Title
Data Driven Approaches for Characterization Techniques with Applications in Materials Chemistry Discovery

Author
Liu, Shuai

Publication Date
2019

Peer reviewed|Thesis/dissertation

Data Driven Approaches for Characterization Techniques with Applications in Materials Chemistry Discovery

by

Shuai Liu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Dr. Alexander Hexemer, Co-chair
Professor Teresa Head-Gordon, Co-chair
Professor Martin Head-Gordon
Professor Mark Asta

Summer 2019

Data Driven Approaches for Characterization Techniques with Applications in Materials Chemistry Discovery

Abstract

Data Driven Approaches for Characterization Techniques with Applications in Materials Chemistry Discovery

by

Shuai Liu

Doctor of Philosophy in Chemistry

University of California, Berkeley

Dr. Alexander Hexemer, Co-chair

Professor Teresa Head-Gordon, Co-chair

Understanding the structure of chemical compounds and nanoscale materials is critical for materials chemistry discovery. In the context of high-throughput technology, automatic chemical synthesis and advanced robotic characterization techniques have been applied to many systems. In contrast, there have been very few explorations to improve and accelerate the process of understanding characterized data, which becomes the bottleneck of next generation materials chemistry discovery. In this dissertation, I combine the experimental and data driven approaches for characterization techniques to perform chemistry-structure relationship understanding, data analysis and management, structure identification, computational prediction and facility optimization. These developments aim to accelerate the characterization processes of materials chemistry systems.

The first part of this dissertation describes the motivation and necessary background. In the second part of this dissertation, I demonstrate a framework for characterizing materials chemistry systems by combining experimental methods and data driven approaches, using X-ray scattering as the characterization technique. There are three aspects to consider within this framework: the experimental methods, the automatic data categorization workflow, and application of machine learning models. Experimental characterization is an important and essential part in this framework. In chapter 3, I study the chemistry-structure-property relationship of a supramolecular system using X-ray scattering. In addition, this chapter describes the experimental data collection and conventional data interpretation process. However, the conventional method is not compatible with high-throughput materials chemistry discovery. To address this bottleneck, in chapter 4, I build a large-scale database and propose a machine learning-based hierarchical method for X-ray scattering data categorization toward high-throughput data analysis. Using the data in chapter 3 as an example, I demonstrate that this method can be potentially utilized in materials chemistry discovery. In many cases, labeling experimental X-ray scattering data requires extensive human input. In chapter 5, I simulate millions of X-ray scattering data to train machine learning models. With this high-quality large-scale dataset, I analyze the performance of machine learning model

under different physical parameters and provide the interpretations of the prediction results.

The third part of this dissertation extends the data driven approaches to other characterization problems in materials chemistry. While the X-ray scattering technique is very powerful, it might not be sufficient to fully characterize all materials chemistry systems due to challenges such as low sensitivity to hydrogen and beam source instability. Nuclear Magnetic Resonance (NMR) is a complementary technique, in that its elemental sensitivities are very different, with better resolution for hydrogen in particular. In chapter 6, I use a deep learning method to predict chemical shifts in NMR crystallography. In comparison to the state-of-art DFT method, the deep learning method is significantly faster for large systems. Moreover, the prediction errors are lower than reported kernel ridge regression method. To improve source stability and characterization data quality, in chapter 7, I demonstrate a model-independent characterization facility optimization method using machine learning. The beam size variance is reduced using the neural network based feed-forward method.

Dedicated to my parents.

Contents

List of Figures

List of Tables

Acknowledgments

It was like a dream for my five years at Berkeley. I owe the word “thanks” to many people at Berkeley so I feel very fortunate to have this opportunity to acknowledge everyone who helped me in these five years. First, I would like to express my deepest gratitude to my research advisors and/or dissertation chairs: Dr. Alexander Hexemer, Dr. Daniela Ushizima and Professor Teresa Head-Gordon for their mentorship and support. I am deeply impressed by their enthusiasms in science and their professions in their research area. They provided me many helps on my research and dissertation writing process. They are all highly esteemed scholars in their research area. They are willing to devote a lot of time to give me the guidance throughout the projects. Moreover, they give me a lot of freedom to explore and highly respect my ideas, which allows me to become an independent researcher. I learned how to think creatively, address the problems uniquely, conduct the experiments systematically, and summarize the results scholarly. Moreover, I feel very lucky to explore this interdisciplinary research area between materials chemistry and machine learning under their guidance, since very few PhD students have this opportunity to explore this interesting but possibly mysterious research area. The PhD training under their guidance will be one of the most valuable experiences in my life, which help me well-prepared for the future career.

Second, I would like to thank my dissertation committee: Professor Martin Head-Gordon and Professor Mark Asta for being my dissertation committee member. They are famous experts in materials science and chemistry and I am very fortunate to get their helps on my dissertation. Moreover, I would like to thank my qualifying exam committee: Professor Felix Fischer, Professor Matthew Francis and Professor Richard Andersen. Especially, I would like to thank Professor Felix Fischer for taking the responsibility as my qualifying exam committee chair. He also provided me many guidance and helps in the later years. I highly appreciate the helps from my committee members, which allows me to become a qualified Berkeley PhD.

Third, I would like to thank all the professors and staffs in Chemistry department who helped me during my graduate career. Professor Richard Saykally, Professor Naomi Ginsberg, Professor Eran Rabani, Professor Sung-Hou Kim and Professor Berend Smit helped me identify my dissertation committee. I would also like to give a special thank to Lynn and the department chair, Professor Matthew Francis, who helped me out when everything was in a mess. Another special thank is to Professor Ting Xu. First, I would like to thank her for support during the first a couple of years. Second, I would like to thank her for helping me on the writing process of several manuscripts and published papers, which contributes to an important chapter of this dissertation. Last, I feel more than lucky to have her compliment, encouragement and recommendation as “smart", even though I understand clearly that a more accurate word for me is “diligent". Any little accomplishment I made through my graduate career was from the thousands of trials even during the mid-nights and weekends. Thank you, all the professors!

Also, I would like to thank the scientists at UC Berkeley and Lawrence Berkeley National Lab: Dr. Chenhui Zhu, Dr. Hiroshi Nishimura, Dr. Simon Leemann, Dr. Cheng Wang and Dr.

Guillaume Freychet for helping me through all the projects and collaborations. Special thanks to Dr. Charles Nathan Melton, Dr. Dinesh Kumar, Dr. Ronald Pandolfi and Dr. Kochise Bennett who pointed out the issues in my dissertation and helped me with the corrections. They are very good mentors who are always willing to help me. I would also like to thank my graduate colleagues: Jerry Li, Brad Ganoe, Dr. Peter Bai, Yihan Xiao and Katherine Evans, and my postdoc colleagues: Dr. Tim Stauch, Dr. Jingyu Huang, Dr. Tao Li and Dr. Tao Jiang for their helps on my research projects.

Moreover, I would like to thank the Berkeley campus. They don't only treat us as graduate students. They support and protect us to guarantee us the bright future. Especially, I would like to thank the support from the International Office and the help from Graduate Division.

Finally but most importantly, I would like to thank my parents. The most lucky thing in my life is to be their child. They give me lots of support during the tough time. Their unconditional love helps me overcome the difficulties, and also makes me be kind to people and helps me be optimistic to life.

Part I

Introduction

Chapter 1

Introduction

In this chapter, the motivations of this dissertation project is discussed. In addition, an overview of the subsequent chapters is provided.

1.1 Motivation and Opportunity

The materials chemistry discovery cycle contains many different components, including synthesis, characterization and data interpretation. In the past few decades, automatic synthesis pipelines have been established for many chemistry and materials systems [1–3]. For characterization, many advanced techniques, such as X-ray scattering [4–6] and NMR crystallography [7–9], have enabled the structure identification of various chemical, biological and materials systems, including polymers [10–12], inorganic materials [13–15] and proteins [16, 17]. These techniques have been developed and improved substantially in the past few decades, which brings high-throughput experimental discovery into reach. Meanwhile, these breakthroughs produce millions of characterization data. However, the process of understanding structural features from these data is labor intensive. It requires many man hours by highly specialized and trained scientific staff to interpret the data and identify the structure. In addition, instrument instability introduces systematic error during the characterization process, which leads to further complication. Therefore, from the experimental side, the next generation of materials chemistry research requires a novel approach to address these problems.

From the computational side, high performance simulation methods have been developed to understand the structures of underlying materials chemistry systems. There are several limitations which obstruct the usage of these methods in a high-throughput fashion. Many first-principle simulations are computationally intensive. For example, predicting chemical shifts for NMR crystallography with DFT calculations takes up to $10^2 - 10^3$ CPU hours or more for a typical chemical system containing more than 100 atoms [18, 19]. Even in the case that the simulation process is not the bottleneck, it is not trivial to map the characterization result to underlying structures. For example, the reverse Monte Carlo method for X-ray scattering data fitting requires a long time to converge even with GPU acceleration [20].

To accelerate materials chemistry research, a different and flexible approach is needed to address these challenges from both experimental and computational sides to enable high-throughput discovery. Recently, machine learning, a branch of artificial intelligence, has demonstrated the capability to tackle many challenging chemistry and materials problems, including machine-learning-assisted materials discovery [21], drug design [22] and crystal structure representations [23]. Herein, I propose a novel and unique approach: integrate machine learning methods into characterization techniques to categorize and manage experimental data, identify the structures, understand the chemistry-nanostructure relationship, approximate state-of-art computational prediction results, and optimize characterization facility parameters.

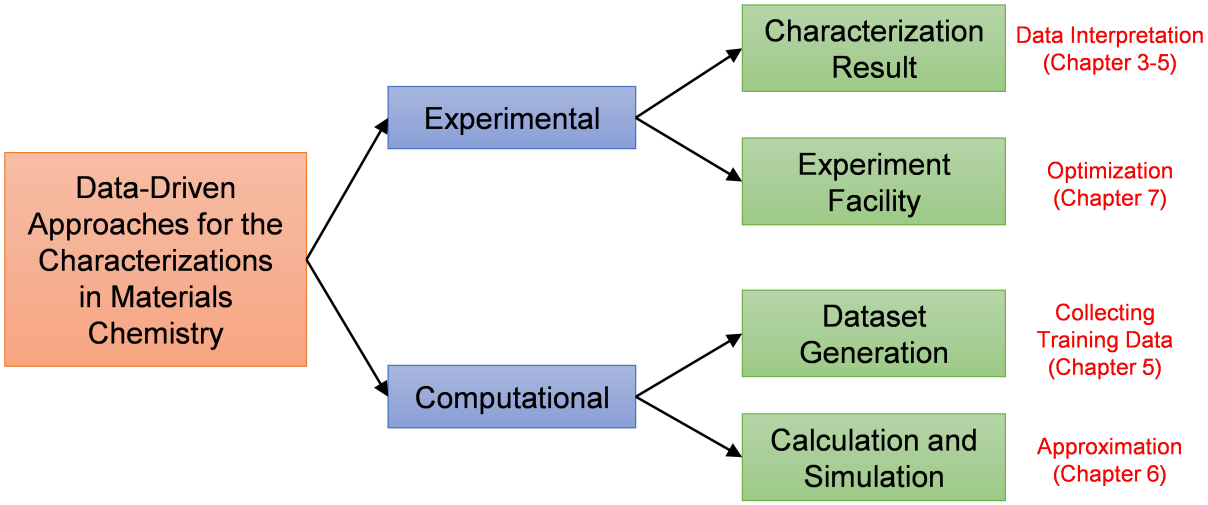


Figure 1.1: Combine data driven approach with characterization techniques in materials chemistry discovery.

Figure 1.1 shows how the data driven approaches interact with existing experimental and computational methods. In detail:

1. Machine learning approaches can be applied to assist, approximate or expedite some of the state-of-art simulation methods [24–26]. In comparison to first-principle computational methods, machine learning models may require less computational resources. For example, the ^{1}H-NMR chemical shift predictions of 500 molecular crystals can be accomplished less than a GPU hour using the MR-3D-DenseNet developed in chapter 6.

2. Machine learning methods have demonstrated the capability and flexibility to build end-to-end modeling for a variety of tasks together with "big data". I construct databases containing simulation and experimental characterization data, together with their underlying structural information. Using these databases, the machine learning models can be easily trained to classify the experimental data to understand the underlying structures.

3. To improve the characterization instrument stability, I demonstrate an proof-of-concept example to predict and stabilize the beam size at Advanced Light Source (ALS).

1.2 Overview of the Subsequent Chapters and Contributions

This dissertation is mainly focused on the efforts and progress made to combine experimental or computational methods with data driven approaches for characterization tasks toward accelerating materials chemistry discovery process. The first part of this dissertation introduces the motivation and background. In the second and third parts of this dissertation, I report several developments toward these goals.

Combine Experimental Method with Data Driven Approach for X-ray Scattering Characterizations

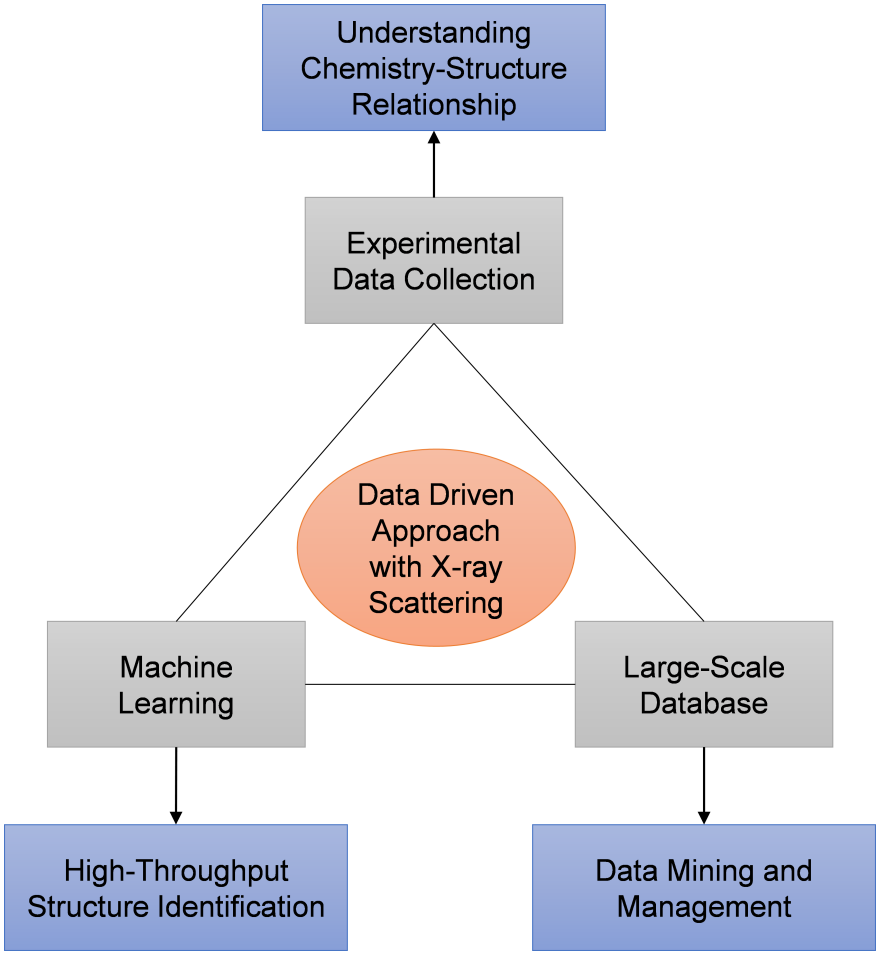


Figure 1.2: A framework by merging X-ray scattering experiments with data mining.

In the second part of this dissertation, I demonstrate a framework by combining experimental method with data driven approaches, with an emphasis on the X-ray scattering technique (shown in Figure 1.2). In detail:

1. Experimental data collection and analysis is an important and essential part in this framework. As an example, in chapter 3, I present an experimental X-ray scattering study of ionic liquid containing polymer based supramolecule. First, using a large number of static X-ray scattering experiments, I investigate the chemistry-structure relationship for supramolecules with respect to many chemical design factors, such as alkyl chain length, counter ions, stoichiometry and polymer backbone. Second, I discover that the configuration of polystyrene chain is more related to the small molecule chemistry whereas the configuration of supramolecule

chain is more related to the stoichiometry. This interesting phenomenon was not observed in other polymer based supramolecules. Third, using *in situ* X-ray scattering experiments, I observe that this system exhibits unusual thermal stability in comparison to other polymer based supramolecules. Some of the data will be reused as examples in subsequent chapters.

2. Conventional X-ray scattering data analysis is time-consuming, which is one of the bottlenecks in high-throughput experiments. In chapter 4, I propose a data driven framework by combining experimental characterization data, machine learning methods and domain knowledge in X-ray scattering. First, a large-scale X-ray scattering experiment database is built with feature based labels from domain experts. Second, I propose a hierarchical approach to analyze the data and implement different machine learning models toward automatic analysis. Third, I demonstrate the application of this framework by applying it to different experimental systems. The training, evaluation and the robustness of the models rely on high-quality and diverse dataset. Finally, I point out the importance of the data with suggestive future improvements. In the future, we plan to integrate this framework into a larger materials chemistry discovery framework by combining it with high-throughput synthetic platform.

3. When a large-scale labeled experimental dataset is unavailable due to high complexity of the system, a machine learning model can be trained using simulation datasets. In chapter 5, I illustrate a machine learning approach to identify the nanostructures of certain materials chemistry systems. First, I construct a GISAXS database containing millions of simulated GISAXS results and the corresponding structural information. Second, I train the machine learning models and evaluate the robustness of the model under different simulated instrumental noise values. Third, I perform an analysis to obtain insights from the prediction results with respect to different physical parameters, such as lattice structure, orientation, and the number of repeating units along different directions.

Data Driven Methods for other Challenges in Characterization

In the third part of this dissertation, I describe how the data driven approaches can be applied to a wide variety of techniques:

1. Another complementary approach for structural characterization is NMR crystallography. In chapter 6, I use a deep learning model to predict the chemical shifts for molecular crystals. First, this approach provides good prediction results with fast speed. The deep learning method is significantly faster than DFT calculation. Also, the prediction accuracy is higher than the kernel ridge regression (KRR) method reported in previous literature. Second, I demonstrate that the prediction performance can be improved by representing the chemical environment with different bounding box sizes. To further exploit the benefit of multi-resolution approach, I also propose a modification of 3D-DenseNet architecture. Third, I report a study of the chemical environment representation using different Gaussian, Slater, and other density functions. Finally, I analyze the correlation of the neural network output with

several pre-designed features in the literature to show the capability of structural information extraction from convolutional neural networks.

2. Instrumental stability is a necessary for high quality data acquisition. In chapter 7, I use machine learning method to optimize the experimental characterization setup, using the accelerator at ALS as an example. The first attempts of the experiments during applied physics shift indicate that the vertical beam size variance is reduced using neural network-based method in comparison to the current baseline method.

To make the dissertation more coherent and concise, several other projects, including the synthesis and characterization of MOF-random copolymer monolayer and a new neural network architecture for periodic data, are not presented in this dissertation. At the end of each chapter, I also present the limitations and future directions, which describes the research opportunities in the next steps.

Chapter 2

Background: Characterization and Machine Learning Methods for Chemistry and Materials Science

In this chapter, a short review is provided on the basic concepts of X-ray scattering and NMR crystallography, as well as their applications in materials chemistry. In addition, I will provide a brief introduction to the applications of machine learning in chemistry and materials discovery.

2.1 Fundamentals of X-ray Scattering

Basic Concepts of X-ray Scattering

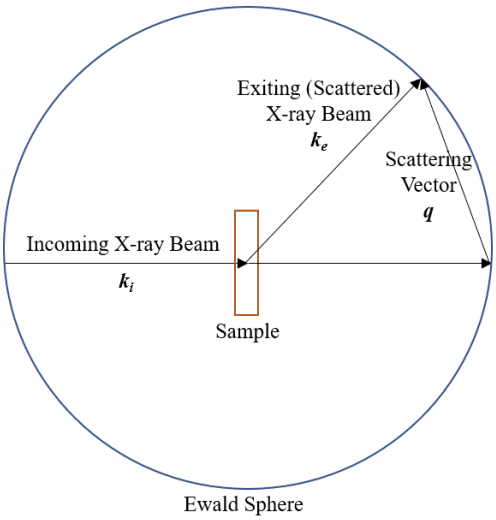


Figure 2.1: The geometry of incoming beam wave vector, exiting beam wave vector and scattering vector.

In this section, I provide a mini-review on the fundamental concepts of X-ray scattering. The scattering vector $\boldsymbol{q}$ is defined as the difference of the exiting beam wave vector $\boldsymbol{k_e}$ and the incoming beam wave vector $\boldsymbol{k_i}$ in reciprocal space (Figure 2.1).

$$\boldsymbol{q} = \boldsymbol{k_e} - \boldsymbol{k}_i \tag{2.1}$$

which is the momentum transfer with magnitude $\frac{2\pi}{\lambda}$. Scattering intensity $I(\boldsymbol{q})$ is

$$I(\boldsymbol{q}) = NP(\boldsymbol{q})S(\boldsymbol{q}) \tag{2.2}$$

where N is the number of repeating units, $P(\boldsymbol{q})$ is the conjugate square of the form factor $F(\boldsymbol{q})$, and $S(\boldsymbol{q})$ is the structure factor. The form factor $F(\boldsymbol{q})$ is the scattering factor of a single unit, such as a single nanoparticle or an atom. Mathematically, the form factor $F(\boldsymbol{q})$ is the Fourier transform of electronic density distribution $\rho(\boldsymbol{r})$ in physical space.

$$F(\boldsymbol{q}) = \int \rho(\boldsymbol{r}) \exp(i\boldsymbol{q}\boldsymbol{r}) \mathrm{d}\boldsymbol{r} \tag{2.3}$$

The structure factor $S(\boldsymbol{q})$ contains the geometric information between the units, which is defined as

$$S(\boldsymbol{q}) = \frac{1}{N}\left|\sum_{i=1}^{N} \exp(-i\boldsymbol{q}\boldsymbol{R_i})\right|^2 \tag{2.4}$$

where $\boldsymbol{R_i}$ is the position of unit i.

Experimentally, the scattering vector $\boldsymbol{q}$ can be calculated from the angle θ between the incoming X-ray beam and the exiting X-ray beam

$$|\boldsymbol{q}| = \frac{2\pi}{\lambda} \sin\theta \tag{2.5}$$

In real experiments, the intensity signal is usually collected on a 2D planar detector. Sometimes, geometric corrections are required to calibrate the $\boldsymbol{q}$ values. Based on the scattering angle and the geometry of the scattering experiment, the experimental techniques can be divided into different categories shown in Table 2.1.

Criterion	Classes
Scattering Angle	Small Angle X-ray Scattering(SAXS) Wide Angle X-ray Scattering(WAXS)
Geometry of the X-ray Scattering Experiment	Transmission X-ray Scattering Grazing incidence X-ray Scattering

Table 2.1: Different X-ray scattering characterization techniques

By collecting the scattering intensity at different $\boldsymbol{q}$ ranges, the X-ray scattering can be utilized to identify the structures in various length scales. WAXS is usually applied to characterize the structures with small repeating units with Angstrom to nanometer scale. In contrast, SAXS can be applied to characterize the structure with relatively larger repeating units, such as the structure of self-assembled block copolymers, which usually contains 10 nm to 100 nm size features.

Grazing Incidence X-ray Scattering

Scattering techniques can also be categorized by the geometry of the X-ray scattering experiment, for example, transmission X-ray scattering or grazing incidence X-ray scattering. Transmission X-ray scattering is usually applied to bulk samples, whereas grazing incidence X-ray scattering is applied to thin film samples. In this section, I provide a short introduction to the grazing incidence small angle X-ray scattering (GISAXS) technique.

The X-ray beam interacts with the surface of the sample with a shallow incidence angle α_i. There are two exiting beam angles: in-plane (sample plane on xy direction) angle ψ and out-of-plane (z direction) angle α_f. In grazing incidence X-ray scattering, the scattering vector is [27]

$$\boldsymbol{q} = \boldsymbol{k_f} - \boldsymbol{k_i} = \frac{2\pi}{\lambda} \begin{pmatrix} \cos\alpha_f \cos\psi - \cos\alpha_i \\ \cos\alpha_f \sin\psi \\ \sin\alpha_i + \sin\alpha_f \end{pmatrix}$$

Grazing incidence X-ray scattering is a widely used technique for the surface characterization for organic and inorganic systems, including polymers, nanocrystals and metal-organic frameworks (MOFs) with differnt ranges of length scales [28]. The scattering result contains the information of the size and pattern of materials, which can be inferred from physics models, such as the Distorted-wave Born Approximation (DWBA) [29, 30]. The scattering intensity can be formulated as a sum of four different components [31]

$$I(q_{//}, q_z) = |F(q_{//}, q_z) + R(\alpha_i)F(q_{//}, p_z) + R(\alpha_f)F(q_{//}, -p_z) + R(\alpha_i)R(\alpha_f)F(q_{//}, -q_z)|^2 \quad (2.6)$$

where $p_z = (\boldsymbol{k_i} + \boldsymbol{k_e})_z$, $q_{//} = (q_x^2 + q_y^2)^{1/2}$. $R(\alpha_i)$ and $R(\alpha_f)$ are the Fresnel reflection coefficients of the substrate.

