978386.

Xie, Tian and Jeffrey C Grossman (2018). "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties". In: *Phys. Rev. Lett.* 120.14, p. 145301.

Xu, Keyulu et al. (2019). "How Powerful are Graph Neural Networks?" In: *arXiv* 1810, 00826, Preprint at http://arxiv.org/abs/1810.00826.

Yang, Kevin et al. (2019). "Analyzing Learned Molecular Representations for Property Prediction". In: *J. Chem. Inf. Model.* 59.8, pp. 3370–3388. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.9b00237.

Yao, Yudong et al. (2022a). "AutoPhaseNN: unsupervised physics-aware deep learning of 3D nanoscale Bragg coherent diffraction imaging". In: *npj Computational Materials* 8.1, p. 124. ISSN: 2057-3960. DOI: 10.1038/s41524-022-00803-w. URL: https://doi.org/10.1038/s41524-022-00803-w.

— (2022b). "AutoPhaseNN: unsupervised physics-aware deep learning of 3D nanoscale Bragg coherent diffraction imaging". In: *npj Computational Materials* 8.1, p. 124.

Yau, Allison et al. (2017). "Bragg coherent diffractive imaging of single-grain defect dynamics in polycrystalline films". In: *Science* 356.6339, pp. 739–742. DOI: 10.1126/science.aam6168. eprint: https://www.science.org/doi/pdf/10.1126/science.aam6168. URL: https://www.science.org/doi/abs/10.1126/science.aam6168.

Ye, Qiuliang, Li-Wen Wang, and Daniel PK Lun (2022). "SiSPRNet: end-to-end learning for single-shot phase retrieval". In: *Optics Express* 30.18, pp. 31937–31958.

— (2023). "Towards practical single-shot phase retrieval with physics-driven deep neural network". In: *Optics Express* 31.22, pp. 35982–35999.

Zhang, Fucai et al. (2016). "Phase retrieval by coherent modulation imaging". In: *Nature Communications* 7.1, p. 13367. DOI: 10.1038/ncomms13367. URL: https://doi.org/10.1038/ncomms13367.

Zhang, Yuhe et al. (2021a). "PhaseGAN: a deep-learning phase-retrieval approach for unpaired datasets". In: *Optics express* 29.13, pp. 19593–19604.

— (2021b). "PhaseGAN: a deep-learning phase-retrieval approach for unpaired datasets". In: *Opt. Express* 29.13, pp. 19593–19604. DOI: 10.1364/OE.423222. URL: https://opg.optica.org/oe/abstract.cfm?URI=oe-29-13-19593.

Zhao, Yong et al. (2023). "Physics guided deep learning for generative design of crystal materials with symmetry constraints". In: *Npj Comput. Mater.* 9.1, p. 38. ISSN: 2057-3960. DOI: 10.1038/s41524-023-00987-9.

Zhuang, Zhong et al. (2022). "Practical phase retrieval using double deep image priors". In: *arXiv preprint arXiv:2211.00799*.