As discussed in the previous section, the scattering data are usually collected on a planar 2D detector, which is a sampling of the Ewald sphere. For grazing incidence small angle X-ray scattering (GISAXS), q_x is much smaller than q_y and q_z and the data correction is not necessary. However, for the grazing incidence wide angle X-ray scattering (GIWAXS), the data processing is necessary due to the curvature of the Ewald sphere [32].

2.2 Applications of X-ray Scattering Technique in Materials Chemistry

X-ray scattering techniques have been applied to the structure characterizations of many different chemistry and material systems, such as polymers and composite materials.

Polymers

Polymers are a type of functional materials with many properties driven by its chemistry and morphology. X-ray scattering is a routine technique for the structural characterization of polymers in bulk and in thin films. In bulk systems, small size features, such as the crystallization, can be characterized by WAXS. The features in a relatively large scale, such as the morphology of block copolymers, are usually characterized by SAXS. For example, Nogales et al. characterized the morphological transition of isotactic polypropylene during shear-induced crystallization process [33] using a combination of SAXS and WAXS methods. For the polymer thin films, the nanostructures are usually characterized by GISAXS and GIWAXS. Liu et al. investigated the morphology control and the aggregations of polymer based organic solar cells using GISAXS and GIWAXS, respectively [34].

Specially, block copolymers can self-assemble into different morphologies in bulk and in thin film (the detailed physics explanation is available in chapter 3.2). The morphology of block copolymers can be identified using X-ray scattering techniques. Park et al. characterized the ordered and ultra-dense PS-PMMA block copolymer arrays on the faceted surfaces of sapphire wafers using

GISAXS [35]. When one of the block is crystalline, WAXS can be applied to characterize to the crystalline structure within the micro domains. Loo et al. studied the crystallization modes in different types of block copolymer micro domains using WAXS [36]. Rancatore et al. characterized the alignment of organic semiconductor molecules in block copolymer thin films using GIWAXS with the periodicity around $1.2 - 1.4$ Å [37].

Other than the static X-ray scattering technique, *in situ* X-ray scattering can be applied to investigate the dynamics of the polymer system and its response to different environments. Bai et al. characterized the irreversible order-order transition in a supramolecular system using *in situ* SAXS [38]. Paik et al. characterized the reversible morphology control of block copolymer thin films under the solvent vapor using *in situ* GISAXS [39]. *In situ* X-ray scattering can also be used to monitor the chemical reactions in the polymer systems. Agzenai et al. characterized the polymerization of diallyldimethylammonium chloride using *in situ* SAXS [40].

Inorganic and Composite Materials

X-ray scattering techniques can also be applied to the characterization of inorganic materials. Polte et al. studied the gold nanoparticles nucleation and growth process using *in situ* SAXS [41]. Other than the single component nanocrystals, X-ray scattering has also been applied to multi-component nanocrystal systems. Kwon et al. studied the growth mechanism of gold nanoparticles on CoPt nanocrystal seeds using SAXS. SAXS can also be applied to core-shell structure identification in multi-component nanocrystal systems [42]. The size of the core and shell can be calculate by fitting the form factor intensity given the electronic charge distribution contrast of core, shell and the environment. Krycka et al. characterized the core-shell structure of $Fe_3O_4|\gamma$-Mn_2O_3 using SAXS by fitting the form factor using the core-shell model [43].

Beside the polymers or nanoparticles alone, polymer-nanoparticle nanocomposites combine the quantum properties of inorganic nanoparticles and the synthetic versatility of polymers. X-ray scattering techniques have been successfully applied to characterize the composite materials. Lin et al. reported the self-assembly of CdSe/polystyrene-block-poly(2-vinylpyridine) mixtures [44]. This self-assembly process was studied in both bulk and thin film, which were characterized using SAXS and GISAXS, respectively. Ye et al. characterized the binary superlattices thin film of polystyrene coated nanoparticles using interfacial assembly [45]. The particle size and the polymer length can be tailored to generate different 2D or 3D lattices, such as Body Centred Cubic (BCC) or Face Centred Cubic (FCC). The crystal lattice and the orientation can be deciphered from the GISAXS experiments. Other than the synthetic polymers, the natural biomolecules can also be utilized as the organic component in the nanocomposite materials. For example, Macfarlane et al. utilized SAXS to characterize the structure of gold nanoparticle-DNA composite with controlled nanoparticle size and/or DNA length [46].

2.3 NMR Crystallography in Materials Chemistry

X-ray scattering or diffraction techniques are not sensitive to hydrogen atoms. A complementary approach is nuclear magnetic resonance (NMR), which can characterize the chemical environment of hydrogen atoms under natural abundance. NMR crystallography also has other advantages. For example, it does not require long range order of the sample. In this section, I will provide a brief discussion on the chemical structure identification using NMR crystallography.

NMR crystallography is a characterization technique to determine the structure of solid-state chemical compounds or materials using NMR spectroscopy. The chemical shifts of atoms (relative to the external standard) are measured. The chemical shift is determined by the local chemical environment. In NMR crystallography, the crystal structures can be verified and/or determined by comparing the calculation results with the experimental NMR spectra. In detail, the basic (only using chemical shift) NMR crystallography has the following steps:

1. Propose the trial coordinates of the atoms. The proposed coordinates could be generated using the structural model from XRD [47], adapting from the structures in the existing database [48, 49], or using some templates with prior knowledge [50]. Optionally, these proposed coordinates may be further optimized using molecular dynamics (MD) or quantum mechanics (QM) methods.

2. Calculate the chemical shifts of the atoms using the density functional theory (DFT), for example, gauge including projected augmented wave (GIPAW) method [51–53].

3. Compare the calculated results with experimental results to verify or identify the structures.

Besides the chemical shift, both spin-spin coupling (J-coupling) and direct dipolar coupling can also provide additional local structural information. However, these techniques are out of the scope of this dissertation. NMR crystallography has been applied to many chemical, biological and materials systems. Elena et al. identified the unit cell structures and the hydrogen atom positions in powder samples using NMR crystallography [54, 55]; Abraham et al. characterized the water molecules in pharmaceutical molecular crystals [56]; Skotnicki et al. characterized the structure of amorphous valsartan [57].

2.4 Application of Machine Learning in Chemistry and Materials Discovery

In this section, I provide a short review on the applications of machine learning methods in chemistry and materials problems in literature.

Machine learning has been applied in many chemistry subfields. In analytical chemistry, machine learning has been applied to interpret and classify spectroscopy data. Zhou et al. built

the gradient boosting decision tree (GBDT) to identify chemical information using the desorption electrospray ionization (DESI) mass spectrometry of fingerprint and forehead lipid [58]. In organic chemistry, machine learning methods have been applied to reaction performance prediction and synthesis planning. Ahneman et al. predicted the C-N coupling using neural network and random forest [59]. Segler et al. combined Monte Carlo Tree Search (MCTS) and neural networks to design the chemical synthesis route using machine learning [60]. The performance of machine learning models outperforms traditional linear regression significantly. These ideas are the first attempts to automate the chemistry research using machine learning approaches.

Machine learning algorithms are also applied to the materials chemistry discovery. Machine learning methods have proved to be effective on assisting materials synthesis. Raccuglia reported the training of a support vector machine (SVM) model to learn the reaction outcome under different experimental conditions [21]. The prediction accuracy is 89%, which is higher than well-trained synthetic chemists. Moreover, the decision process can be rationalized and visualized by deriving the decision trees. In addition to materials synthesis, machine learning has also been applied to materials property prediction. Ghanshyam Pilania applied kernel ridge regression (KRR) to predict materials properties such as atomization energy, bandgap, and electron affinity of quasi-1-d material motifs using DFT calculation data [61]. Fischer et al. predicted the crystal structure of an alloy using a generalized cumulant expansion probabilistic model [62]. Jong et al. built a machine learning framework to predict elastic moduli of k-nary inorganic polycrystalline compounds [63].

There are relatively few reports on machine learning assisted characterization techniques, especially for experimental data, and only over the last 3-4 years. There are several reports on characterization data processing, such as independent component analysis (ICA) to unmix the signal of nanoparticle clusters [64], simulated XRD data classification [65, 66] and experiment type categorization in scattering data [67]. However, based on our best knowledge, there is no machine learning used for structure identification using dedicated grazing incidence X-ray scattering data. Moreover, for solid-state NMR crystallography, there are only two reports on the chemical shift prediction in crystalline states using fully connected neural networks with symmetric descriptors [68] and KRR [69], respectively.

In chapter 4 and 5, I will illustrate how machine learning methods can be applied to assist the structure identification using X-ray scattering data. These methods enable the high-throughput automatic structure identification in inorganic, organic and composite systems. In chapter 6, I will demonstrate how to apply deep learning and chemistry knowledge to learn the local chemical environment and predict the chemical shift. In chapter 7, I will show an example to illustrate how the machine learning models can be applied to stabilize the experiment setup for the chemistry and materials characterization facility.

Part II

A Data Driven Framework of Merging X-ray Scattering Experiments and Data Mining

Chapter 3

X-ray Scattering Experimental Method: A Study of Polymer based Supramolecules

Experimental data collection is an important and essential part in the data driven discovery framework. In this chapter, using a large number of X-ray scattering experiments, I characterize the nanostructures of supramolecules to understand the chemistry-structure-property relationship with different small molecule chemistry, stoichiometry and polymer backbone structures. All of the data are analyzed using conventional methods. Some of the data will be reused as examples in subsequent chapters to illustrate the automatic data interpretation method.

3.1 Introduction

Chemistry-Structure Relationship in Polymer based Supramolecules

In this chapter, we systematically investigate the self-assembly of ionic liquid containing block copolymer based supramolecular system in bulk using SAXS. The morphology of the supramolecules can be flexibly tuned by the small molecule chemistry (such as alkyl chain length and counter ion type) and the stoichiometry (shown in Figure 3.1). Moreover, this supramolecular system has unusually high thermal stability revealed from *in situ* X-ray scattering.

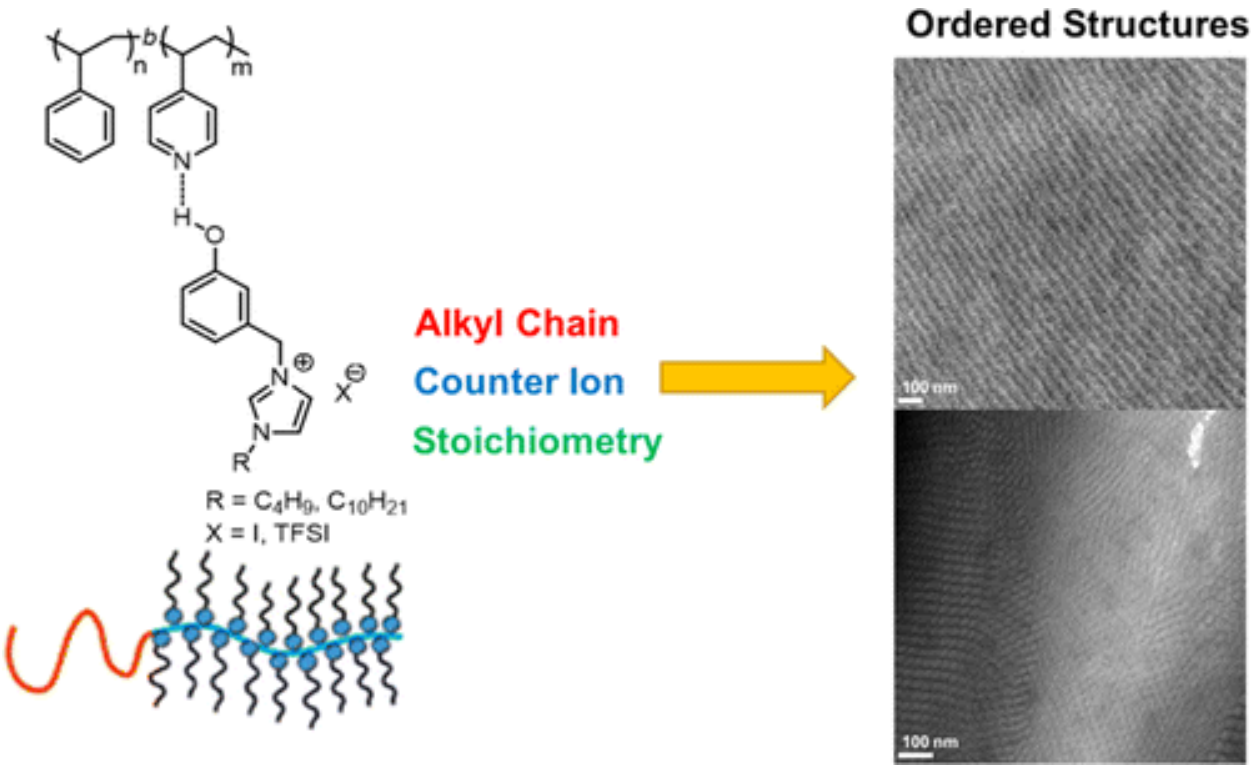


Figure 3.1: A schematic explanation of the chemical and nano-structures of ionic liquid containing block copolymer based supramolecules

Introduction to Ionic Liquid-Polymer Systems

Ionic liquids (ILs) have low melting temperature, low vapor pressure, high thermal stability, high ion mobility and high dielectric constant [71–74]. IL-containing materials have aroused broad interests, such as carbon capture [75–77], conductive membranes [78, 79] and energy conversion [80, 81]. For imidazolium based ILs, it's convenient to modify the counter ions and alkyl chain lengths to tailor their properties [82–84], such as density, viscosity and self-diffusion constant. However, their rheological properties and processibilities are not suitable for many applications. IL-containing polymer and/or polymer composites have been developed, such as polymer-ILs ion

gels [85–89] and poly(ionic liquid)s (PILs). In polymer-ILs ion gels, polymers were added as rheology modifiers to form a chemically or physically cross-linked polymer-ILs network with enhanced mechanical strength. In ion gels, ILs showed high mobility and thus high ion conductivity. Ion gels can be processed to form films for gate dielectrics layer [90–92] and gas separation membranes [93]. However, it remains challenging to control the alignments of IL-containing phase and manipulate their nanostructures to optimize the direction and dimensionalities of ion transport. PILs are a family of polymers where ILs are covalently linked to polymer side chains [94–96]. The alignments or distributions of ILs can be manipulated upon forming block-copolymer (BCP) where one block is PIL. However, the property was limited by the low mobility of PIL segments, especially at temperature below glass transition temperature (T_g) of PILs. Moreover, it is difficult to adjust the volume fraction of ILs post synthesis. This leads to difficulties to modulate overall morphology and ion conductivity.

BCP-based supramolecules, comprised of small molecules non-covalently linked to polymer sidechains, represent a facile way to incorporate functionalities into polymer systems without complex synthesis [97–107]. IL-containing supramolecule may offer a new route to simultaneously optimize the rheological property, morphology and mobility. Small molecules of different chemistry can be readily incorporated to access certain properties such as alkyls [97–99], liquid-crystals [104], and organic semiconductors [106, 107]. Also, the interactions and packings between small molecules can be modified to control the self-assembly behavior and thermal property of the resultant supramolecule. Moreover, the volume fraction of each component, the morphology and the feature size can be adjusted by tuning polymer chain length and/or stoichiometry ratio between the small molecules and polymer repeat unit.

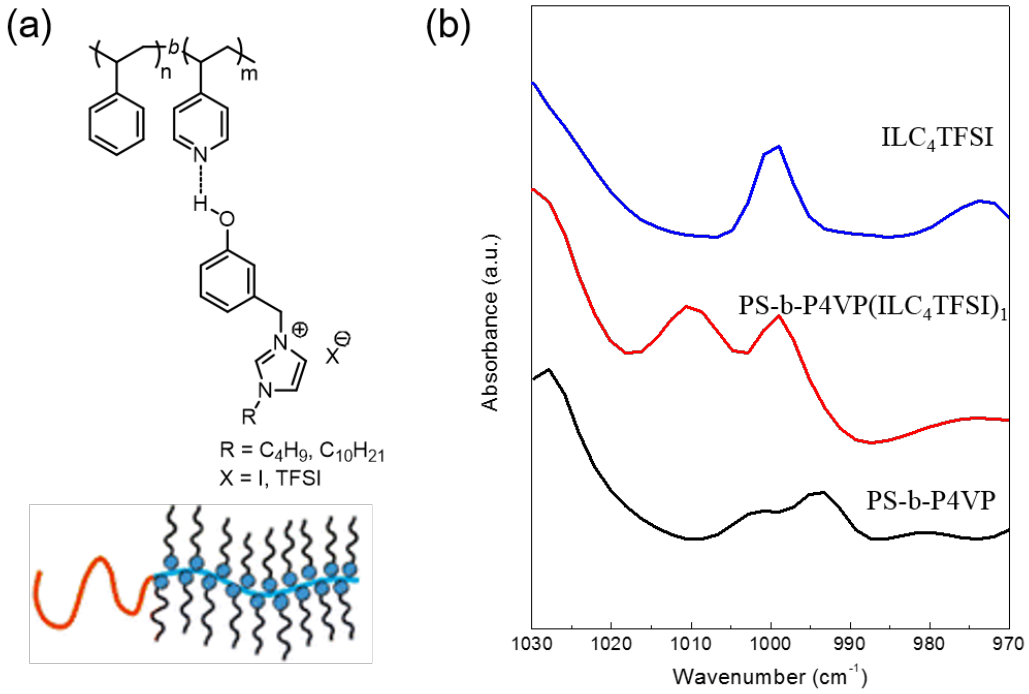


Figure 3.2: Design of PS-*b*-P4VP(ILC$_4$TFSI)$_1$ supramolecules (a) chemical structure of supramolecules (b) FTIR characterizations supramolecule (PS-*b*-P4VP(ILC$_4$TFSI)$_1$).

We designed a family IL-containing BCP-based supramolecules where phenol functionalized ILs are hydrogen bonded to the polystyrene-block-poly(4-vinyl pyridine) (PS-*b*-P4VP) BCP (Figure 3.2). Structure and thermal behaviors of the supramolecules were investigated as a function of alkyl chain lengths, counter ion, and the IL/4VP stoichiometry ratio (r). IL-containing supramolecules provided a diverse and flexible platform to generate nanostructured IL-containing materials. Supramolecules with different morphology and periodicity can be obtained by varying r. The chemistry and composition of ILs can be modified to further tailor the supramolecular morphology. This can be attributed to the changes in the size of comb block and the interactions between IL and each BCP block. Furthermore, the IL-4VP hydrogen bonding in IL-containing supramolecule has higher thermal stability in comparison to other supramolecular systems based on alkyl or liquid-crystal, opening up temperature window to treat and process supramolecules. Thus, the IL-containing supramolecules provide a viable platform to control the spatial arrangement of IL in a processible form and may open up more opportunity to achieve morphological control to improve the properties of IL-based materials.

Theory of Micro-phase Separation in Block Copolymer

The micro-phase separation of block copolymer is controlled by the thermodynamics. The free energy difference ΔG between mixed state and micro-phase separation state is

$$\Delta G = \Delta H - T\Delta S \tag{3.1}$$

where ΔH and ΔS are enthalpy difference and entropy difference, respectively. In the micro-phase separation state, the enthalpic term between component A and component B is propotional to the interfacial energy γ_{AB} and interfacial area S_{AB} in unit volume. The interfacial energy is related to Flory-Huggins parameter χ_{AB}. For example, in lamellae morphology, the interfacial energy is

$$\gamma_{AB} = \frac{kT}{a^2}\sqrt{\frac{\chi_{AB}}{6}} \tag{3.2}$$

where k is Boltzmann constant, T is temperature, and a is Kuhn length. Here, we consider a simple case by assuming the Kuhn length of two blocks are the same: $a_A = a_B = a$. The interfacial area is

$$S_{AB} = \frac{Na^3}{\lambda/2} \tag{3.3}$$

where N is the number of segment of polymer chain and λ is the domain periodicity. By considering the interfacial energy and area, the enthalpic contribution is

$$\Delta H = \frac{kT}{a^2}\sqrt{\frac{\chi_{AB}}{6}}\frac{Na^3}{\lambda/2} - N\chi_{AB}\phi_A\phi_B kT \tag{3.4}$$

where the first term is the multiplication of the interfacial energy and the interfacial area, and the second term is the mixing enthalpy calculated from Flory-Huggins theory. ϕ_A, ϕ_B are the volume fractions of component A and component B, respectively.

The entropic term is related to the polymer chain conformation. Considering the simple Gaussian coil model, the entropy difference in microphase separation in lamellae morphology is

$$\Delta S = \frac{3}{2}kT[\frac{(\lambda/2)^2}{Na^2} - 1] \tag{3.5}$$

The physical parameters, such as χ, N, a and ϕ are determined by the chemical nature of the polymer. The self-assembly behavior is also related to the environment, such as temperature T. By varying these parameters, different morphologies can be achieved, such as lamellae, hexagonal, gyroid and spheres. In block copolymers, the volume fraction is fixed once the polymer is synthesized. Supramolecular strategy provides a unique platform to tune the morphology by varying the chemistry and the stoichiometry of small molecules in a post-synthesis fashion.

3.2 Experiment Results

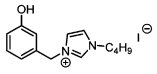
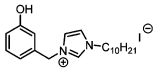
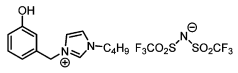
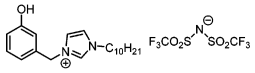
Abbreviation	Name	Chemical Structure
ILC_4I	1-butyl-3-(3-hydroxybenzyl) imidazolium iodide	
$ILC_{10}I$	1-decyl-3-(3-hydroxybenzyl) imidazolium iodide	
ILC_4TFSI	1-butyl-3-(3-hydroxybenzyl) imidazolium bis(trifluoromethylsulfonyl)imide	
$ILC_{10}TFSI$	1-decyl-3-(3-hydroxybenzyl) imidazolium bis(trifluoromethylsulfonyl)imide	

Table 3.1: Small molecules investigated in this chapter

For IL containing supramolecules, phenol functionalized ILs are hydrogen bonded to 4-vinyl pyridine (4VP). The chemical structures of ILs are shown in Table 3.1. ILC_4TFSI was selected for initial study because butyl and TFSI containing IL is the one of the most investigated ILs[108, 109]. The hydrogen bond between 4VP and ILC_4TFSI is shown in Figure . The absorption peak of free 4VP is at 993 cm^{-1}. After hydrogen bonded with ILC_4TFSI, the adsorption peak is shifted to 1010 cm^{-1}. This peak is absent in spectra of both BCP and ILC_4TFSI, which indicates the stretched pyridine ring and the formation of hydrogen bond [103].

Stoichiometry Effect

The morphologies of IL-containing supramolecules were explored as a function of ILC_4TFSI to 4VP ratio, r and are shown in Figure 3.3. The volume fractions are calculated based on the molecular weight, stoichiometry and density of different components: PS, P4VP, phenol and ILs in bulk. The densities of ILs are based on ref [110, 111]. When r is 0.5, the volume fraction of $P4VP(ILC_4TFSI)_{0.5}$ block is 0.42. The SAXS profile showed peaks at q = 0.021 A^{-1}, 0.042 A^{-1}, 0.063 A^{-1} and 0.084 A^{-1}. The SAXS peak ratio is 1:2:3:4 and TEM result (Figure 3.3a) indicates the lamellar morphology with periodicity of 29.9 nm. No internal structure within lamellae microdomain was seen. The volume fraction of $P4VP(ILC_4TFSI)_r$ increases to 0.56 at r = 1, and 0.64 at r = 1.5, respectively. The SAXS profile of $P4VP(ILC_4TFSI)_1$ showed q = 0.018 A^{-1}, 0.031 A^{-1}, 0.035 A^{-1}, 0.047 A^{-1}, 0.052 A^{-1} with a peak position ratio of $1:\sqrt{3}:2:\sqrt{7}:3$. Together with TEM image shown in Figure 3.3b, the morphology is determined to be hexagonally packed PS cylinders embedded in $P4VP(ILC_4TFSI)$ matrix and the periodicity is 35.7 nm. For $P4VP(ILC_4TFSI)_{1.5}$, the SAXS profile showed q = 0.018 A^{-1}, 0.030 A^{-1}, 0.036 A^{-1}, 0.047 A^{-1}, which indicates cylindrical morphology with a periodicity of 35.3 nm. The TEM images of $P4VP(ILC_4TFSI)_{1.5}$ (Figure 3.3c) showed hexagonally packed PS cylinders within P4VP(IL) domain with a periodicity of 40 nm, which are consistent with SAXS results.

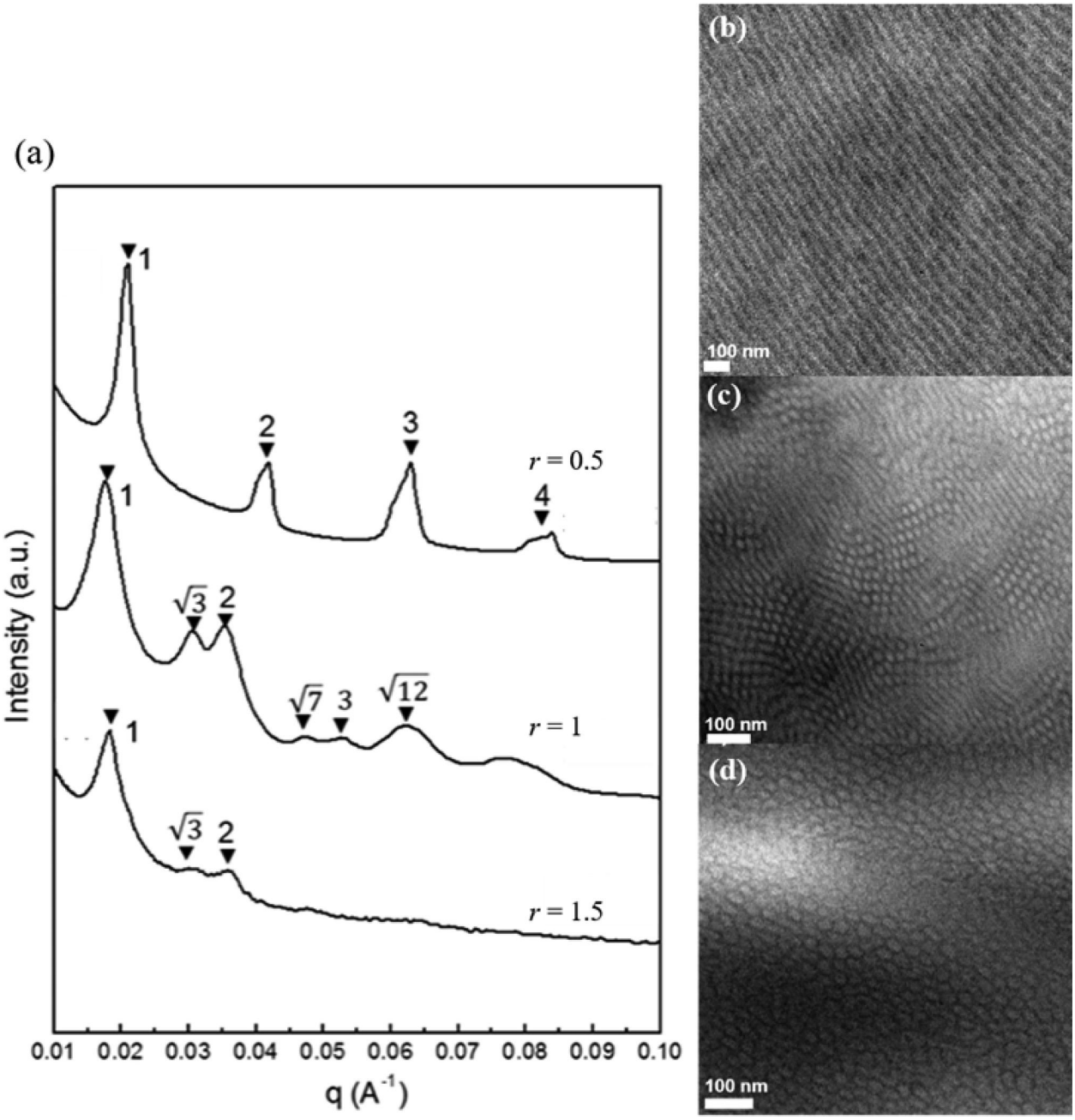


Figure 3.3: Small angle X-ray scattering, TEM images of (a) PS-*b*-P4VP(ILC$_4$TFSI)$_{0.5}$ (q = 0.021 A^{-1}) (b) PS-*b*-P4VP(ILC$_4$TFSI)$_1$ (q = 0.018 A^{-1}) (c) PS-*b*-P4VP(ILC$_4$TFSI)$_{1.5}$ (q = 0.018 A^{-1}). Samples were stained by iodine before TEM test and dark phases were P4VP(ILC$_4$TFSI)$_x$.

Thermal Behavior

The thermal behavior of PS-*b*-P4VP(ILC_4TFSI)$_1$ was characterized using *in situ* FTIR and *in situ* SAXS in Figure 3.4. The *in situ* FTIR spectrum (Figure 3.4a) showed that the intensity of peak corresponding to 4VP/ILC_4TFSI H-bond drops only 13% upon heating from 40 °C to 150 °C. A large fraction of H-bond is still present even at 150 °C. During the cooling process, the hydrogen bond recovers. *In situ* SAXS indicates that the periodicity of PS-*b*-P4VP(ILC_4TFSI)$_1$ only changes 1 nm during the heating (Figure 3.4b) and cooling cycle (Figure 3.4c) and q values of first peak are shown in Figure 3.4d. IL-containing supramolecule is not as temperature sensitive as all of supramolecular systems studied previously.

The thermal behavior property of supramolecules were further evidenced by DSC and rheology experiments shown in Figure 3.5. DSC curve (Figure 3.5a) showed a decreasing T_g as increasing stoichiometry r. The T_g of PS-*b*-P4VP, PS-*b*-P4VP(ILC_4TFSI)$_1$, PS-*b*-P4VP(ILC_4TFSI)$_{1.5}$ are 110 °C, 100 °C, 95 °C, respectively. From rheological analysis, we observe that the PS-*b*-P4VP(ILC_4TFSI)$_1$, PS-*b*-P4VP(ILC_4TFSI)$_{1.5}$ always have solid-like behavior ($G' > G$"). DSC and rheological analysis (Figure 3.5b, 3.5c) also showed more ambiguous glass transition processes of PS block as increasing stoichiometry r.

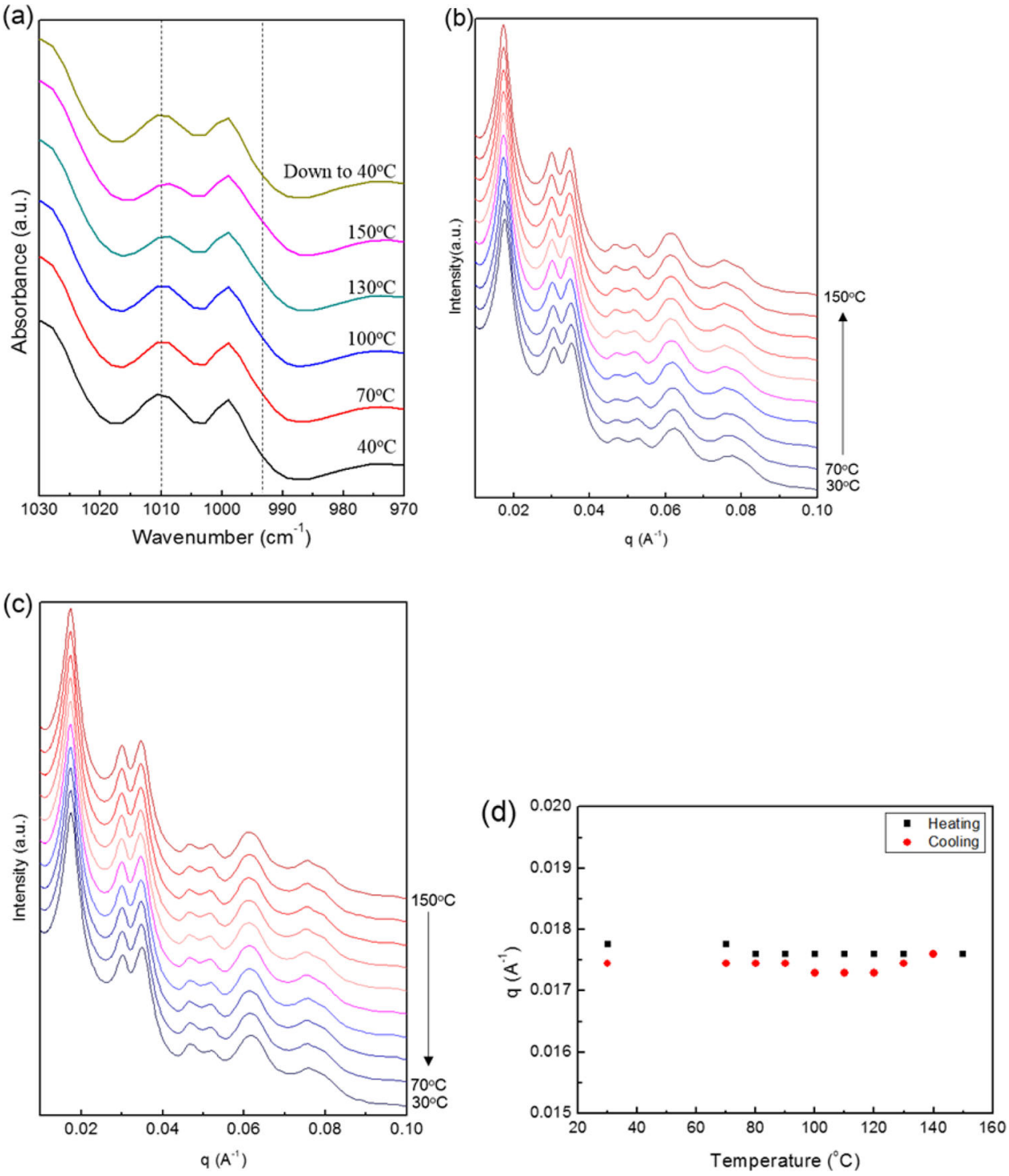


Figure 3.4: Thermal behavior characterizations of PS-*b*-P4VP(ILC_4TFSI)$_1$. (a) *In situ* FTIR of PS-*b*-P4VP(ILC_4TFSI)$_1$ during the heating and cooling process from 40 °C to 150 °C. *In situ* SAXS of PS-*b*-P4VP(ILC_4TFSI)$_1$ during the (b) heating process and (c) cooling process and the (d) q value of the first order peak as a function of temperature.

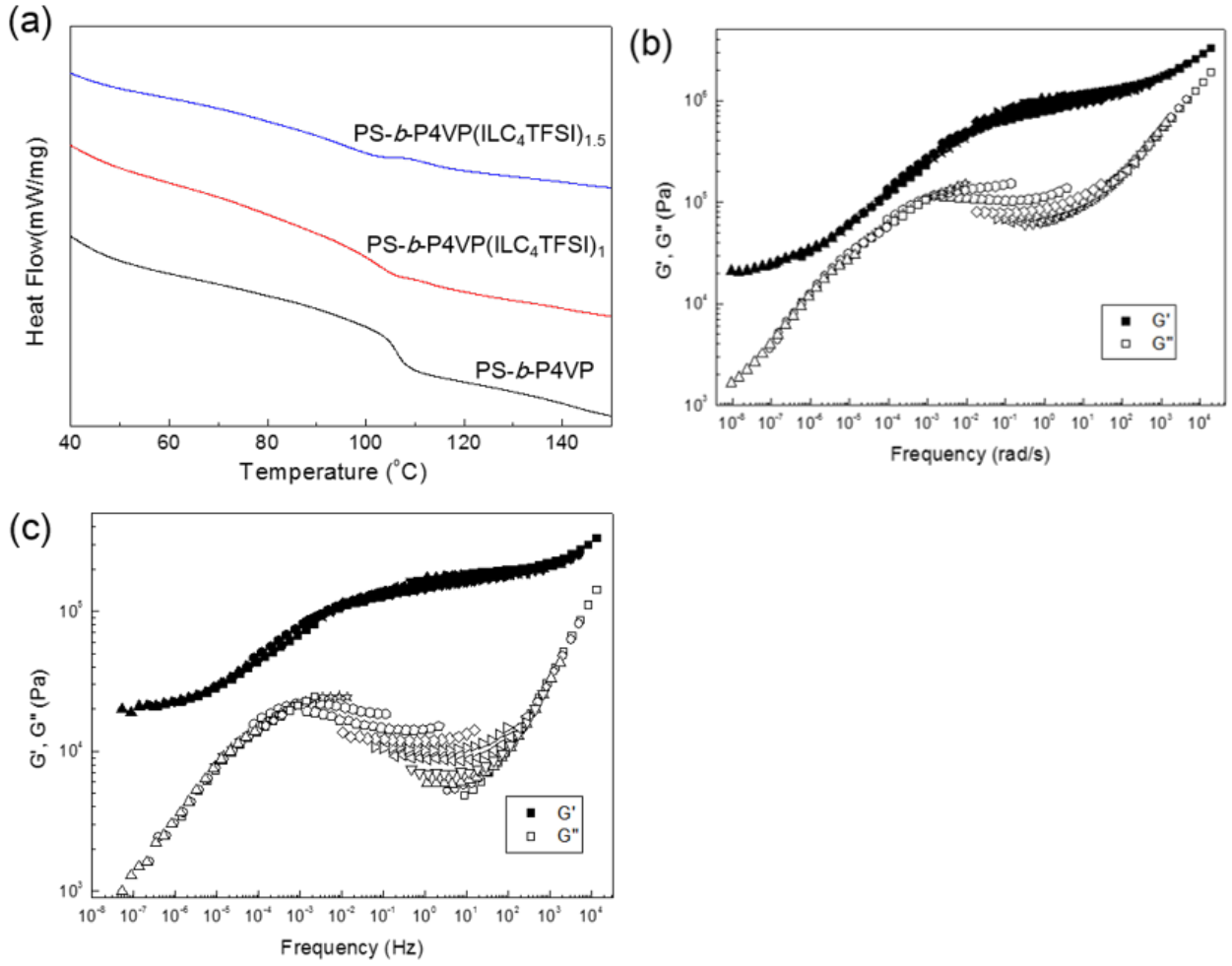


Figure 3.5: (a) DSC curves of PS-*b*-P4VP, PS-*b*-P4VP(ILC_4TFSI)$_1$ and PS-*b*-P4VP(ILC_4TFSI)$_{1.5}$. Time-temperature superposition (tTS) master curve of (b) PS-*b*-P4VP(ILC_4TFSI)$_1$ (c) PS-*b*-P4VP(ILC_4TFSI)$_{1.5}$ using 80 °C as reference temperature.

Effect of Small Molecule Chemistry

The IL chemistry may affect the stability of hydrogen bond between the IL and 4VP. Supramolecules with different ILs listed in Table 3.1 were investigated (Figure 3.6). Two parameters were varied, i.e. the alkyl chain length (butyl (C_4) vs. decyl (C_{10})) and the counter ion (iodide vs. TFSI). The thermal behavior was characterized in Figure 3.6. For all four IL-supramolecules, the H-bond weakens upon heating and recovers upon cooling. The H-bond thermal stability is much better than previously studied supramolecules containing alkyls [98, 99, 103]. For butyl containing ILs, intensity of peak corresponding to hydrogen bonded of P4VP with ILC_4I (Figure 3.6a) or ILC_4TFSI (Figure 3.4a) drop 12% and 11%, respectively upon heating from 40 °C to 150 °C. However, when the alkyl chain longer (C_{10}), the peak intensity drops 24% and 18% respectively for $ILC_{10}I$ (Figure 3.6b) and $ILC_{10}TFSI$ (Figure 3.6c) containing BCP under the same condition. The relative peak

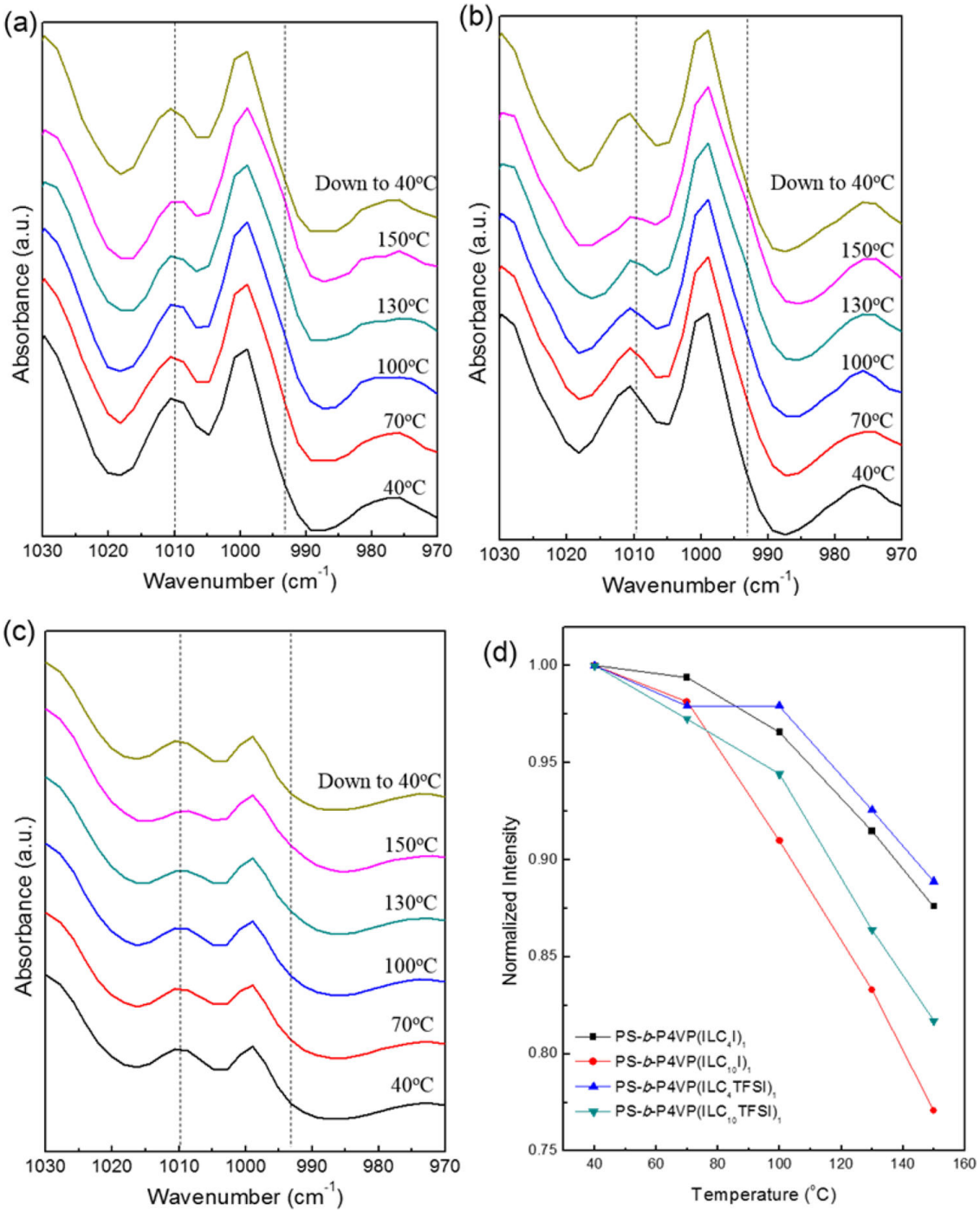


Figure 3.6: *In situ* FTIR of supramolecules with different polymer chain structures (a) PS-*b*-P4VP(ILC_4I)$_1$ (b) PS-*b*-P4VP($ILC_{10}I$)$_1$ (c) PS-*b*-P4VP($ILC_{10}TFSI$)$_1$. Dash lines are at 1010 cm^{-1} and 993 cm^{-1}, which are corresponding to hydrogen bonded P4VP and free P4VP. (d) Ratio of intensity at various temperatures to the intensity at 40 °C (A_{1010}(T)/A_{1010}(40 °C)).

intensities corresponding to 4VP-IL hydrogen bonded are shown as a function of temperature by using the ratio of intensity at various temperature to the intensity at 40 °C (A_{1010}(T)/A_{1010}(40 °C)) in Figure 3.6d. The hydrogen bonds in supramolecules containing ILs with C_{10} alkyl chain, and iodide as counter ion are more temperature sensitive.

The morphologies of supramolecules were investigated as a function of IL chemistry. To exclude the influence of stoichiometry, the stoichiometry r was set as 1. To exclude the effect of thermal history (an example shown in Figure 3.14 in supplementary information), we first raise the temperature above T_g for 10 hours. SAXS and TEM revealed the lamellae and hexagon morphologies of IL-supramolecules (Figure 3.7). Both periodicity and morphology depend on the IL chemistry, i.e. alkyl chain length and counter ion chemistry. The SAXS profile of PS-*b*-P4VP(ILC_4I)$_1$ showed peaks at q = 0.017 A^{-1}, 0.034 A^{-1}, 0.052 A^{-1}, 0.068 A^{-1}, 0.086 A^{-1}, which indicates the lamellar morphology with a periodicity of 36.7 nm. This is confirmed via the TEM image (Figure 3.7a). The SAXS profile of PS-*b*-P4VP($ILC_{10}I$)$_1$ showed peaks at 0.019 A^{-1}, 0.032 A^{-1}, 0.038 A^{-1}, which indicates the hexagonal packed cylindrical morphology with a periodicity of 32.9 nm. The SAXS profile of PS-*b*-P4VP($ILC_{10}TFSI$)$_1$ showed peaks at 0.021 A^{-1}, 0.035 A^{-1}, 0.041 A^{-1}, which indicates the hexagonal packed structure with periodicity of 30.3 nm. The TEM images of P4VP($ILC_{10}I$)$_1$ (Figure 3.7b) and P4VP($ILC_{10}TFSI$)$_1$ (Figure 3.7c) showed hexagonally packed PS cylinders embedded within P4VP(IL) matrix with a periodicity of 38 nm and 35 nm, respectively, The SAXS and TEM results are consistent. Moreover, in-situ SAXS studies were performed to evaluate temperature dependence of IL-supramolecule. Figure 6d shows the q value of the first order peak in SAXS profile as a function of temperature for each supramolecule. From 30 °C to 150 °C, the change in the supramolecular periodicity is 1 nm. This is quite different from that of other supramolecules investigated previously [107, 108].

To understand the structure-IL chemistry relationship, the stoichiometry r was set to 0.5 to obtain lamellar structures for all PS-*b*-P4VP(IL)$_{0.5}$. The SAXS profile of PS-*b*-P4VP(ILC_4I)$_{0.5}$ showed a series of diffraction peaks at q = 0.018 A^{-1}, 0.036 A^{-1}, 0.055 A^{-1}, 0.072 A^{-1}, which indicates the lamellar morphology with a periodicity of 33.8 nm. This agrees with the TEM result (Figure 3.8a). The SAXS profile of PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$ showed peaks at 0.022 A^{-1}, 0.044 A^{-1}, 0.067 A^{-1}, which indicates the lamellar morphology with a periodicity of 28.3 nm. The result is consistent with TEM image (Figure 3.8b). The SAXS profile of PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ showed peaks at 0.023 A^{-1}, 0.045 A^{-1}, 0.068 A^{-1}, which indicates the lamellar morphology with a periodicity of 27.4 nm and is consistent with TEM image (Figure 3.8c).

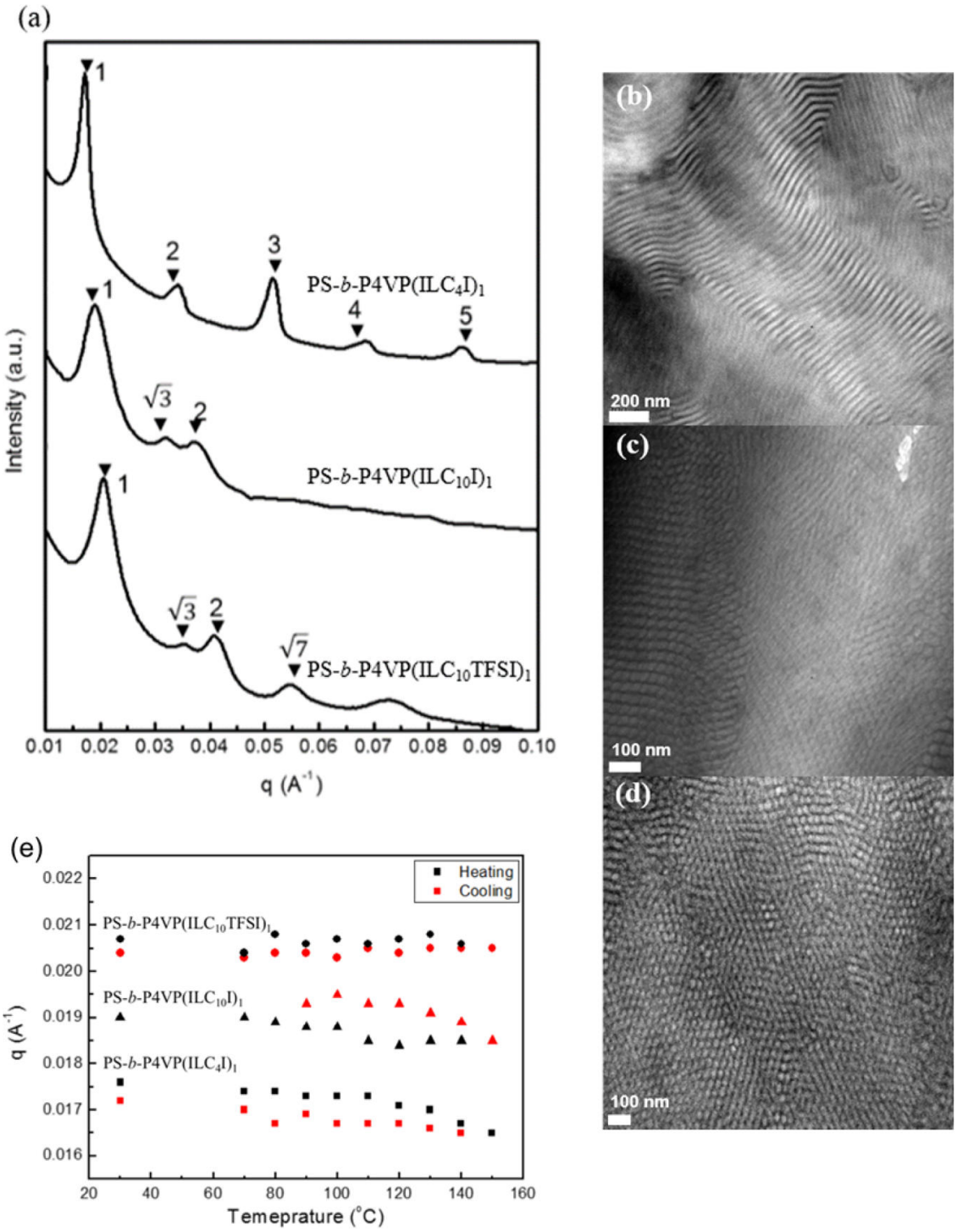


Figure 3.7: Small Angle X-ray Scattering and TEM images of (a) PS-*b*-P4VP(ILC$_4$I)$_1$ (q = 0.017 A^{-1}). (b) PS-*b*-P4VP(ILC$_{10}$I)$_1$ (q = 0.019 A^{-1}). (c) PS-*b*-P4VP(ILC$_{10}$TFSI)$_1$ (q = 0.021 A^{-1}). Samples were stained by iodine before TEM test and dark phases were P4VP(IL)$_1$. (d) q value of different supramolecules under various temperature characterized by *in situ* SAXS.

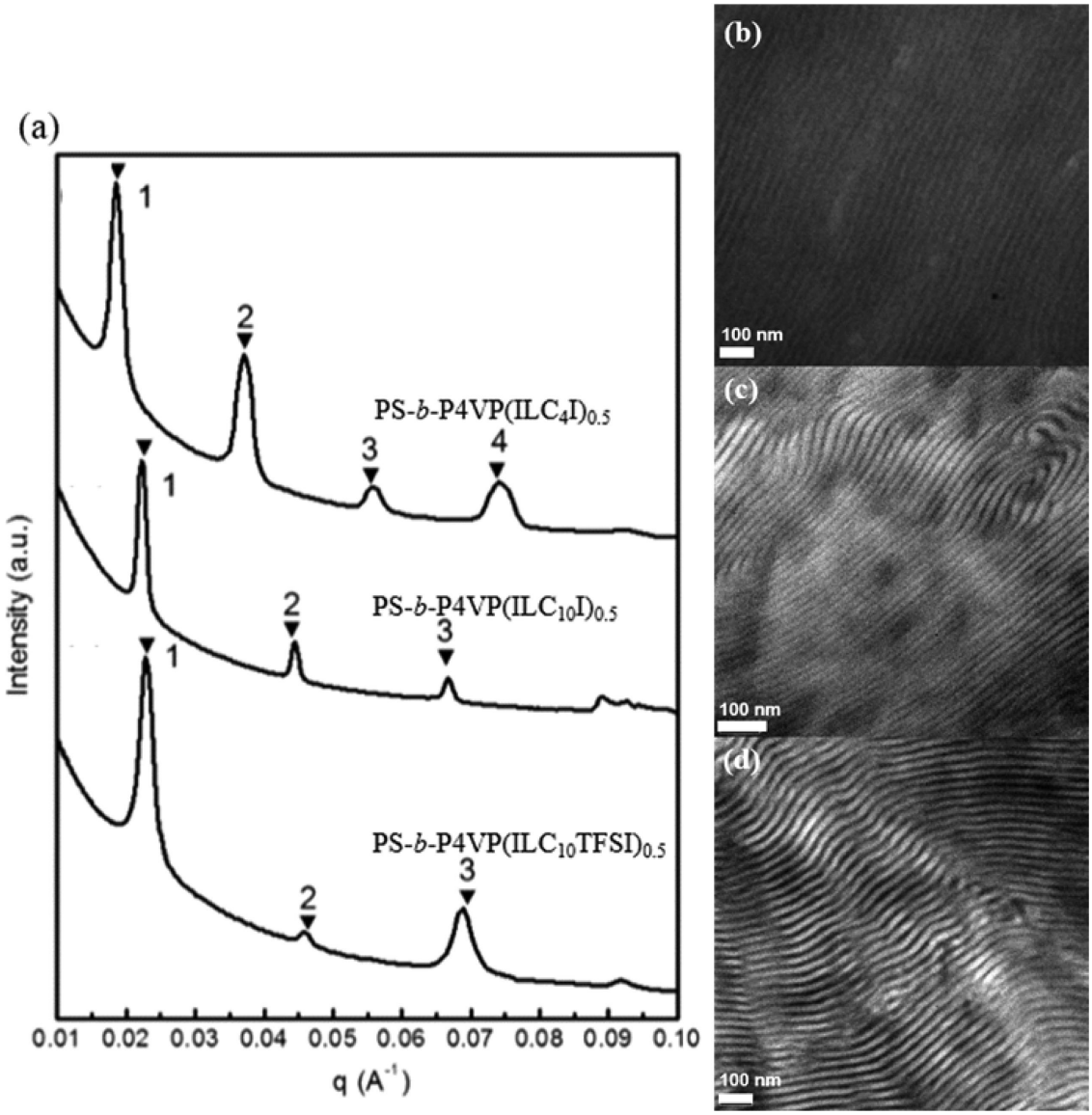


Figure 3.8: Small Angle X-ray Scattering and TEM images of (a) PS-*b*-P4VP(ILC_4I)$_{0.5}$ (q = 0.018 A^{-1}) (b) PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$ (q = 0.022 A^{-1}). (c) PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ (q = 0.023 A^{-1}). Samples were stained by iodine before TEM test and dark phases were P4VP(IL)$_{0.5}$.

Small Molecules	**H-Bond Sensitivity**	**IL/4VP ratio** r	**f**(P4VP-IL)	**Structure**	**Periodicity (nm)**
ILC_4I	0.12	0.5	0.36	Lam[b]	33.8
		1	0.46	Lam	36.7
$ILC_{10}I$	0.24	0.5	0.41	Lam	28.3
		1	0.55	Hex[b]	32.9
ILC_4TFSI	0.11	0.5	0.42	Lam	29.9
		1	0.56	Hex	35.7
$ILC_{10}TFSI$	0.18	0.5	0.46	Lam	27.4
		1	0.61	Hex	30.3

Table 3.2 IL-supramolecule with different small molecules.

(a) Peak intensity decrease from 40 °C to 150 °C relative to the peak intensity at 40 °C.

$$\text{Sensitivity}=\frac{A_{1010}(40\,^{\circ}\text{C}) - A_{1010}(150\,^{\circ}\text{C})}{A_{1010}(40\,^{\circ}\text{C})} \tag{3.6}$$

(b) Lam: Lamellae, Hex: Hexagonally packed PS cylinder.

We summarize the H-bond thermal behavior, stoichiometry r, volume fraction and nanostructure in Table 3.2. H-bond thermal sensitivity is calculated using the ratio of peak intensity decrease from 40 °C to 150 °C normalized by the peak intensity at 40 °C. Volume fractions are calculated based on the molecular weight, stoichiometry and density of different components: PS, P4VP, phenol and ionic liquids in bulk. The periodicities ($2\pi/q$) are calculated based on the first order peak in SAXS.

Effect of Polymer Backbone Structure

To compare with the ordered block copolymer based supramolecules, we designed the homopolymer and random copolymer based supramolecules. The structure of homopolymer and random copolymer based supramolecules and the corresponding *in situ* FTIR are shown in Figure 3.9. At room temperature, the peaks at 1010 cm^{-1} present in all three supramolecules, which proves the formation of hydrogen bonds. When the supramolecules were heated up above 100 °C, the FTIR of RCP based supramolecules showed that the peak at 1010 cm^{-1} was weakened. This indicated that the hydrogen bond in RCP based supramolecules was partially dissociated at high temperature. On the contrary, the peak intensities of HP and BCP based supramolecules only slightly decreased upon temperature even to 170 °C. This indicates that the thermal behavior of hydrogen bonds is correlated with the structure of polymer backbone. During the cooling process, all the peaks at 1010 cm^{-1} were recovered, which indicates that this dissociation process is reversible.

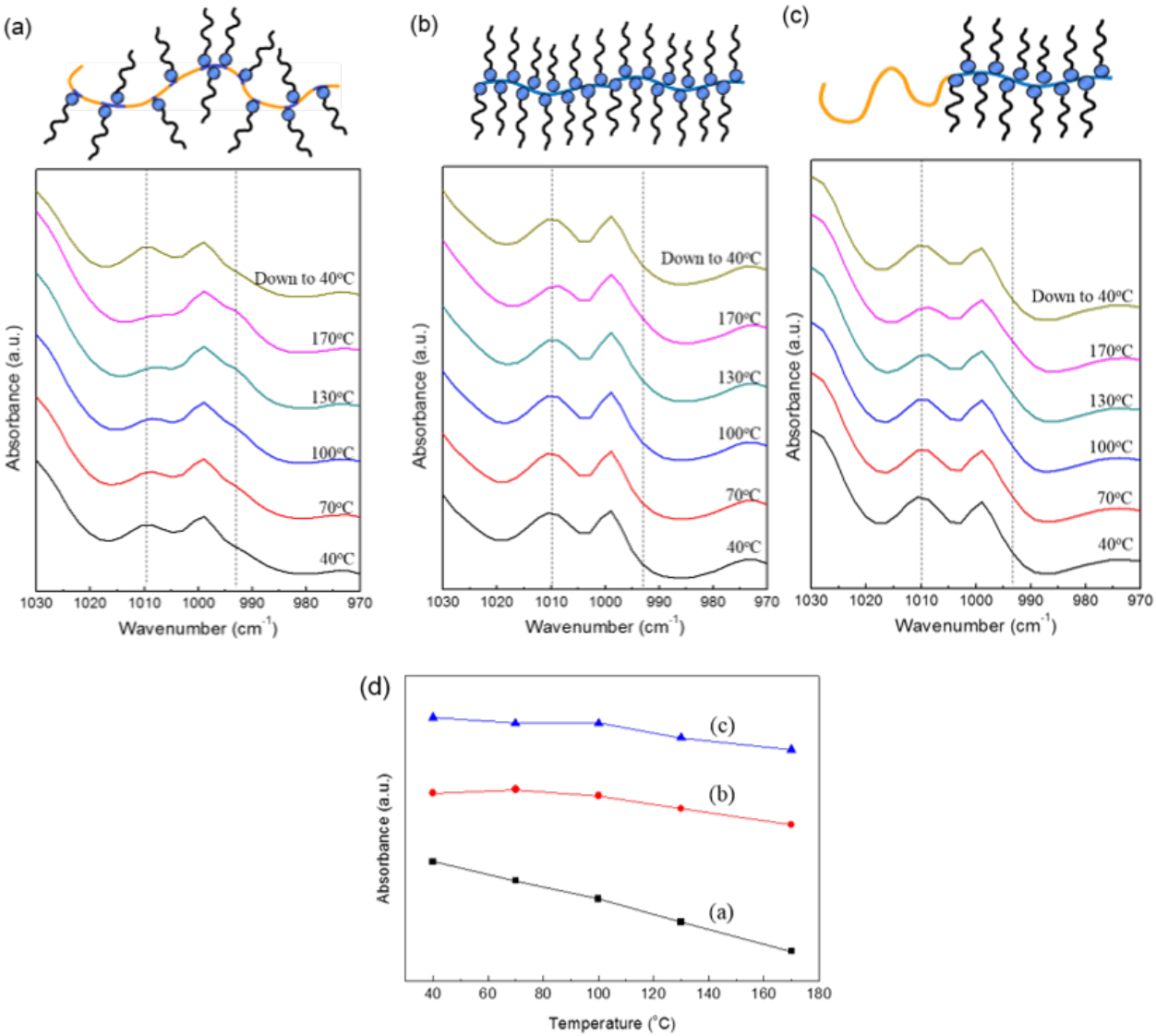


Figure 3.9: *In situ* FTIR of supramolecules with different polymer chain structures (a) PS-*r*-P4VP(ILC$_4$TFSI)$_1$ (b) P4VP(ILC$_4$TFSI)$_1$ (c) PS-*b*-P4VP(ILC$_4$TFSI)$_1$. Dash lines are at 1010 cm^{-1} and 993 cm^{-1}, which are corresponding to hydrogen bonded P4VP and free P4VP (d) The integration of peak at 1010 cm^{-1} as a function of temperature with different polymer chain structures.

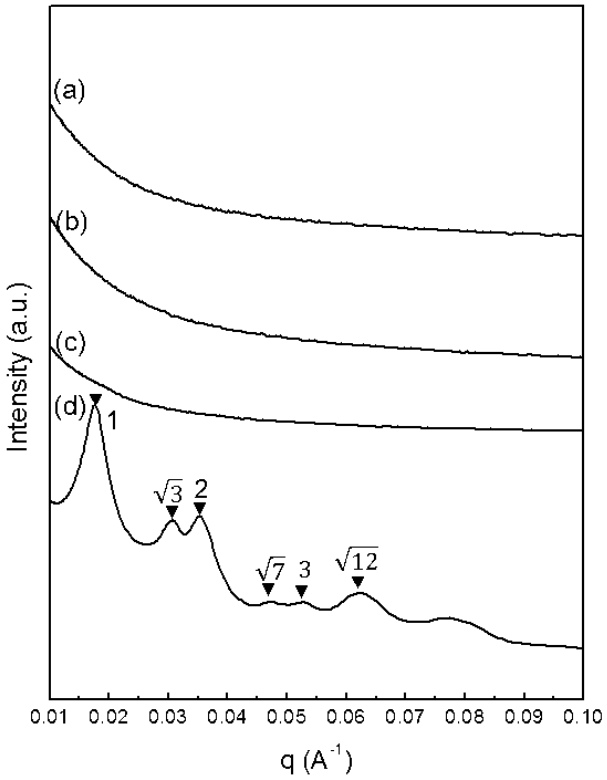


Figure 3.10: Small Angle X-ray Scattering of (a) ILC_4TFSI (b) P4VP(ILC_4TFSI)1 (c) PS-*r*-P4VP(ILC_4TFSI)$_1$ (d) PS-*b*-P4VP(ILC_4TFSI)$_1$ (q = 0.0176 A^{-1}).

We also performed Small Angel X-ray Scattering (SAXS) to characterize the nanostructures of ionic liquid containing supramolecules with different polymer backbone structures shown in Figure . Pure small molecules, HP and RCP based supramolecules did not show any features in the window of tens of nanometers. In BCP based supramolecules, the ionic liquids were incorporated into P4VP domain via hydrogen bond. This verified that the ordered structures were formed by the phase separation of the P4VP-ionic liquid supramolecules block and the PS block.

The structure differences also lead to differed thermal behavior. The thermal behaviors are characterized using DSC and rheometer in supplementary information (Figure 3.15 and Figure 3.16). RCP based supramolecules has a T_g around 30 °C and HP based supramolecule is in liquid phase at room temperature, where BCP based supramolecules has a T_g around 90 °C-100 °C.

3.3 Discussion

For the supramolecule PS-*b*-P4VP(ILC_4TFSI)$_r$, three different IL stoichiometries are investigated, r = 0.5, 1.0, and 1.5, with IL weight fractions of 31%, 47%, and 57%, respectively. Ordered lamellar (r=0.5, $f_{P4VP-IL}$=0.42) and hexagonally packed cylindrical ($f_{P4VP-IL}$=0.56, 0.64) morphologies are obtained (shown in Table 3.2). Moduli of supramolecules are 10^6 - 10^7 Pa at room temperature and stay solid-like in a wide range of temperature (30 °C-150 °C). There are two relaxation regimes below and above T_g. This two-relaxation-mode behavior was also observed in BCP-based ion-gel systems[87, 89]. As the fraction of IL increases, the glass transition process of IL-containing supramolecule becomes more ambiguous and modulus decreases according to both DSC and rheological analysis. In summary, IL-containing BCP-based supramolecules can form ordered structure over a wide range of IL content and the morphology, moduli and IL loading can be readily tuned.

The liquid nature of ILs also distinguishes the structures of the IL-containing supramolecules from that of supramolecules based on crystalline small molecules (CSMs). In all the CSM-containing supramolecules we investigated, such as PS-*b*-P4VP(PDP), P4VP(CSM), they formed comb blocks and packed with a periodicity of about 4 nm which is evident from peak at 0.1 - 0.2 A^{-1} in SAXS.[98, 99, 104, 107] The molecular packing of PDP and other CSMs leads to periodic assemblies of the P4VP(CSM) comb blocks. The liquid-like nature of ILs makes P4VP(IL) chain configuration more bottle-brush like, and thus there is no inter-chain packing (SAXS in high q range were shown in Figure 3.15) in supplementary information. This behavior is similar to that of CSMs-containing supramolecules above the melting temperature of CSM.

Supramolecule	**Etimated Lamellar Domain Size (nm)**			
	Based on SAXS		**Based on Volume Fraction**	
	PS	**P4VP(IL)**	**PS**	**P4VP(IL)**
PS-*b*-P4VP(ILC_4I)$_{0.5}$	22.5	11.3	21.7	12.1
PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$	-	-	16.7	11.6
PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$	-	-	17.0	12.9
PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$	13.7	13.7	14.8	12.6

Table 3.3: Estimated Domain Length of Supramolecules in Lamella Structure

Supramolecular strategy is also compatible with diverse IL chemistry, such as different alkyl chain lengths and counter ions. We investigated the ILs with butyl (C_4H_9) or decyl ($C_{10}H_{21}$) as the alkyl chain and iodide or TFSI as the counter ion. The result clearly shows that the IL chemistry affects the final supramolecule structures. The different IL chemistries alter their molecular weights, densities and interactions with PS block. An IL with shorter alkyl chain and iodide as counter ion has a lower molecular weight but higher density. The volume of small molecule in each assembly is in the order of $ILC_4I < ILC_4TFSI \sim ILC_{10}I < ILC_{10}TFSI$. For a given small molecule stoichiometry r, upon increasing the volume of IL molecule, the volume fraction of P4VP(IL) block increases. However, the periodicity of supramolecule decreases in the order of PS-*b*-P4VP(ILC_4I)$_{0.5}$ < PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$ < PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$ < PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$. Thus, the periodicity of PS domain length greatly decreases whereas P4VP does not. To quantitatively correlate the chemistry-structure relationship, we estimate the domain size of two blocks using SAXS data and volume fractions in Table 3.3. First, we can estimate the domain sizes of each block in the supramolecules based on the coincidence of the form factor minimum and structure factor maximum. For instance, the third order peak of PS-*b*-P4VP(ILC_4I)$_{0.5}$ is weak because it coincides the form factor minimum when the volume ratio of the two blocks is close to 1:2. The volume fraction of P4VP(ILC_4I)$_{0.5}$ in PS-*b*-P4VP(ILC_4I)$_{0.5}$ is 0.36 (Table 3.2). Thus, the PS domain size is about 2/3 of the periodicity whereas the P4VP(ILC_4I)$_{0.5}$ domain size is about 1/3 of the periodicity. Similarly, the second order peak of PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ is weak when the volume ratio of two blocks is close to 1:1. In PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$, the domain sizes of two blocks are almost the same. Based on these estimations, the PS domain size in PS-*b*-P4VP(ILC_4I)$_{0.5}$ and PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ are 22.5 nm and 13.7 nm whereas the P4VP(IL)$_{0.5}$ domain sizes are 11.3 nm and 13.7 nm. The volumes of IL molecules are strongly correlated to the PS domain sizes but not to the P4VP(IL)$_{0.5}$ domain sizes.

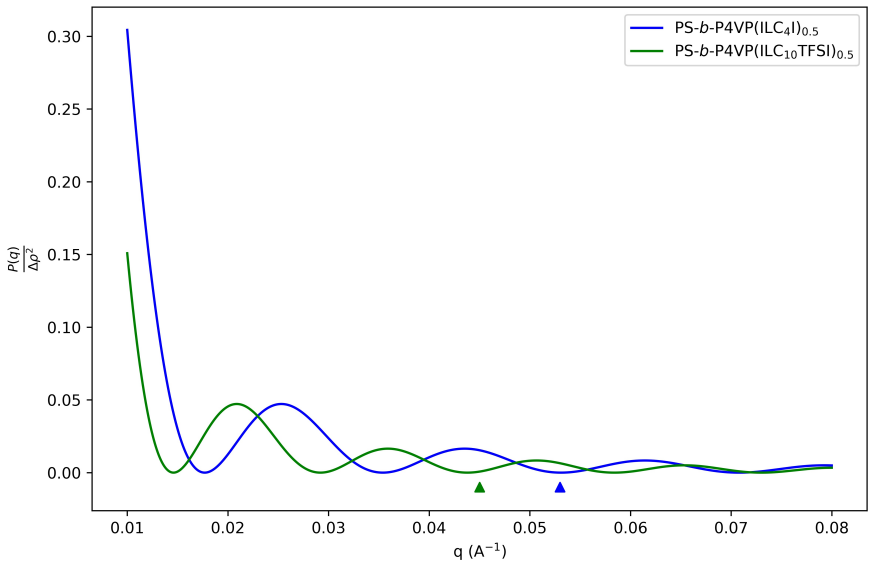


Figure 3.11: Form factor intensity ($P(q)$) of PS-*b*-P4VP(ILC_4I)$_{0.5}$ (blue) and PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ (orange) calculated based on the result in the second and third columns. The blue and green triangles are the markers of the dip of form factor square at 0.054 A^{-1} and 0.046A^{-1} respectively, which are corresponding to the weakening of the second order peak of and third order peak of PS-*b*-P4VP(ILC_4I)$_{0.5}$ and PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$.

To further illustrate this estimation, we calculate the form factor intensity quantitatively based on the estimation. The electron density function of lamellar block copolymer can be estimated by rectangle function. The form factor amplitude $F(q)$ is the Fourier transform of rectangle function

$$F(q) = \Delta\rho \text{sinc}(\frac{qR}{2}) \exp(-\frac{iqR}{2}) \tag{3.7}$$

where R is the size of rectangle and $\Delta\rho$ is the electron density contrast between two blocks. The form factor intensity (P) is

$$P(q) = |F(q)|^2 = \Delta\rho^2 \text{sinc}^2(\frac{qR}{2}) \tag{3.8}$$

Based on this model, we plot the form factor intensity (normalized by $\Delta\rho^2$) of lamellar PS-*b*-P4VP(ILC_4I)$_{0.5}$ and PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ based on the estimate in the second and third columns shown in Figure 3.11. The triangles denote the peak positions we observed in 3.8, which are in the dip positions in the corresponding form factors.

Second, we estimate the domain sizes based on the volume fractions calculated in Table 3.2 and the periodicities from SAXS. The periodicity of PS-*b*-P4VP(ILC_4I)$_{0.5}$, PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$, PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$, PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$ is 33.8 nm, 28.3 nm, 29.9 nm, and 27.4 nm,

respectively. The volume fractions of the P4VP(IL)$_{0.5}$ block in these supramolecules are 0.36, 0.41, 0.42, and 0.46, respectively. Then, the P4VP(IL)$_{0.5}$ domain sizes are 12.1 nm, 11.6 nm, 12.9 nm, 12.6 nm, respectively; the PS domain sizes of supramolecules are 21.7 nm, 16.7 nm, 17.0 nm, 14.8 nm, respectively. The calculation shows that the P4VP domain sizes are all about 12 nm whereas the PS domain sizes range from 15 to 22 nm. Two estimation methods are coincident and both support that the P4VP domain sizes do not change much but the PS domain sizes are altered by the IL chemistry. The order of IL volumes are $ILC_4I < ILC_4TFSI \sim ILC_{10}I < ILC_{10}TFSI$. Since the P4VP domain lengths are almost the same, different volumes of ILs will alter cross-sectional areas of BCP chains. The BCP chaining with smaller molecules have smaller cross-sectional areas.

The volume order is $ILC_4I < ILC_4TFSI \sim ILC_{10}I < ILC_{10}TFSI$. Thus, BCP cross sectional area order is PS-*b*-P4VP(ILC_4I)$_{0.5}$ < PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$ PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$ < PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$. Since the degree of polymerization of PS is the same in all supramolecules, the PS must change its chain configuration to accommodate the cross sectional area difference. Thus, the PS domain sizes inversely correlate to the cross sectional areas. The PS domain size order is PS-*b*-P4VP(ILC_4I)$_{0.5}$ > PS-*b*-P4VP($ILC_{10}I$)$_{0.5}$ > PS-*b*-P4VP(ILC_4TFSI)$_{0.5}$ > PS-*b*-P4VP($ILC_{10}TFSI$)$_{0.5}$. For PS-*b*-P4VP(ILC_4I)$_1$, the supramolecule still forms a lamellar structure at r=1. The periodicity is 36.7 nm and the volume fraction of P4VP(ILC_4I)$_1$ block is 0.46. The domain size of P4VP(ILC_4I)$_1$ is 16.9 nm, which is much larger than that in the supramolecules with r=0.5. As more small molecules were added, the P4VP chain becomes more stretched. Based on the results, the P4VP chain configuration is dependent on the stoichiometry but does not appear to be dependent on the IL chemistry. However, the PS chain configuration is strongly correlated to the IL chemistry. Thus, it is feasible to maintain the domain size of P4VP(IL) block using a wide range of IL chemistry and to control the domain size by tuning the stoichiometry.

3.4 Conclusions

In this study, we investigated the BCP-based supramolecules comprised of phenol-functionalized ILs and PS-*b*-P4VP. Hydrogen bonds in IL-containing supramolecules have higher thermal stability in comparison to other CSMs-containing supramolecules. IL-containing supramolecules microphase separate and form ordered lamellar and hexagonal morphologies under different stoichiometry or ILs chemistry, such as counter ion and alkyl chain length of ILs. This study highlights that BCP-based supramolecules provide a platform to control nanostructures of ILs with different stoichiometry, different chemistry, and to achieve IL-containing assemblies with structural stability at elevated temperatures.

3.5 A Short Overview of Other Scientific Discoveries

The main focus of this chapter is to understand chemistry-structure relationship using X-ray scattering. To make the dissertation coherent and concise, I only provide a brief overview.

1. Together with collaborators, we measured the ion conductivity of ionic liquid containing polymer based supramolecules. Not surprisingly, the result suggests that the ion conductivity is significantly higher than poly(ionic liquids) due to the non-covalent bonding.

2. When the alkyl chain length is short (C4), the BCP based supramolecules form micelle structure in chloroform at room temperature with the size of 20-30 nm. This provides a new platform of micelle structure in organic solvents with tunable dielectric properties in the core.

3. When the alkyl chain length is large (C10), the BCP based supramolecules can form different nanostructures on thin film. The structure can also be tailored by the small molecule chemistry and stoichiometry. Moreover, the nanostructure can be further tuned by solvent annealing. Using a mixture of solvent (chloroform and methanol), the surface of the thin film can be reconstructed to form different structures.

Most of the data and discussions have been summarized and will be available when they are published. However, I will move on to the data driven methods in the subsequent chapters.

3.6 Experiment Method

Materials

Chemicals and reagents were purchased from Sigma-Aldrich and used as received unless mentioned. 1-butylimidazole (98%), imidazole (≥99%), 3-Hydroxybenzyl alcohol (99%), potassium iodide (≥99%,), Boron trifluoride methyl etherate (99%), 1-bromodecane (97%), bis(trifluoromethane) sulfonimide lithium salt, (99.95%), toluene (≥99.5%), methanol (99.8%), ethyl acetate (98%), acetone (98%), N,N-dimethylformamide (≥99.8% anhydrous), tetrahydrofuran (≥99.9%, anhydrous), chloroform (contains 100-200 ppm amylenes as stabilizer, ≥99.5%), d-chloroform (99.8% D) and dimethyl-d6 sulfoxide (DMSO, 99.9 atom% D, contains 0.03% v/v TMS) were used as received from Sigma-Aldrich. PS-*b*-P4VP(19k-*b*-5.2k *Da*) was purchased from Polymer Source Inc.

General Methods for Chemical Characterizations

All 1H, 13C, and 19F NMR spectra were recorded on Bruker AVQ-400 MHz spectrometers and are referenced to residual solvent peaks ($CDCl_3$ 1H NMR δ = 7.26 ppm, 13C NMR δ = 77.16 ppm; DMSO-d6 1H NMR δ = 2.50 ppm, 13C NMR δ = 39.60 ppm; Acetone-d6 1H NMR δ = 2.05 ppm). ESI mass spectrometry was performed on a Finnigan LTQFT (Thermo) spectrometer in positive ionization mode. Gel permeation chromatography (GPC) was carried out on a LC/MS Agilent 1260 Infinity set up with a guard and two Agilent Polypore 300 mm × 7.5 mm columns at 35 °C and calibrated to narrow polydispersity polystyrene standards ranging from Mw = 100 to 4,068,981.

Sample preparations

Polymers and ILs in were weighed, mixed and stirred overnight in THF to form 1 – 2% (w/v) stock solutions. Bulk samples were obtained by evaporating solvent and drying in vacuum oven at room temperature. Bulk samples were thermally annealed before SAXS and TEM studies. Detailed conditions are discussed for each measurement.

Fourier-Transform Infrared (FT-IR) Spectroscopy.

Samples were cast between two NaCl pellets, and the absorption spectra were collected using a Nicolet 6700 FT-IR spectrometer. For *in situ* FT-IR, samples on NaCl pellets were heated from room temperature to 10 °C and keep the samples at each temperature for 10 minutes before measurement. The intensity at 1010 cm^{-1} of hydrogen bonded P4VP was calculated by integrating the peak intensity from 1008 cm^{-1} to 1012 cm^{-1} after baseline calibration.

Differential Scanning Calorimetry (DSC)

Differential scanning calorimetry measurements were performed on a TA Instruments DSC Q200. The samples (about 2 mg) were heated from 0 °C to 200 °C at a heating rate of 15 °C/min under nitrogen gas. Three heating and cooling cycles were performed to eliminate the thermal history of the samples. The transitions were collected from the third heating and cooling cycle. Small-Angle X-ray Scattering (SAXS). SAXS studies were carried out at the Advanced Light Source beamline 7.3.3. X-ray source has a wavelength of 1.240 A (10 keV). Spectra were collected on an ADSC Quantum 4u CCD detector with an area of 188 mm × 188 mm (2304 pixels × 2304 pixels) or a Pilatus 1 M detector with an area of 169 mm × 179 mm (981 pixels × 1043 pixels). The 1D SAXS profiles were obtained by circularly averaging the 2D data. Prior to SAXS experiment, samples were mounted in standard differential scanning calorimetry pans and annealed under 120 °C for 5 hours then slowly cooled down to room temperature. For *in situ* SAXS, the DSC pans containing samples were loaded to a heating stage. All SAXS profiles were measured after keeping the samples at each temperature for 20 min.

Small-Angle X-ray Scattering (SAXS)

SAXS studies were carried out at the Advanced Light Source beamline 7.3.3. X-ray source has a wavelength of 1.240 A (10 keV). Spectra were collected on an ADSC Quantum 4u CCD detector with an area of 188 mm × 188 mm (2304 pixels × 2304 pixels) or a Pilatus 1 M detector with an area of 169 mm × 179 mm (981 pixels × 1043 pixels). The 1D SAXS profiles were obtained by circularly averaging the 2D data. Prior to SAXS experiment, samples were mounted in standard differential scanning calorimetry pans and annealed under 120 °C for 5 hours then slowly cooled down to room temperature. For *in situ* SAXS, the DSC pans containing samples were loaded to a heating stage. All SAXS profiles were measured after keeping the samples at each temperature for 20 min.

Transmission Electron Microscopy (TEM)

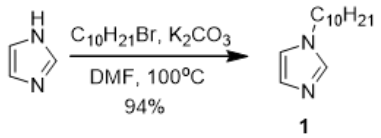
For TEM observations, samples mounted in the DSC pan were annealed at 120 °C for 5 hours then slowly cooled to room temperature. Samples were embedded in resin (Araldite 502, Electron Microscopy Sciences) and cured at 60 °C overnight. Thin sections about 70 nm in thickness were microtomed using an RMC MT-X Ultramicrotome (Boeckler Instruments) and picked up on carbon-coated Cu grids on top of water. The thin sections were exposed to iodine vapor for 1 hour to stain the P4VP domain selectively and imaged using a FEI Tecnai 12 TEM operating at 120 kV accelerating voltage or a JEOL 2100 TEM operating at 200 kV.

Rheological measurements

The samples were loaded on the rheometer at 120 °C and then equilibrated for 30 minutes to eliminate the thermal history and then cool down to starting measuring temperature by 1 °C/min. Oscillatory shear measurements were performed at 0.5% strain amplitude (in linear viscoelastic regime) with a 15 mm diameter parallelled plate based on modulus at a gap height of 0.5 mm using a stress controlled oscillatory rheometer (Physica MCR 302 Modular Compact Rheometer, Anton Paar, Ashland, VA). Frequency sweeps from 0.01 to 10 Hz were applied to determine storage (G') and loss (G") modulus. Time-temperature superposition (tTS) master curve were obtained by frequency sweep after keeping the samples at each temperature for 10 minutes. The temperature ranges are from 30 °C to 150 °C.

Syntheses of Small Molecules

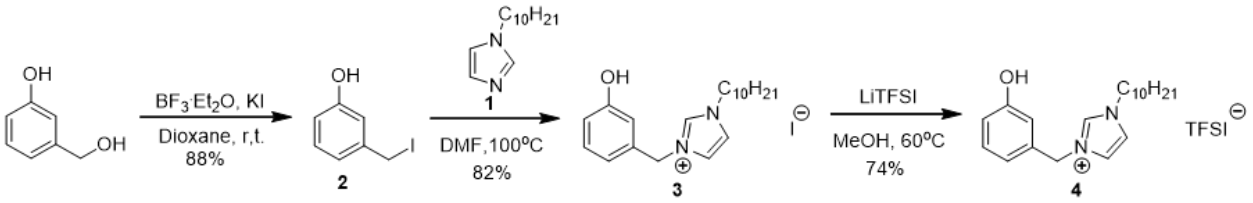
All of the NMR and ESI spectrum are available in the supporting information of Shuai Liu et al., "Ionic Liquids Containing Block-Copolymer Based Supramolecules" from Macromolecules, **2016**, 49 (16), 6075-6083.



1

1-decylimidazole (1) A 20mL glass vial was charged with imidazole (0.75 g, 11 mmol, 1.1 equiv.) and K_2CO_3 (1.78 g, 16.5 mmol, 1.65 equiv.) in DMF (8 mL). The reaction was stirred at 100 °C overnight. Then, 1-bromodecane (2.21 g, 10 mmol, 1.0 equiv.) was added and the mixture was stirred for another 24 h. The solvent was removed under vacuum. Chloroform (15 mL) was added in residue and washed by water (15 mL) for three times. The organic layer was collected and dried by anhydrous Na_2SO_4. The solvent was removed under vacuum to yield 1 as light yellow oil (2.34 g, 94%). 1H NMR (400 MHz, $CDCl_3$) δ 7.42 (s, 1H), 7.01 (s, 1H), 6.87 (s, 1H), 3.88 (t, J = 7.2 Hz, 2H), 2.74 (m, 2H), 1.94 - 1.54 (m, 2H), 1.43 - 1.13 (m, 12H), 0.84 (t, J = 6.9 Hz, 3H).

13C NMR (101 MHz, CDCl3) δ 137.03, 129.26, 118.77, 47.05, 31.85, 31.08, 29.48, 29.43, 29.26, 29.07, 26.54, 22.67, 14.12., which is consistent with literature[112].

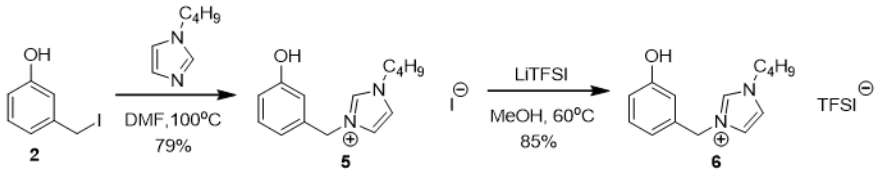


3-(iodomethyl)phenol (2) A 250mL round bottom flask was charged with 3-hydroxymethylphenol (5.01 g, 40 mmol, 1.0 equiv.) and KI (7.03 g, 42 mmol, 1.05 equiv.) in dry 1,4-dioxane (100 mL). The reaction mixture was stirred at room temperature for 30 min under nitrogen gas. Then, boron trifluoride diethyl etherate (5.3 mL, 42 mmol, 1.05 equiv.) was added into the mixture. The reaction mixture was stirred for another 5 h at room temperature. Then the solvent was removed under vacuum. The residue was dissolved in chloroform (300 mL) and wash by water (200 mL) for three times. The organic layer was collected and dried by anhydrous Na_2SO_4 and concentrated on a rotary evaporator. Column chromatograpy (silica gel; 10:1 hexane/ethyl acetate) yielded 2 (8.27g, 79%) as white powder. 1H NMR (400 MHz, $CDCl_3$) δ 7.17 (m, 1H), 6.96 (d, J = 7.6 Hz, 1H), 6.86 (s, 1H), 6.73 (d, J = 8.1 Hz, 1H), 5.44 (s, 1H), 4.39 (s, 2H). 13C NMR (101 MHz, $CDCl_3$) δ 155.50, 141.00, 130.18, 121.34, 115.75, 115.21, 5.51. which were consistent with report[113].

1-decyl-3-(3-hydroxybenzyl) imidazolium iodine ($IL_{10}I$, 3) A 20mL glass vial was charged with 3-(iodomethyl)phenol (2) (0.92 g, 4 mmol, 1.0 equiv) and 1-decylimidazole (1) (0.83 g, 4 mmol, 1.0 equiv.) in DMF (10 mL). The reaction mixture was stirred at 100 °C overnight. The solvent was removed under vacuum. Chloroform (20 mL) was added into the residue and washed by water (20 mL) for three times. The organic layer was collected and dried by anhydrous Na_2SO_4 and concentrated on a rotary evaporator. Column chromatograpy (silica gel; gradient elution from 1:1 ethyl acetate/hexane to 20:1 hexane/methanol) yielded 3 (1.44g, 82%) as yellow oil. 1H NMR (400 MHz, $CDCl_3$) δ 9.77 (s, 1H), 7.37 (m, 1H), 7.33 (m, 1H), 7.24 (s, 1H), 7.15 (d, J = 7.8 Hz, 1H), 7.04 - 6.98 (m, 1H), 6.90 - 6.84 (m, 1H), 5.40 (s, 2H), 4.36 - 4.14 (m, 2H), 2.11 - 1.75 (m, 2H), 1.44 - 1.18 (m, 14H), 0.91 (t, J = 6.9 Hz, 3H). 13C NMR (101 MHz, CDCl3) δ 157.27, 135.64, 134.29, 130.44, 122.48, 120.06, 119.63, 117.24, 116.05, 52.95, 50.42, 31.85, 30.14, 29.47, 29.40, 29.26, 28.98, 26.27, 22.67, 14.15. FTMS (HR-ESI positive): $[C_{20}H_{31}O_1N_2]^+$ cal. 315.2431; found, 315.2427

1-decyl-3-(3-hydroxybenzyl) imidazolium bis(trifluoromethylsulfonyl)imide (IL_{10}TFSI, 4) A 4 mL glass vial was charged with 1-decyl-3-(3-hydroxybenzyl) imidazolium iodine (3) (221 mg, 0.5 mmol, 1.0 equiv), lithium bis(trifluoromethylsulfonyl)imide (214 mg, 0.75 mmol, 1.5 equiv.) in methanol (2 mL). The reaction mixture was stirred at 60 °C overnight. The solvent was removed under vacuum. Chloroform (5 mL) was added in residue and washed by water (5 mL) for three times. The organic layer was collected and dried by anhydrous Na_2SO_4. The solvent was removed under vacuum to yield 4 as light yellow oil (218 mg, 92%). 1H NMR (400 MHz, $CDCl_3$) δ 8.68

(s, 1H), 7.30 - 7.15 (m, 3H), 6.93 - 6.79 (m, 3H), 5.17 (s, 2H), 4.17 - 4.05 (m, 2H), 2.02 - 1.64 (m, 2H), 1.26 (m, 14H), 0.87 (t, J = 6.9 Hz, 3H). 13C NMR (101 MHz, $CDCl_3$) δ 157.07, 134.74, 133.58, 130.62, 122.27, 120.18, 116.79, 115.44, 99.85, 53.22, 50.15, 31.69, 29.81, 29.25, 29.13, 29.07, 28.68, 25.97, 22.51, 13.94.



1-butyl-3-(3-hydroxybenzyl) imidazolium iodine (IL_4I, 5) A 20 mL glass vial was charged with 3-(iodomethyl)phenol (2) (1.15 g, 4 mmol, 1.0 equiv) and 1-butylimidazole (0.62 g, 5 mmol, 1.0 equiv.) in DMF (10 mL). The reaction mixture was stirred at 100 °C overnight. The solvent was removed under vacuum. Chloroform (20 mL) was added into the residue and washed by water (20 mL) for three times. The opaque organic layer was collected and dried by anhydrous Na_2SO_4 and concentrated on a rotary evaporator. Column chromatograpy (silica gel; gradient elution from 1:1 ethyl acetate/hexane to 20:1 hexane/methanol) yielded 7 (1.40g, 79%) as yellow oil. 1H NMR (400 MHz, DMSO) δ 9.70 (s, 1H), 9.30 (s, 1H), 7.81 (m, 2H), 7.20 (d, J = 7.8 Hz, 1H), 6.86 âĂŞ 6.69 (m, 3H), 5.34 (s, 2H), 4.19 (t, J = 7.2 Hz, 2H), 3.50 (s, 28H), 2.51 (s, 2H), 1.83 - 1.71 (m, 2H), 1.25 (dt, J = 14.8, 7.4 Hz, 2H), 0.89 (t, J = 7.4 Hz, 3H). 13C NMR (101 MHz, DMSO) δ 157.92, 136.30, 136.21, 130.34, 122.93, 122.82, 118.80, 115.83, 115.12, 52.13, 48.92, 31.47, 19.00, 13.49. FTMS (HR-ESI positive): $[C_{14}H_{19}O_1N_2]^+$ cal. 231.1492; found, 231.1490

1-butyl-3-(3-hydroxybenzyl) imidazolium bis(trifluoromethylsulfonyl)imide (IL_4TFSI, 6) A 4 mL glass vial was charged with 1-butyl-3-(3-hydroxybenzyl) imidazolium iodine (5) (179 mg, 0.5 mmol, 1.0 equiv), lithium bis(trifluoromethylsulfonyl)imide (214 mg, 0.75 mmol, 1.5 equiv.) in methanol (2 mL). The reaction mixture was stirred at 60 °C overnight. The solvent was removed under vacuum. Chloroform (5 mL) was added in residue and washed by water (5 mL) for three times. The organic layer was collected and dried by anhydrous Na2SO4. The solvent was removed under vacuum to yield 6 as light yellow oil (215 mg, 85%). 1H NMR (400 MHz, DMSO) δ 9.71 (s, 1H), 9.25 (s, 1H), 7.77 (m, 2H), 7.21 (m, 1H), 6.86 âĂŞ 6.70 (m, 3H), 5.31 (s, 2H), 4.17 (t, J = 7.2 Hz, 2H), 1.87 - 1.65 (m, 2H), 1.24 (dt, J = 14.8, 7.4 Hz, 2H), 0.89 (t, J = 7.4 Hz, 3H). 13C NMR (101 MHz, DMSO) δ 158.04, 136.31, 130.38, 122.95, 122.90, 121.34, 118.83, 118.14, 115.92, 115.20, 52.25, 49.00, 31.54, 19.05, 13.43. 19F NMR (376 MHz, DMSO) δ -78.02.

3.7 Acknowledgement

This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, of the U.S. Department of Energy under Contract DE-AC02-05CH11231. The Advanced Light Source is supported by the Director, Office of Science,

Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract DE-AC02-05CH11231.

3.8 Supplementary Information

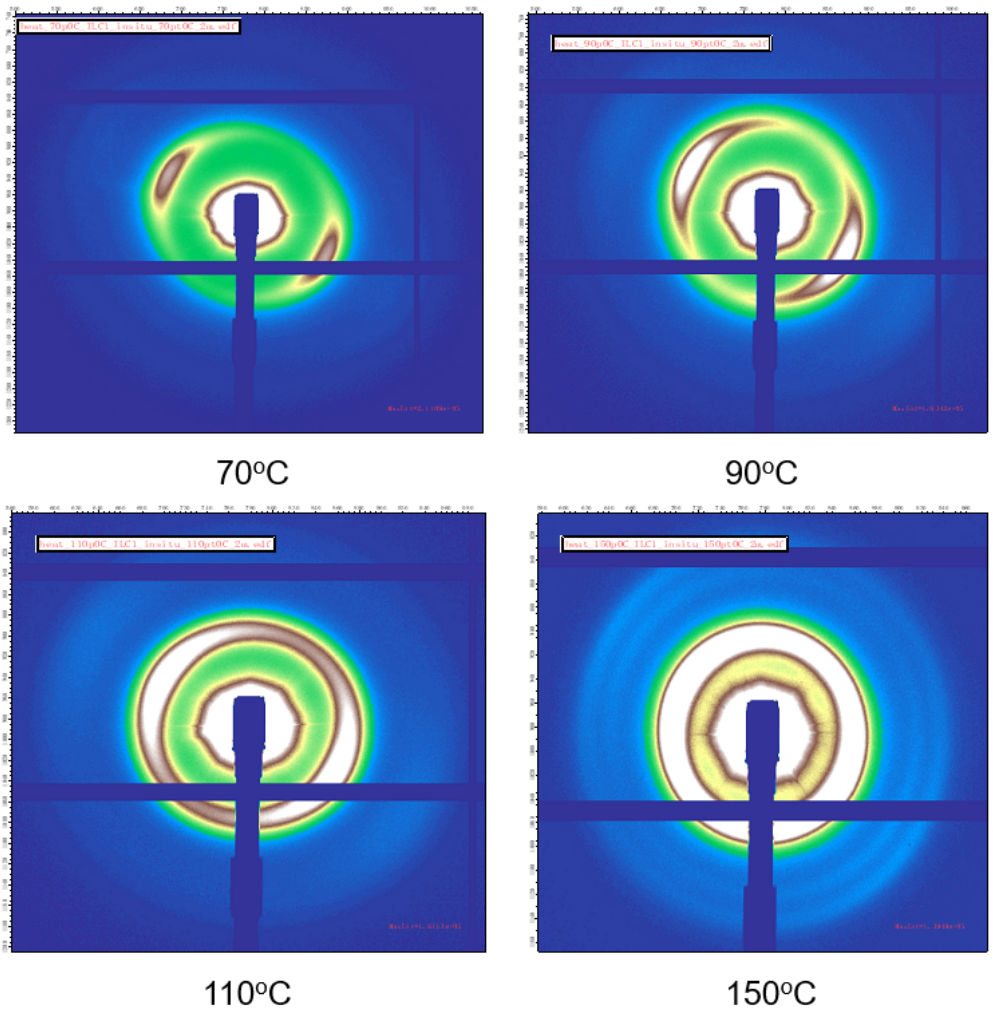


Figure 3.12: Small angle X-ray scattering data during thermal annealing process of PS-*b*-P4VP(ILC_4TFSI)$_1$. Before the thermal annealing, the supramolecule has anisotropic structure based on sample processing history. By thermal annealing above the T_g, the anisotropic behavior is eliminated.

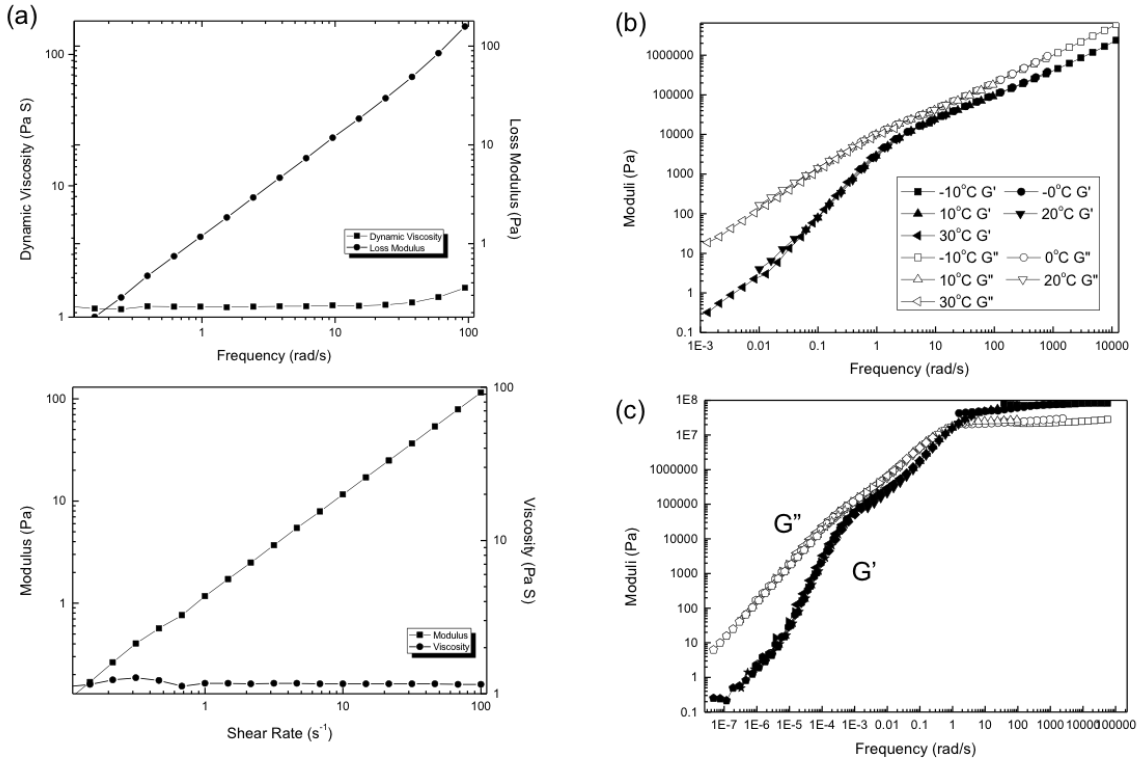


Figure 3.13: Rheology behavior of ILC_4TFSI and supramolecules. (a) Rheology behavior of small molecule(ILC_4TFSI) under different shear rate and oscillate frequency at room temperature. (b) Time-temperature superposition master curve of P4VP(ILC_4TFSI)$_1$ from -10 °C to 30 °C (10 °C as reference). (c) Time-temperature superposition master curve of PS-*r*-P4VP(ILC_4TFSI)$_1$ from 20 °C to 120 °C (80 °C as reference)

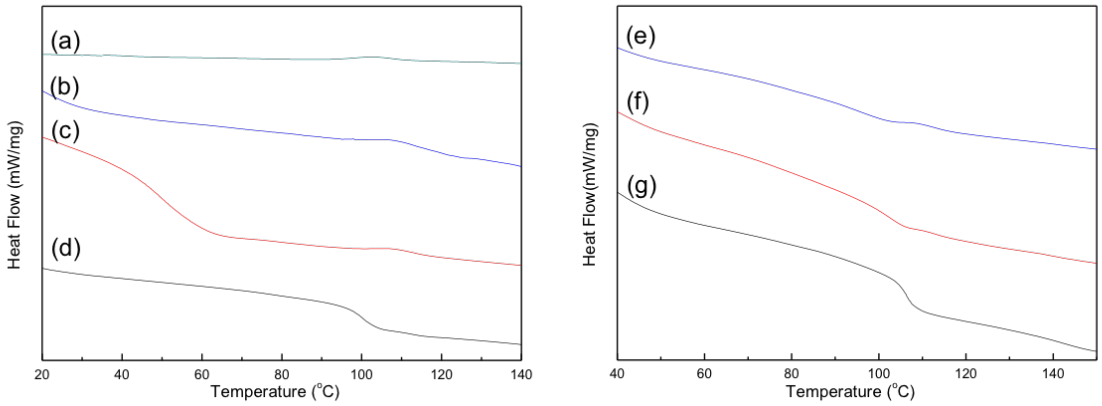


Figure 3.14: DSC scan of (a) ILC_4TFSI (b) $P4VP(ILC_4TFSI)_1$ (c) PS-*r*-P4VP(ILC4TFSI)1 (d) PS-*r*-$P4VP(ILC_4TFSI)_1$ Temperature ramp at 10 °C/min, using the third heating-cooling-heating cycle for analysis.

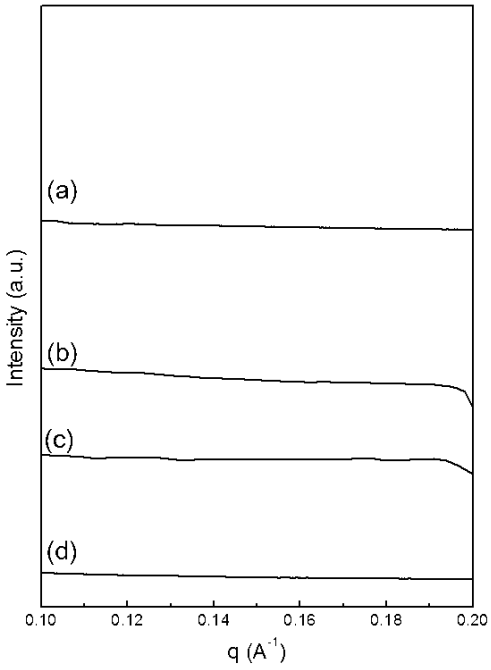


Figure 3.15: Small Angle X-ray Scattering profiles at q from 0.1 A^{-1} to 0.2 A^{-1}. (a) PS-*b*-P4VP(ILC_4I)$_1$ (b) PS-*b*-P4VP(ILC_4TFSI)$_1$ (c) PS-*b*-P4VP($ILC_{10}I$)$_1$ (d) PS-*b*-P4VP($ILC_{10}TFSI$)$_1$

Chapter 4

A Data Driven Framework: Hierarchical X-ray Scattering Experimental Discovery

In chapter 3, I presented an experimental X-ray scattering study on the chemistry-structure relationship of supramolecules, and described how structural information was derived from the data. However, this conventional method is time-consuming, and hence does not apply to high-throughput materials chemistry discovery. In this chapter, I start to investigate data driven approaches with applications on materials chemistry systems. In the first part of this chapter, I build a database containing a large number of experimental X-ray scattering data with feature based labels. In the second part of this chapter, I demonstrate a hierarchical categorization method by combining machine learning methods and domain knowledge. In the third part of this chapter, I apply this system to two X-ray scattering studies. Finally, I discuss the importance of the data and the future directions of this platform.

4.1 Introduction

X-ray scattering can be used to characterize different materials chemistry systems. At present, X-ray scattering experiments can be conducted in a high-throughput manner using high speed detectors and high flux sources [114–119], hence it is therefore important to develop scientific procedures to manage and analyze large-scale datasets. In this chapter, I propose a machine learning based hierarchical categorization approach to manage and classify the data, so that an appropriate analysis pipeline can be applied autonomously in near real-time. More importantly, this framework can be potentially integrated into an automatic materials chemistry discovery process, together with high-throughput synthesis and robotic X-ray scattering experiments.

X-ray scattering data can be categorized based on different criteria discussed in chapter 2.1. For example, depending on the geometry of experiment, it can be categorized as transmission or grazing incidence X-ray scattering data. In addition, the data can also be categorized by its features. Different features, such as rings, arcs, rods and Bragg peaks, need their corresponding toolkits [27, 120]. For example, if ring features exist in transmission X-ray scattering data, the radial integration is commonly applied to extract the information in reciprocal space [121, 122]. If X-ray scattering data does not have any obvious feature, there is no need for further evaluations.

Recently, histogram of gradient (HOG) feature extraction with SVM classifier have been applied to predict X-ray scattering experiment configurations with more than 80 classes [67, 123]. However, this approach does not provide further insight on the underlying structures. Herein, we propose a novel and more general framework for materials chemistry discovery using X-ray scattering platform by leveraging the large-scale experiment database and different machine learning methods. First, we start with organizing scattering data into a flexible database containing experiment information, labels from domain experts, and predicted labels from trained machine learning models. Together with our collaborators, we build a database containing more than 500,000 images. A convenient web application for data labeling was developed[1], where we obtain 10,994 labeled experimental images. Later, we build machine learning models using a hierarchical approach, which allows us to categorize each X-ray scattering data's features individually, starting from the coarse-grain information (such as geometry of X-ray scattering experiment), to the fine-grain information (such as ring or crystalline features). Last, we apply this model to materials chemistry systems to demonstrate its application.

[1]This application was developed by Dr. Ronald Pandolfi. I built the database and contributed to the data labeling. All other materials present in this chapter are my own work.

4.2 X-ray Scattering Database with Feature based Labels

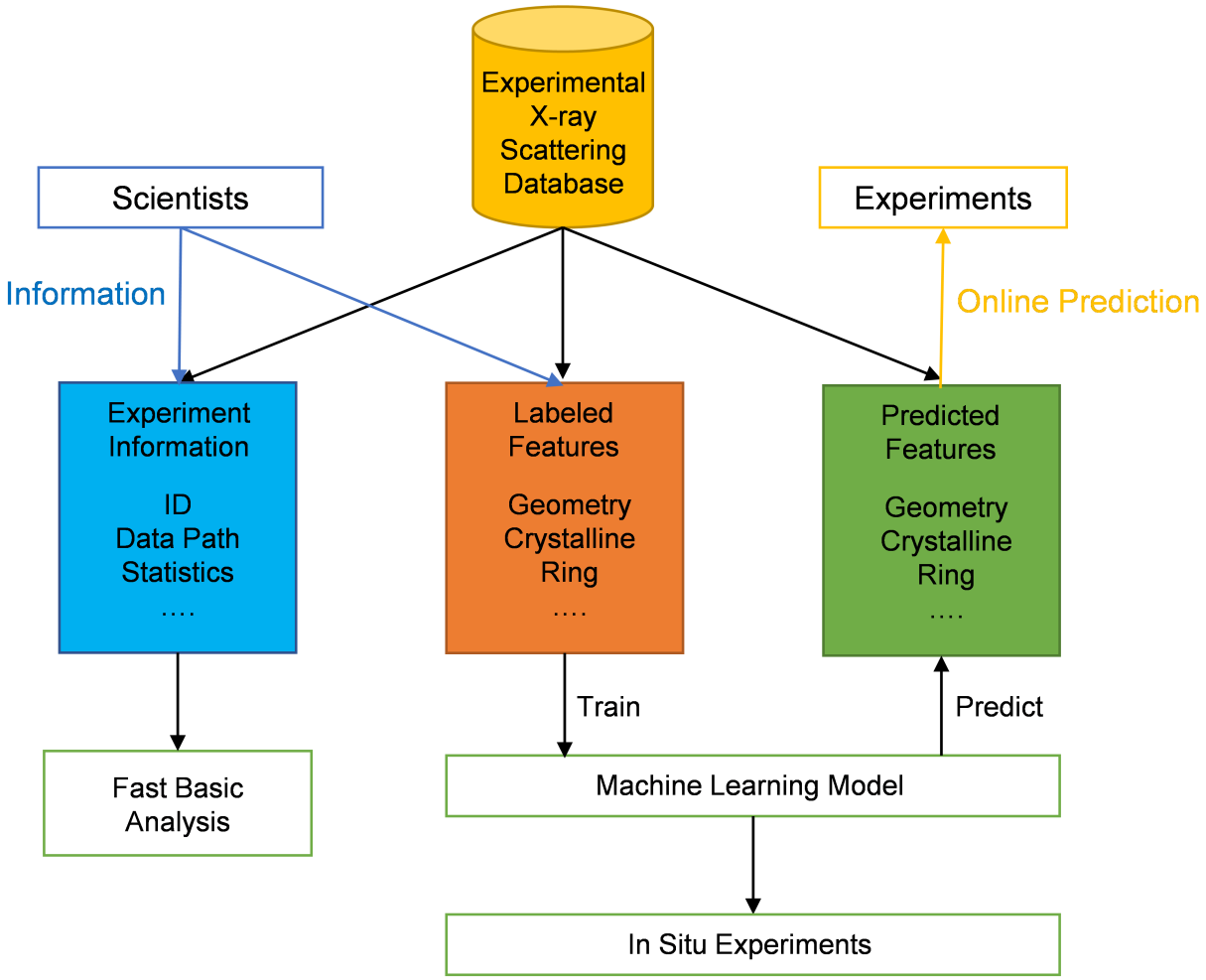


Figure 4.1: The designed database containing basic experimental information, labels from domain experts and predictions from machine learning models.

Stage	Description
1	Identify the geometry of scattering experiment
2	Given a certain geometry of scattering experiment, classify if scattering data has feature or not
3	Given the scattering data has feature, perform binary classification on each feature
4	Given the scattering image has certain feature, identify certain parameters (e.g., crystal lattice and orientation)

Table 4.1: Descriptions of the four stages.

We built a database of a large number of experimental X-ray scattering patterns using MongoDB as the backend. Each X-ray scattering image has its own basic information (metadata), such as a unique ID, data path, and some basic statistics from the data. A subset of the data (10,994 images) were labeled by domain experts. Then, we trained the machine learning models using X-ray scattering data and the corresponding labels. Later, we applied trained machine learning models to predict the features of unlabeled experimental data. In this chapter, the supervised learning methods are trained and evaluated using 10,994 X-ray scattering patterns labeled by domain experts. First, we shuffle the dataset and divide them into training and testing dataset with a ratio of 4 to 1. The size of training dataset is 8,975 and the size of testing dataset is 2,019. To expand dataset, we augment each data to 10-fold for training and testing dataset separately[2]. The size of the final training dataset is 89,750. In the feature extraction stage, we combine both training dataset and some unlabeled images to train the autoencoder model. The hypothesis is that including unlabeled dataset is helpful to generalize the feature extraction network. The detailed data preprocessing and augmentation procedure are discussed in supplementary information.

[2]The purposes of augmenting training and testing datasets are different. Training dataset augmentation is designed to make the machine learning model more generalizable. Augmentation of the testing dataset is to test the robustness of machine learning model under different simulated conditions.

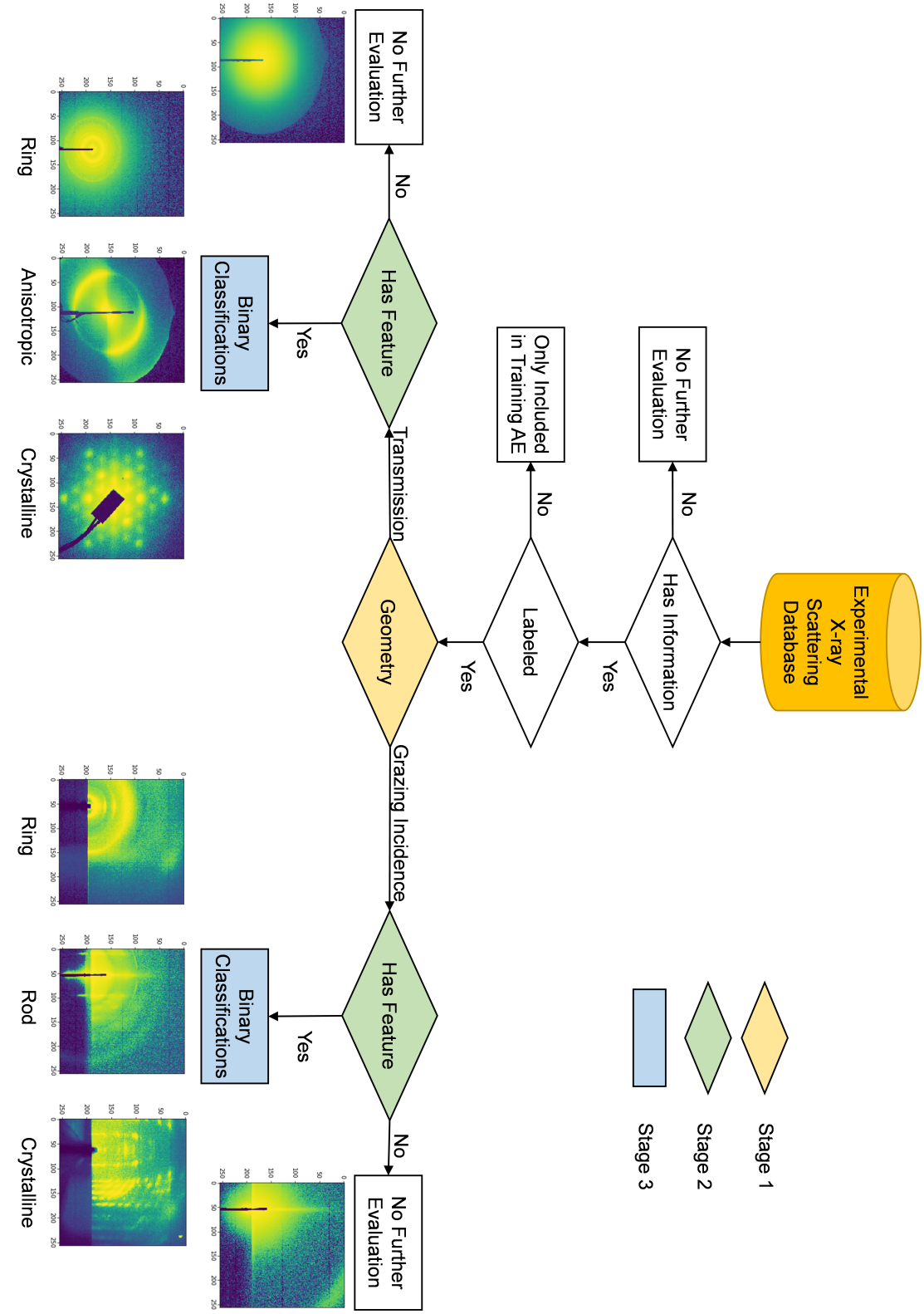


Figure 4.2: The hierarchical categorization framework for X-ray scattering data.

4.3 The Hierarchical Categorization Method

Hierarchical Categorization

There are four stages in our proposed hierarchical categorization method, from coarse-grain features to fine-grain features shown in Figure 4.2. Table 4.1 describes each of the four stages. During the first stage, the scattering data is categorized by the geometry of X-ray scattering experiment. Under each geometry, the data will be categorized as featureless or not. If the data is featureless, it does not require further processing in our pipeline. An important note is that "featureless" is not equivalent to no information. Featureless refers to not containing the feature defined in this framework. Otherwise, it will be passed to the next stage. Then, the data will be categorized for each possible feature in a binary way. During the final stage, for each feature, a dedicated machine learning model will be trained to understand the underlying structures. However, the label in this stage cannot be easily obtained. Alternatively, we propose to generate the data using simulation, which will be discussed in chapter 5.

The Probabilistic Interpretation

At each stage, a dedicated model is trained to calculate the probability in each binary classification task (except SVM, which gives a score but not explicit probability). For example, the probability that a data is a GISAXS image with ring and rod features (but no other features, such as crystalline) can be formulated as:

$$\begin{aligned} P(\text{GISAXS with ring and rod features}) = {} & P(\text{Geometry=GISAXS}) \\ & \times P(\text{Feature=True}|\text{Geometry=GISAXS}) \\ & \times P(\text{Ring=True}|\text{Feature=True,Geometry=GISAXS}) \\ & \times P(\text{Rod=True}|\text{Feature=True,Geometry=GISAXS}) \\ & \times P(\text{Crystalline=False}|\text{Feature=True,Geometry=GISAXS}) \end{aligned} \tag{4.1}$$

The pros and cons of general hierarchical method has been discussed in literature [124]. There are several considerations that we utilize the hierarchical modeling framework for X-ray scattering experiment by the domain knowledge. The model can be easily modified given different experimental pre-information. For example, if we know the experimental geometry is transmission X-ray scattering, we can simply set P(Geometry=SAXS)=1 without further complication. In addition, if we are only interested in certain features (e.g., crystalline or not), the probability of that feature can be easily predicted. This provides the opportunity for exploring the structure of certain materials chemistry systems under different environment or synthesis conditions. Two examples will be presented in section 4.5.

Details of Each Stage

Currently, this pipeline and database are highly specialized for the SAXS/WAXS beamline at ALS with a limited number (11k) of labeled data. They are from the materials chemistry systems contributed by several research groups. Therefore, the size and variety of the dataset is still limited. However, this chapter is focused on illustrating this feature-based database, hierarchical framework and the applications in materials chemistry. Here, we present the preliminary classification results and the discussions using the current dataset.

Since the data are from several research groups, even though each sample is from a unique measurement, the training and testing dataset still share many similarities. Therefore, the "*testing accuracy*" in this section is better described as "*validation accuracy*" within certain materials chemistry systems by conducting validation on hold-out dataset. For example, we observe extensive Poly(3-hexylthiophene-2,5-diyl) (P3HT) characterization data in both training and testing dataset. To mitigate this issue, we utilize the data augmentation to simulate the effect of different beamstop positions and/or feature sizes. Moreover, we also tested the trained model using a small dataset with different underlying materials chemistry. We are planning to get access to a larger dataset with more diverse scattering data from a variety of materials chemistry systems to make this study more comprehensive. We will present a more detailed comparison of accuracy between different stages and different machine learning algorithms using the comprehensive dataset. The details of the limitations and the future improvements are available in section 4.7.

In this chapter, we evaluate the first three stages in this hierarchical framework. Due to the complexity of the last stage, we will train the machine learning model using simulation data, which will be illustrated in chapter 5. The first stage in our framework identifies the geometry of X-ray scattering experiment. The geometry of the X-ray scattering experiment can be identified by both yoneda peak, the mirroring symmetry across the specular plane and even beam stop/feature position in many cases, which is a simple task. The prediction of scattering geometry is the first and also very important step in our hierarchical categorization framework because it will direct the data into two different branches: transmission and grazing incidence X-ray scattering data for the further data processing. We obtain > 95% accuracy on the testing dataset with similar underlying materials chemistry (significantly higher than the baseline that 63% of the data are grazing incident X-ray scattering data) and 86% accuracy on the testing dataset with different underlying materials chemistry.

After geometry identification, two models using either transmission or grazing incidence X-ray scattering data are built separately to identify if the scattering data contains important features. In comparison to the previous task, we observe slightly lower accuracy (about 92% accuracy on the testing dataset with similar underlying materials chemistry and 80%-85% accuracy on the small dataset with different underlying materials chemistry). We hypothesize that this is due to the interference from background and noise, such as Poisson shot noise. This stage is important for data processing and storage: the featureless data can be filtered and possibly moved to cold storage.

An example of its application on the data processing in chapter 3 will be demonstrated in section 4.5.

The third stage is to perform binary classification for each feature. The motivation is that a single scattering image can contain multiple features which need to be identified separately. However, there are several challenges in this stage. First, due to the filtering process in the first and second stage, there are only limited number of data left in this stage. Therefore, we can only train and evaluate the machine learning model on GISAXS branch with a subset of features (ring and crystalline). Second, in comparison to the labeling at the first two stages, during the data labeling process, the feature in this stage is sometimes difficult to be labeled. Third, the data in this stage becomes imbalanced: one class is usually dominant. Moreover, due to limited number of training data in this stage, we also observe that the machine learning model is not very robust in the real applications. Details will be discussed in section 4.5. We expect these issues can be addressed when we obtain a larger dataset.

4.4 Potential Applications to Experimental Systems

In this section, we provide two preliminary examples to illustrate how our framework can be potentially applied to materials chemistry discovery.

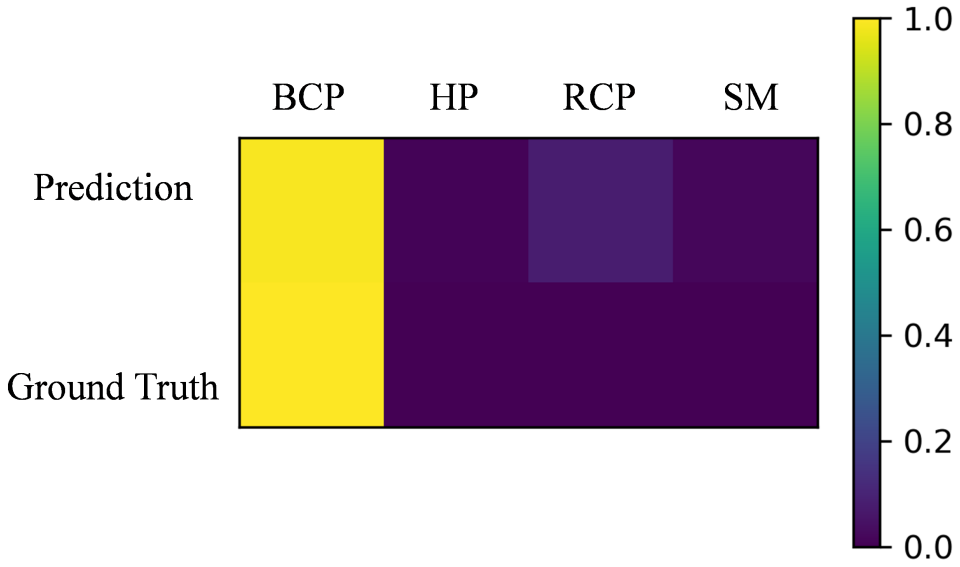


Figure 4.3: Predicted probability by CNN that the SAXS data has feature and its ground truth. BCP, HP, RCP, SM are ionic liquid containing block-copolymer based supramolecule, homopolymer based supramolecule, random copolymer based supramolecule and small molecule, respectively.

We demonstrate the applications of hierarchical model to the ionic liquid containing polymer based supramolecule in chapter 3. The model in transmission branch at stage 2 can be applied

to distinguish if the SAXS data has feature or not. Based on the discussion in chapter 3, only BCP based supramolecules has the order (by microphase separation). The prediction from the CNN model is consistent with these analysis. Using this example, we demonstrate the potential application of stage 2 (featureless or not) in materials chemistry systems.

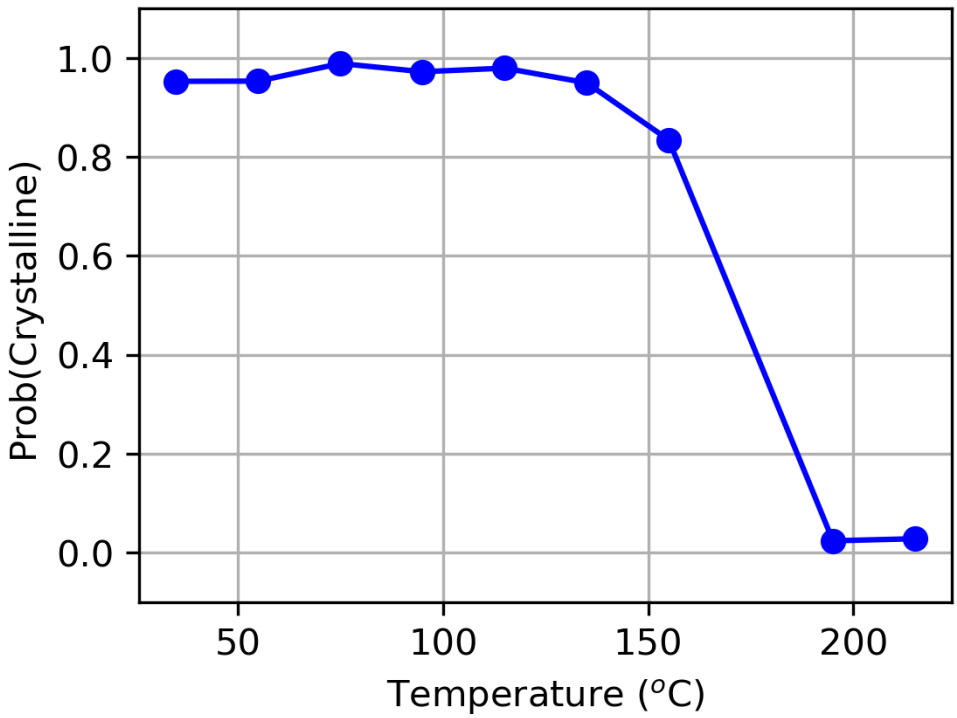


Figure 4.4: Decomposition process of metal-organic chalcogenolate during *in situ* GIWAXS experiment predicted by CNN model. Above the decomposition temperature, the crystalline feature vanishes in the scattering data. The original decomposition data is published in [125].

Another example is the decomposition process of metal-organic chalcogenolate [125]. Figure 4.4 shows the probability of crystalline feature predicted by the machine learning model[3]. The prediction made by the machine learning model is generally consistent with GIWAXS results. However, we found that the predicted probability is dependent on the data preprocessing. When we move the beamstop position on the vertical direction by image processing, the predicted probability becomes different. In that case, the result at temperature 155 °C is predicted as amorphous (with probability 0.3), which is misclassified. Even though we conducted the data augmentation on the training dataset by zoom in and random cropping, this example indicates that the robustness of the model still needs to be improved for future applications.

[3]An important note is that, we are aware that there are existing packages for crystallinity analysis from scattering or diffraction data [126]. However, our framework can be easily generalized to the analysis of other features.

4.5 Conclusions

In this chapter, I present a novel hierarchical approach for X-ray scattering data toward accelerating materials chemistry characterization data analysis. We build the first experimental X-ray scattering database containing feature-based labels, which is motivated by the hierarchical X-ray scattering analysis pipeline. Our approach consists of four stages: scattering geometry, featureless or not, binary classification of each feature, and structure identification given the feature information. In comparison to conventional categorization approaches, our method has higher flexibility, where different models can be utilized or combined easily for different tasks. We perform a preliminary classification study using this system and point out that the number and variety of current dataset need to be improved. Finally, we present two preliminary examples to demonstrate its potential applications in materials chemistry systems using SAXS and GIWAXS experiments, respectively. This pipeline requires high-quality, large-scale and diverse dataset. During the model training and evaluation, we identified several problems due to limited number of data, such as the= generalization problem, imbalanced dataset and the robustness issue. We plan to improve this platform by taking more data and materials chemistry systems into the database in the near future.

4.6 Future Directions and Progresses

Improvement of the Datasets and Model

Currently, the number of labeled data is still limited (11k samples), which is the most significant limitation in the study. In addition, it only contains the data from research groups utilizing beamline 7.3.3 at the ALS. These two limitations open up opportunities for future improvements:

1. More balanced dataset. The distribution of samples in this database may not perfectly reflect the distribution over all the scattering data. For example, in our database, about 80% of SAXS images with features contain the rings, which is imbalanced. To mitigate this issue, we attempted to subsampling on the dominate class. However, the ideal solution is to include more diverse data.

2. Better generalization. Even though each of the image in the database is from unique measurement, the training and testing dataset still share many similarities because the data are from a limited number of research groups. The model may fail when it is tested on the experimental data that is significantly different than the current dataset (e.g., with different chemistry or different beamlines). The generalization issue is very common in the data science field. For example, Recht et al. reported that the ImageNet could also have the generalization issue [127]. A consequent problem is that the evaluation process is difficult. Under the current setting, rather than reflecting the accuracy in the real experiment scenarios, our testing accuracy only serves as a benchmark that reflects the accuracy within limited materials systems. We also considered to use the evaluation metric calculated by training on one materials system and testing on another. However, this evaluation is also problematic because it takes the

assumption that the testing materials chemistry system is significantly different than any of the materials chemistry systems in the training dataset. The ideal solution is to have better understanding on the distribution of the data and obtain a dataset that is more diverse to make the model more generalizable.

3. Robustness of the model. Some noises, or even simple experimental configuration (e.g. beamstop position) can influence the prediction result. Our next step is to improve the robustness of the model.

4. More stages and features. We plan to extend this framework to include more stages and/or more categories to reflect more detailed features.

Another effort has been involved to build an easy-to-use tagging pipeline to obtain more labeled experiment image. Moreover, we are building the systems to label images during experimental data collection and update the model in an online fashion. These extensions will further improve the capability of the current method.

Improve Machine Learning Models

In this chapter, we only implemented a naive autoencoder architecture. There are several different strategies to extract the features using autoencoder, which may further improve the performance of current strategy. Moreover, the architecture of neural networks and hyperparameters of machine learning models can be furtehr optimzied toward better performance.

In recent years, many region-proposal CNNs (R-CNNs) have been developed for object recognition [128–131]. In the future, we plan to build a more advanced database with the bounding boxes of the local features to deploy these models. This will further expedite the X-ray scattering data analysis and materials discovery process.

Collaborations with Materials Project

Together with the automatic materials synthesis pipelines, our platform opens up the opportunity for high-throughput experimental discovery. We aim to generate high-throughput experimental data (containing both chemistry and characterized structure information of the materials) for the materials project database.

Toward Automatic Screening of Chemistry-Structure Relationship

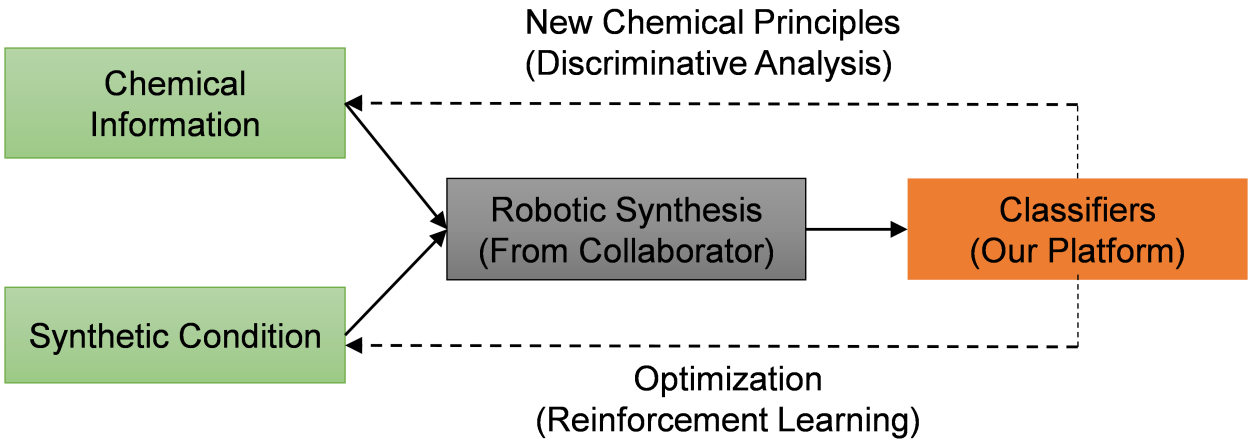


Figure 4.5: Future automatic materials chemistry discovery based on this framework.

We are designing systematic experiments to explore the chemistry-structure relationship of different materials chemistry systems using this platform. In collaboration with experimental groups, we are integrating the automatic synthesis and chemical information extraction into this framework. With large number of data and discriminative analysis, we hope to gain more chemistry-structure relationship insights in a high-throughput fashion. Figure 4.5 shows a diagram of the materials chemistry discovery cycle containing this framework. By combining robotic materials synthesis platform and our methodology, we are able to generate large database containing both the chemical reaction conditions and the corresponding characterization results. We aim to correlate the chemistry-structure relationship automatically through this framework. Moreover, using this framework, we plan to optimize the reaction conditions automatically using deep reinforcement learning methods. We plan to close this loop toward next-generation data-driven materials chemistry discovery. In principle, our framework is compatible with large-scale screening of experimental samples. We are working on the first proof-of-concept example to abstract the chemical knowledge by combining the robotic synthesis and our platform.

4.7 Acknowledgement

We acknowledge J. Nathan Hohman group for allowing us to use their data in section 4.5. This work was supported by the Center of Advanced Mathematics for Energy Research Applications (CAMERA) through the Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Early Career Program. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

4.8 Supplementary Information

Data Preprocessing

To adjust the experimental image appropriately and augment the size of training data, we perform the following procedure:

1. Remove ar interpolate the masked region.
2. Resize the image and perform random crop (at most 5%) to constant 256×256 size.
3. Adjust the image to mitigate the effect of different intensity/integration time.

Machine Learning Model

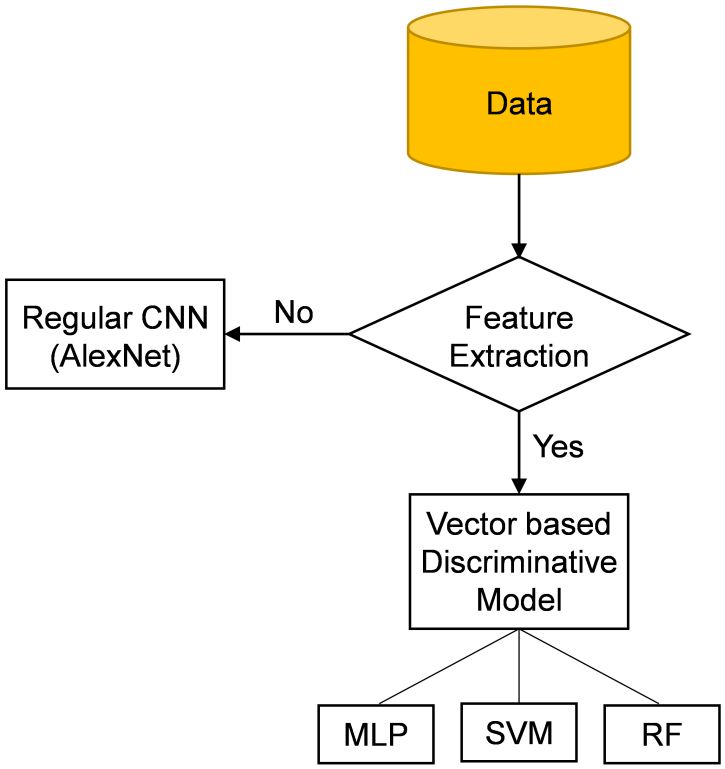


Figure 4.6: Two types of machine learning models.

We apply two different types of machine learning methods (Figure 4.6).

1. Regular CNN: only labeled data are used to train the model. In this approach, we use convolutional neural networks for classification.

2. Feature extraction and supervised learning: we first train an autoencoder using both labeled training dataset and unlabeled dataset to extract the middle layers as the features for the images. Using the extracted features and associated labels, a separate model is trained as the classifier, such as multi-layer perceptron (AE+NN), random forest (AE+RF) and SVM (AE+SVM, proposed in the literature [67, 123]). Among all the models, CNN generally gives the better performance. However, the reason might be that the training and testing data are from limited number of materials chemistry system, where the generalization benefit of autoencoder approach is not reflected in the testing result. Moreover, we only implemented the naive autoencoder for feature extraction. The architecture of autoencoder and the feature extraction strategies can be potentially further optimized toward better performance.

Chapter 5

Machine Learning for GISAXS: Thin Film Structure Identification

It is still challenging to obtain large-scale labeled experimental data for difficult tasks, such as the identification of the unit cell structure and orientation from GISAXS. In this chapter, I propose to generate the labeled datasets through simulation using an existing simulation toolkit, HipGISAXS. Using simulation data, the effect of physical parameters, such as instrumental noises and number of repeating unit cells, can be easily screened.

5.1 Introduction

Application of Machine Learning for Structure Identification

Properly classifying several structural classes automatically continues to be a challenge as the rate of data collection increases. Machine Learning (ML) has shown to be a valuable tool to handle such large data sets, as it has had success in the areas of regression analysis [132], image classification [133, 134], and optimization [135]. Recently, ML has been used to handle small data sets as well for situations where there is limited records to analyze [136, 137]. After ML's ability to handle both large and small datasets successfully, the scientific research community started to inspect both experimental and observational data in terms of ML. For example, these new processing capabilities have enabled the discovery of new chemical compounds by predicting the presence of certain species after chemical reactions [138].

Much of the data obtained from high brilliance lab sources and X-ray facilities is in the form of images [139], and computational methods such as Convolutional Neural Networks (CNN) bring new opportunities for image analysis and interpretation at the current data acquisition regimes [140]. One of the major hurdles of using CNNs is the dependence on labels that describes the acquired data, and to properly interpret data from high-throughput experiments using models based on simulated labeled sets. However, the use of CNNs to categorize and examine new data allows for the implementation of efficient code utilizing both CPUs and GPUs[141], which speeds up the data analysis, and the ability to scan large parameter spaces. An automated CNN based analysis approach would provide quicker turn-around time on image analysis that would otherwise take weeks for manual human categorization.

Data acquired at the Advanced Light Source (ALS), Lawrence Berkeley National Laboratory, comes from many different techniques, such as GISAXS and Grazing Incidence Wide Angle X-ray Scattering (GIWAXS). The data acquired from these experiments exhibit a variety of features, such as rings, peaks, arcs, and yoneda lines [142]. Crystalline lattices (i.e. Simple Cubic, Body-Centered Cubic, and Face-Centered Cubic) can be uniquely identified from scattering patterns. As the speed of this data collection has grown due to improvements in detector technology, new optics, and brighter sources, the necessity of an automated image processing program became evident.

Deep Learning on GISAXS Patterns

Deep learning was recently utilized to categorize features seen in X-ray data [143–146], as well as to maintain the state of the X-ray beam through adjusting the accelerator [147]. ML has also been applied in a variety of experiments, such as to categorize biomacromolecule solutions based on SAXS data [148], to separate and characterize mixed signals obtained from nanoscale X-ray experiments [149], to detect differences in lattices on the nano-scale based on diffraction images [150], and to categorize three-dimensional structures of nanoparticles based on X-ray absorption spectroscopy measurements [151]. However, a systematic study of various CNNs with varying

data quality has yet to be conducted.

To the best of our knowledge, the literature lacks investigations applying CNNs to categorize and predict different orientations of 3D nanoparticle lattices from materials characterized by GISAXS. Previous work proposed the use of shallow CNNs to categorize simulated GISAXS data based on crystal structure [139], followed by further investigation on reverse image search to categorize only four GISAXS patterns [152].

The Overview of this Chapter

In this chapter, I summarize the development of CNN-based classification schemes to categorize seven different 3D lattices and orientations of nanoparticles from X-ray data based on observable features in the scattering pattern. Training data is obtained using the HipGISAXS [153] scattering simulator. The scattering patterns were generated for four nanoparticle crystal lattices at varying orientations, crystal repetitions, and lattice parameters. First, we describe how training on various nanoparticle lattice orientations with various Miller Indices can allow for rapid automated analysis. Second, we show the robustness of the trained CNNs by drastically decreasing image quality and noise levels, and validate our CNN models by presenting successfully classification of materials using the most undesirable (low signal-to-noise ratio) datasets. Finally, we point out the future directions with a preliminary study of applying the trained model on experimental dataset to demonstrate its potential applications and possible future improvements.

5.2 Materials and Methods

Synthetic Data Production

To construct viable CNNs, we created a dataset spanning seven combinations of unit cells and orientations. HipGISAXS, a high-performance X-ray scattering simulator developed at the Advanced Light Source (ALS), Lawrence Berkeley National Laboratory (LBL), was used to generate the diverse collection of image samples. We simulated scattering patterns of various unit cells with different Miller Indices defining the crystal orientation relative to the substrate. We list the different combinations in Table 5.1, and schematics of the unit cells are illustrated in Figure 5.1. Figure 5.2 illustrates the experimental geometry.

Unit Cell	Miller Index
BCC	100
BCC	110
Simple Cubic	100
Simple Cubic	110
FCC	100
FCC	111
HCP	0001

Table 5.1: Different Unit Cells and Miller Indices for Classification.

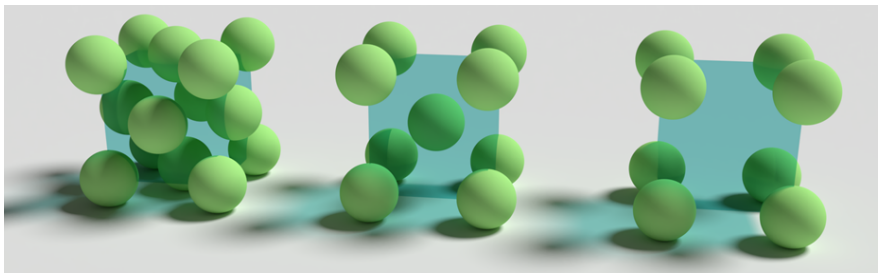


Figure 5.1: Schematics of FCC, BCC, and Simple Cubic unit cells.

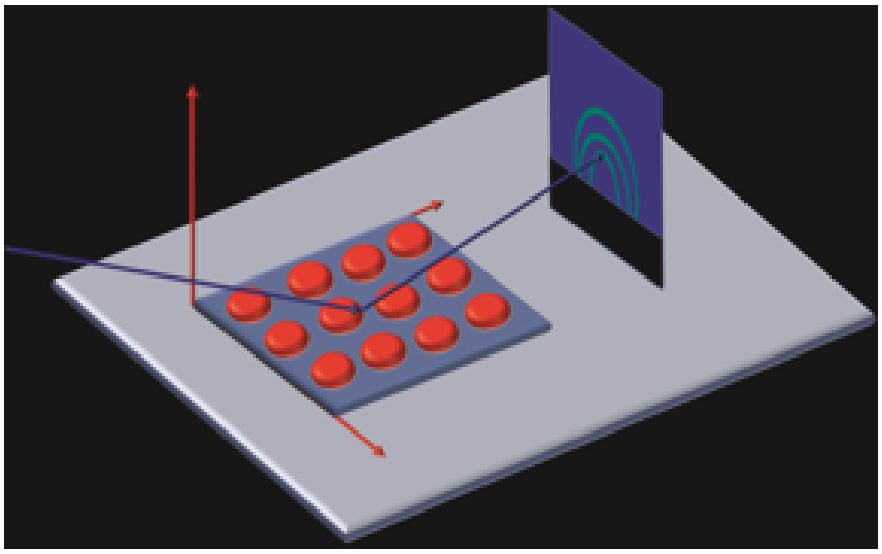


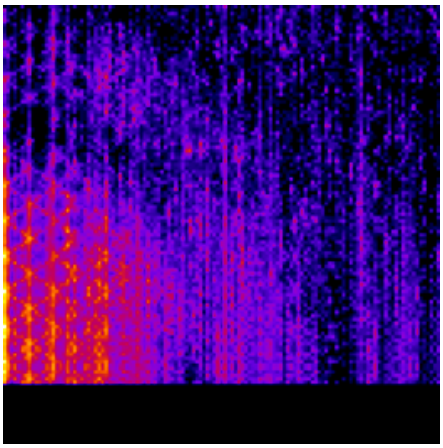
Figure 5.2: Diagram of experimental setup used in HipGISAXS. The incoming x-ray hits the substrate and scatters off the surface, hitting the detector. The collected image is a reciprocal space representation of the material. This diffraction pattern is simulated in HipGISAXS.

Simulation time took 500,000 core hours running on the NERSC super computer Edison at the Lawrence Berkeley National Laboratory. To assess the robustness of our classification model,

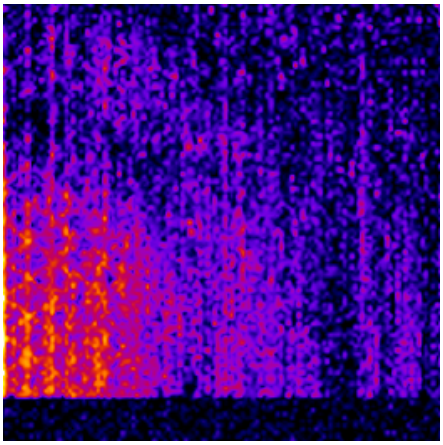
we explored a wide parameter space that influences the image quality. During the simulation, we varied experimental variables such as orientation, lattice parameter, and lattice repetition number (defining how many repeating unit cells occurred along each direction). To simulate possible errors in an experimental condition, we also incorporate different sources of noise:

1. Varying smear scales
2. Different Gaussian noise levels
3. Multiple resolutions to scale the image to smaller size and then scale back using interpolation

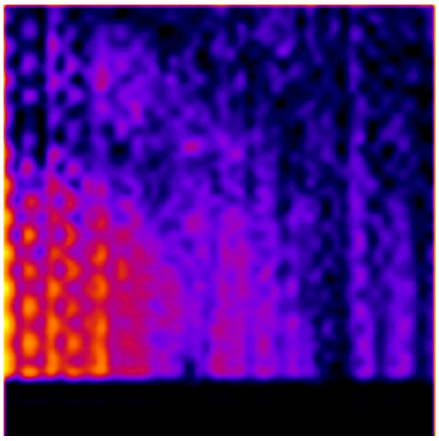
Smearing is used to mimic error in the experimental setup, e.g., X-ray optical element misalignment. The Gaussian noise addition aims to simulate the intensity fluctuations of the incidence beam, e.g., for an experiment performed under vacuum, this is the main source of noise. The rescale is used to simulate different image qualities obtained from possible experiments. We only applied one type of error to one specific data image. For example, the Gaussian or resolution error is only added to the 0 smear dataset to exclude multiplicative effects; the goal is to verify which noise type most affects the results of the CNN classification. Figure 5.3 shows an example of different error sources.



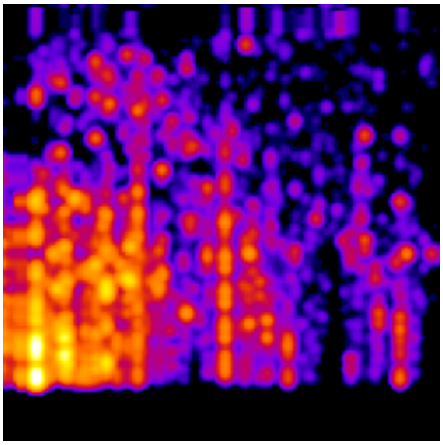
(a) Simulation data without smear and Gaussian noise



(b) With Gaussian noise



(c) With reduced resolution



(d) With smear noise

Figure 5.3: Examples of simulation data with different noise sources. The image resolution is 125×125. The vertical axis is the reflected beam $\vec{q}_{f_z}$ and the horizontal axis is $\vec{q}_{||}$.

Neural Network Models

Deep learning models have been applied to many classification tasks dependent on pictorial information, frequently using cross-entropy as a loss function. Cross-entropy is defined as the negative log-likelihood of the distribution calculated from the SoftMax function predicted from deep learning models. As CNNs provide non-linear models for complex image classification, we constructed a CNN based on the AlexNet architecture, motivated by AlexNet's previously reported accuracy and relatively low computational requirements. The CNN's input was adapted to accommodate the GISAXS image sizes, and we also adjusted the size of the convolutional layers accordingly. For each type of noise listed previously, we built a dedicated model. The training of CNNs are performed on Tesla P100 GPU server at ALS. Each model was trained over 20 epochs using Adam as optimizer.

Data Analysis Method

Within each noise level, we divide data into a training and testing dataset with the ratio 5 to 1. We stratify the data under different x and y repetition numbers accordingly to ensure that the ratio of training and evaluation data are consistent over different x and y repetition numbers.

5.3 Results on Simulation Dataset

We adjusted input size and network architecture to accommodate the simulated GISAXS patterns. The training and testing accuracy on the simulation dataset without any artificial noise is higher than 98%. Training and prediction accuracy denotes the accuracy achieved for training and testing datasets, respectively.

Classification with Different Noise Sources: Under Ideal Condition

As observed in routine experiments, GISAXS images rarely have sharp features. This can be attributed to multiple factors, such as short-range order in soft matter, imperfections in collimation, non-monochromaticity of the X-ray beam, and different types of read-out noises in the detectors [32]. Mathematically, this effect can be approximated by smearing the images and adding Gaussian noise. The smear effect can be formulated as

$$I_s(q_x, q_y) = \frac{1}{2\pi\sigma_x\sigma_y} \int_{-\infty}^{\infty} dx\, W(x; 0, \sigma_x) \int_{-\infty}^{\infty} dy\, W(y; 0, \sigma_y)\, I(q_x + x, q_y + y) \tag{5.1}$$

I_s and I are the intensity of beam with and without smearing effect, respectively. W is a zero-centered Gaussian distribution. The standard deviations σ_x and σ_y are defined as the smear scale. In our study, the standard deviation of the x and y directions are the same.

We smeared the image with different scales to investigate how the smear artifacts affected the prediction performance and the results are shown in Table 5.2. When the smear scale is small (0.5 px), the smear does not affect the testing accuracy. When the smear scale increases to 1 px, both training and testing accuracy decrease. The testing accuracy dropped from 98.12% to 97.48%.

Smear Scale (px)	Training Accuracy	Testing Accuracy
0	98.57%	98.12%
0.5	98.52%	98.12%
1.0	98.18%	97.48%

Table 5.2: Prediction accuracy under different smear scales.

Poisson shot noise can be translated into pixel intensity variations due to the particle distribution in the beam during experiments. Moreover, there are some other noise sources which can be modeled by a Gaussian distribution of noise. We added Gaussian noise (with different scale

multiplier α) to each pixel and investigated how this error affected the training and testing of the CNN. As expected, testing accuracy decreased with the increased amount of Gaussian noise. When $\alpha = 0.2$, the testing accuracy dropped to 97.30%. At the highest noise level, a CNN was trained to identify 88.27% of the images successfully. All the results are tabularized in Table 5.3.

Noise	Training Accuracy	Testing Accuracy
No Noise	98.57%	98.12%
Gaussian Noise (α=0.1)	98.90%	98.09%
Gaussian Noise (α=0.2)	98.82%	97.30%
Gaussian Noise (α=0.5)	98.85%	94.20%
Gaussian Noise (α=1.0)	96.54%	88.27%

Table 5.3: Prediction accuracy under different pixel-wise noises.

In some cases, the resolution of GISAXS images is limited by experimental instruments, such as the pixel size of the detector. To understand the performance of deep learning models under different image resolutions, we scaled the images to a lower resolution and then rescaled it back to 125×125 using interpolation to be able to insert the image into the CNN. The training and testing accuracies are summarized in Table 5.4. The testing accuracy was not compromised if when the image was scaled down to 50×50 resolution. This indicates that the structural information in the scattering pattern was still retained even though we rescaled the image to 4/25 of the original area. We propose that this ability to handle data reduction was due to the quality of the simulation results, and may differ when applied to real experimental data. Moreover, this effect is dependent on the classification task.

Resolution	Training Accuracy	Testing Accuracy
125×125	98.57%	98.12%
100×100	98.57%	98.08%
75×75	98.56%	98.08%
50×50	98.40%	97.95%

Table 5.4: Prediction accuracy under different resolutions.

More General Results: Testing Error using a Single Model

In the previous sections, we highlighted the training and testing of various CNNs with a variety of different noises. An important note is that the results were obtained by exposing both the training and testing datasets to certain noises. In practice, however, we need a single model to test all the data under different conditions. To examine this case, we took this into account by taking various instances of our data that was subjected to high noise scales and lower resolutions and had them classified by the CNN that was trained with the data not subjected to any noise. The results are highlighted in Table 5.5.

Noise	Testing Accuracy
Smear Scale = 1 px	84.53%
Gaussian Noise Scale α = 1	62.37%
Resolution = 100×100 px	97.39%
Resolution = 50×50 px	94.63%

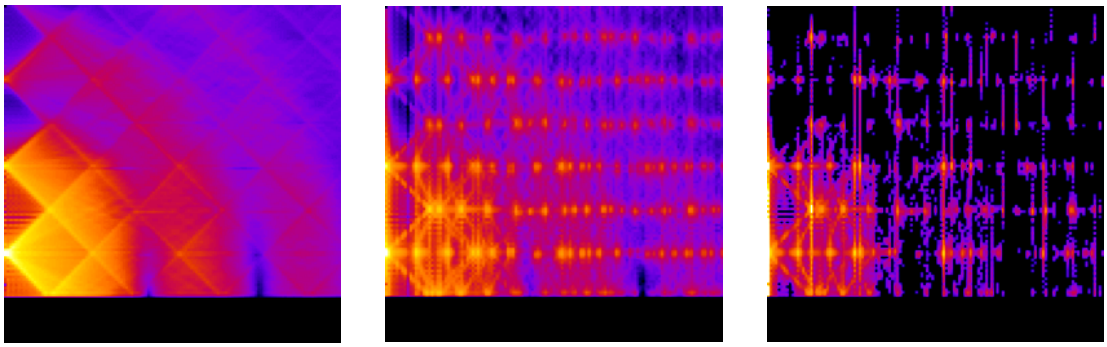
Table 5.5: Prediction accuracy using model trained by data without noise.

First was the dataset subjected to the maximum amount of smear. For reference, we were comparing these new values to the accuracy on data with no noise, which was 98.12%. For the highest smear scale used, the accuracy dropped to 84.53%. Next, we classified the Gaussian noise data. The highest Gaussian noise resulted in a lower accuracy of 62.37%. We also applied the same CNN to the dataset that has been reduced to lower resolution. When the image was lowered to a 100×100 px image resolution, the accuracy dropped to 97.39%, followed by a 94.63% accuracy for a 50×50 px resolution.

Classification Under Different Repetition Scales

The challenge in identifying the crystal structure of materials with a limited number of repeating units in x and y direction lies on the fact that the scattering signal, especially higher order peak, is weakened. Examples of GISAXS images with repetition scales 1, 100, and 1000 are shown in Figure 5.4; this image emphasized the impact of a larger number of repetitions. To investigate the classification variation on a limited number of repetitions, we designed an experiment to evaluate the testing accuracy under different scenarios.

Figure 5.5 shows the testing accuracy under different repetition scales on both the x and y direction. An important note is that, for this report, the repetition values were the same for both directions (1 repetition in both x and y, 10 repetitions for both x and y, etc.), which is a subset of the whole dataset. The experiment was conducted to test for the lower limit of accuracy for the CNN model. As expected, the accuracy improved with more repetition due to an increase in the signal strength of the diffraction image in Figure 5.4.



(a) Repetition 1 (b) Repetition 10 (c) Repetition 100

Figure 5.4: GISAXS simulation image under different repetition numbers: (a) has one repetition of the unit crystal, (b) has 10 repetitions, and (c) has 100 repetitions.

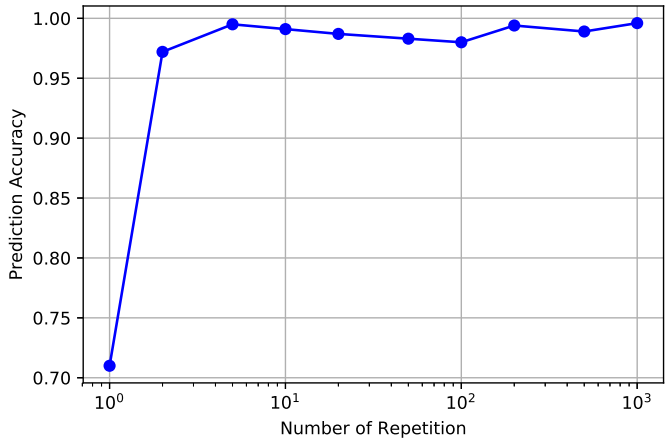


Figure 5.5: Testing accuracy under different repetition numbers for equal x and y repetitions. A sharp increase in testing accuracy is quickly obtained for increased repetitions of the unit cell.

5.4 Discussion

Our initial hypothesis was that materials with diverse orientations of nanoparticle lattices could successfully be identified via ML algorithms. Using synthetic data generated via HipGISAXS, we developed a trained model by first testing various different CNN architectures. We further tested this hypothesis by modeling artifacts over the data through increasing noise addition and varying resolutions, which for example, may change the appearance of features which distinguish one class from another. In this section, we provide several analysis to further understand these classification result.

Understanding the Learning Mechanisms

We first studied which unit cell the model had the most trouble successfully classifying by assembling the confusion matrix [154], as shown in Table 5.6. The element $M_{i,j}$ is calculated as the ratio of the number of samples with label j classified as i over the number of samples with label j. The diagonal of the confusion matrix highlights the accuracy of the model. This indicates that there was high confidence in the model for each unit cell. However, we noticed situations for which the model exhibits inaccurate classification. Particularly, there was some confusion between the Simple Cubic 110 and BCC 110 labels.

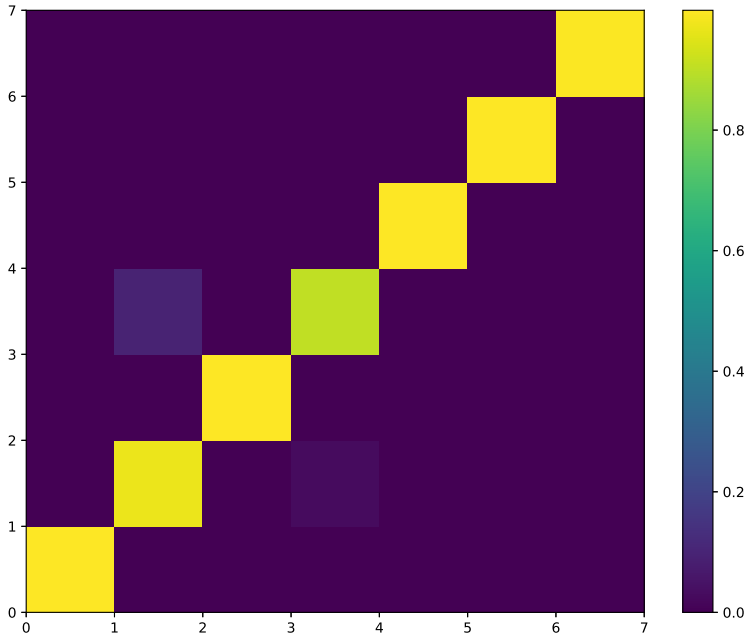


Figure 5.6: Visualization of the confusion matrix.

Visualization of Neural Network Results

To understand the operations in convolutional neural network, we visualize some of the convolution filters in Figure 5.8 in supplementary information. We also visualize some of the results after the convolution operators in Figure 5.9 in supplementary information. In order to visualize a more compact representation of the data, we performed Principle Component Analysis (PCA) [155] of the classified data from the last fully connected layer of the CNN. This allows us to map features to a lower dimensionality space; the two most significant components, PC1 and PC2, as illustrated in Figure 6.5. Mathematically, we can measure the information carried in principle components using

explained variance ratio. The explained variance ratio of the top k components from d dimensional data is defined as

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{d} \sigma_i^2} \tag{5.2}$$

where σ_i is the singular value of ith component. From the visualization of PCA, we observe that each class has been clustered in the last fully connected layer. PCA is a powerful tool for data visualization and analysis. However, it also has limitations. The first two components we extracted only contribute to 65% of explained variance [156], therefore some classes that seem to be overlapping might be well separated in a higher dimensional space. Moreover, as the non-convexity of the neural network, we observe some variations of the PCA results in different models, possibly due to the convergence to different local minimum.

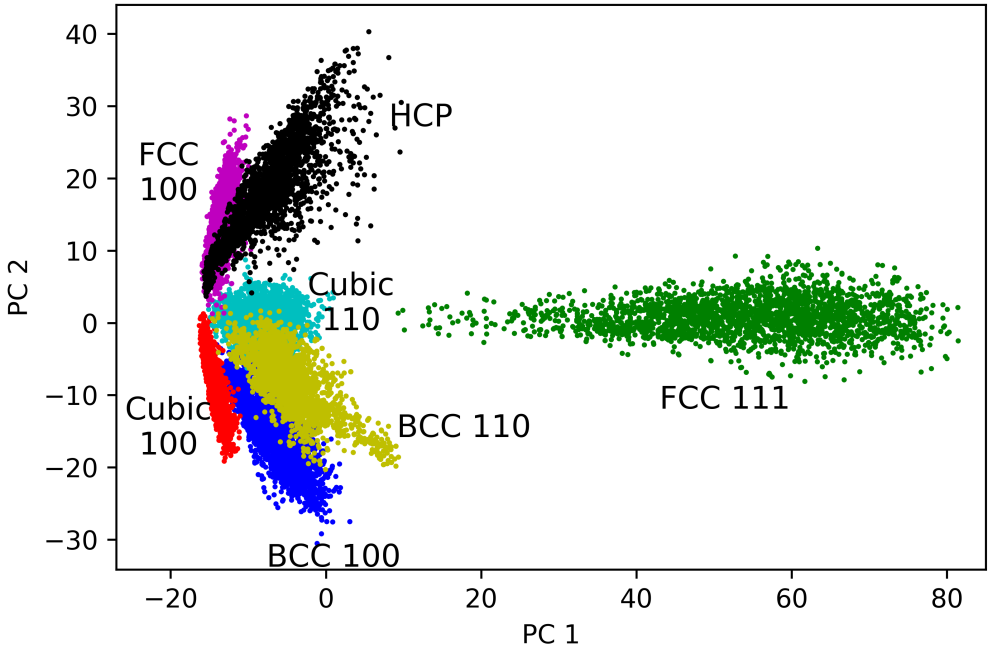


Figure 5.7: Scatter plot for different X-ray scattering patterns in terms of two most significant PCA components.

5.5 Conclusions

In conclusion, we have successfully built an image classification scheme using CNNs for the purpose of classifying various 3D nanoparticle lattice and orientations based on simulated X-ray data with different noise sources. Our CNN was trained on large amount of simulated images

acquired from the use of HipGISAXS. The training accuracy achieved on the simulated data was over 94%. To study the robustness of the model, various changes were made to the simulated GISAXS data such as decreased resolution and by adding noise to the data. When introducing the data to increased smear and Gaussian noise, the accuracy dropped, at its lowest to 64%. The use of this methodology will highly impact the X-ray and neutron science communities by speeding up GISAXS data analysis of new materials at future data regimes.

5.6 Future Outlooks

We applied the CNN to a set of real experimental X-ray patterns of nanocrystal superlattices and the accuracy was about 50% with a limited amount of data (about 30 images labeled by collaborators), which is significantly lower than the testing accuracy we can achieve in the simulation dataset (based on the ideal cases with synthetic noises). Hence, the current predictive models for the real data still requires improvements to be used, for example, to tackle pattern details associated to crystal imperfections. One of the possible direction is to include more diverse q and incidence angle ranges to augment the training dataset. Another possible direction is to further increase the complexity of the GISAXS model such as different space groups, crystal sizes and defects in materials, to simulate high complex morphologies and therefore expand the samples. Further developments will focus on crystals formed from various materials, different nanoparticle size, and the variation of the form factor by the addition of cylinders. The proposed CNN scheme is very flexible, and it could be potentially extended to other materials by constructing simulation data using HipGISAXS and/or by exploiting labeled experimental data.

5.7 Acknowledgement

We acknowledge Ye, Xingchen and Zhu, Chenhui and Ercius, Peter and Raja, Shilpa N. and He, Bo and Jones, Matthew R. and Hauwiller, Matthew R. and Liu, Yi and Xu, Ting and Alivisatos, Paul for allowing us to use their data for our real experiment analysis. This work was supported by the Center of Advanced Mathematics for Energy Research Applications (CAMERA) through the Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Early Career Program. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

5.8 Supplementary Information

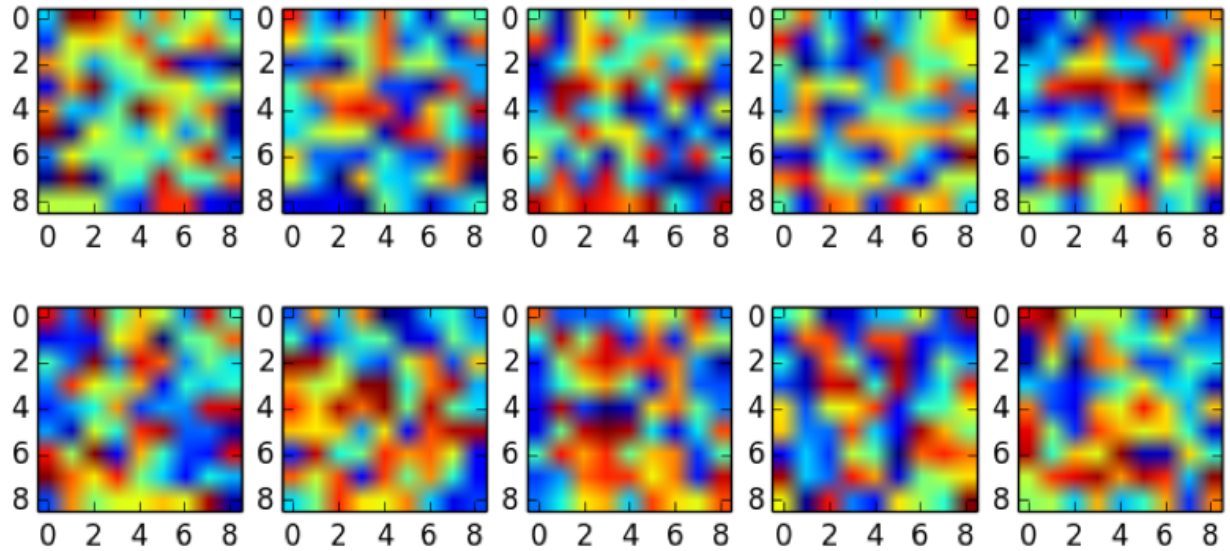


Figure 5.8: Visualization of Filters in Different Layers of Trained Alex-Net.

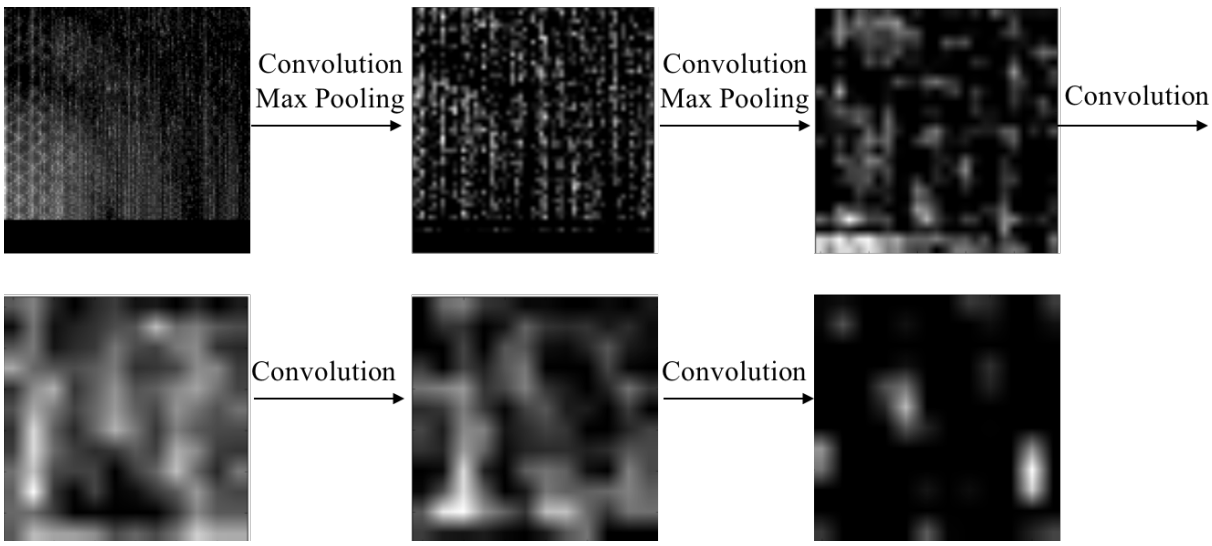


Figure 5.9: Visualization of convolution operations in trained Alex-Net.

Part III

Data Driven Approaches for NMR Crystallography and Characterization Facility Optimization

Chapter 6

Data Driven Approach for NMR Crystallography: Chemical Shift Prediction

In this chapter, I present a data-driven framework for NMR crystallography characterization and its application on materials chemistry involving the solid-state crystals. I have developed a deep learning method for chemical shift prediction for atoms in molecular crystals that utilizes an atom-centered Gaussian density model for the 3D data representation of a molecule. We define multiple channels that describe different spatial resolutions for each atom type that utilizes cropping, pooling, and concatenation to create a multi-resolution 3D-DenseNet architecture (MR-3D-DenseNet). Because the training and testing time scale linearly with the number of samples, the MR-3D-DenseNet can exploit data augmentation that takes into account the property of rotational invariance of the chemical shifts, thereby also increasing the size of the training dataset by an order of magnitude without additional cost. Good agreement between machine learning predictions and DFT calculations are obtained for ^{13}C, ^{15}N, and ^{17}O chemical shifts, with the highest accuracy found for ^{1}H chemical shifts.

This chapter is adapted with permission from Liu et al., "A Multi-Resolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography", J. Phys. Chem. Lett., 2019, 4558. [157], Copyright 2019 American Chemical Society.

6.1 Introduction

Nuclear magnetic resonance (NMR) crystallography is an experimental technique to determine the structure of complex materials [158, 159], biomolecules such as proteins [160, 161], as well as small molecules and pharmaceuticals [56, 57, 162] in the solid state. In practice, NMR crystallography is a structural model building procedure that depends on a number of NMR data types, of which chemical shifts in particular play a prominent role. A strength of NMR chemical shift data is its excellent sensitivity to hydrogen [47] which, given the importance of hydrogen-bonding in most molecular systems, makes it very complementary to X-ray diffraction techniques.

In the case where little is known about the chemical bonding of an unknown structure, the experimental measurements for chemical shifts are compared to the results of *ab initio* methods based on density functional theory (DFT), typically using Gauge-Including Projector-Augmented Waves (GIPAW) methods [163]. However, because of the cubic computational complexity scaling with the number of atoms ($O(N^3)$), alternative methods are being actively investigated to mitigate its large computational cost, especially for large systems. Many of these more inexpensive approaches are focused on fragment models that incorporate the long-range many-body polarization effects of the lattice environment via electrostatic embedding, such as the self-consistent reproduction of the Madelung potential (SCRMP) [164], which has yielded very high quality results when combined with hybrid DFT functionals.

An alternative approach is to apply machine learning methods to predict the experimental and/or DFT results for systems ranging from proteins in solution [165–169] to solid-state materials [158, 159, 170]. Cuny et al. reported a fully connected shallow neural network to predict the quadrupolar couplings and chemical shifts in silica materials for ^{17}O and ^{29}Si using symmetric functions of the Cartesian coordinates as the input [171]. Paruzzo et al. applied the kernel ridge regression (KRR) using a smooth overlap of atomic positions (SOAP) kernel, that also directly incorporates rotational invariance of the chemical shift value to applied magnetic field, for molecular crystal systems [19]. However, the KRR approach requires $O(N^2)$ complexity for calculating the similarity kernel matrix, and quadratic-to-cubic complexity for kernel matrix inversion, which is ultimately not tenable for large training and testing datasets.

Convolutional neural networks (CNNs) have been applied to several problems in chemistry and biology, such as enzyme classification [172], molecular representation [173], amino acid environment similarity [174], and potential energy prediction [175]. They have not to the best of our knowledge been applied to NMR crystallography property prediction. There are a number of deep network variants that have been developed to address important deficiencies of a vanilla CNN, which are hard to train because of the vanishing (or exploding) gradient problem. This is because the repeated application of non-linear activation functions cause later outputs in the deep layers to flatten out, and back-propagated gradients are then diminished.

Residual networks (ResNets) were developed to precondition the network to learn the residual

of a non-linear mapping by referencing it to an identity mapping, which is easier to train due to the presence in the network architecture of "identity shortcut connections".[176] Because these network connections skip layers, there is more direct information flow from the loss function to correct the weight parameters of earlier layers. DenseNets build on these ideas by also utilizing skipped connections for better gradient flow, while at the same time also performing concatenation of feature maps that permits greater propagation and reuse of features in what is termed "deeper supervised learning".[177]

Here we report a deep learning approach to predict chemical shifts in the solid-state for hydrogen (^{1}H), carbon (^{13}C), nitrogen (^{15}N) and oxygen (^{17}O) that outperforms KRR while also allowing for chemical interpretation of the results using principal component analysis (PCA). The deep learning approach is based on a multi-resolution (MR) spatial data representation, where each resolution level and atom type is formulated as an independent channel of a deep learning 3D-DenseNet architecture, augmented with special concatenation of pooling layers (at reduced resolution) with cropping of feature maps (retaining high resolution features with reduced size) of the transformed spatial data. The resulting MR-3D-DenseNet removes the restrictions imposed by KRR, i.e. the need to build in rotational invariance of chemical shifts as well as the limitations to small data sets [19], in order to take advantage of a data augmentation procedure in which we rotate the chemical environment for each atom in a sample, thereby increasing the data set size by close to an order of magnitude with little computational expense.

Using the greater capacity of the MR-3D-DenseNet deep network, we obtain significant improvements for ^{13}C, ^{15}N, and ^{17}O chemical shifts over KRR, with excellent agreement for ^{1}H chemical shift prediction with RMSE error of 0.37 ppm, which is the same level of error between *ab initio* calculations and experimental measurements. The PCA allows us to both understand these improvements, as well as interpreting the remaining deficiencies, for chemical shift prediction for all atom types. Based on our PCA analysis and the prediction performance compared to *ab initio* calculation, we emphasize the importance of size and variety of training samples. Given the far better computational scaling of the multi-resolution 3D-DenseNet, we can afford to address this deficiency with much larger data sets than currently available in future studies.

6.2 Data Representation

The molecular crystal structures are from the Cambridge Structural Database (CSD) [178], comprising 2,000 crystal structures in the training dataset and 500 crystal structures in the testing dataset. The coordinates of atoms in the unit cell, and the corresponding calculated chemical shieldings, are as given in the reported literature by Paruzzo and co-workers [19]. In that paper, the training data was generated by conducting farthest data sampling to yield 2,000 crystal structures from which to extract training examples, whereas for testing they utilized uniform sampling to yield 500 crystal structures from which to extract testing examples, although there are 61,000 available structures in the CSD. This is reasonable given the cost of the underlying DFT calculations for

chemical shift values, and also because of limitations due to the unfavorable scaling of the KRR which we discuss later. This resulted in the number of unaugmented 3D samples for training and testing for each of the atom types as given in Table 6.1. No further data selection or cleaning procedures are applied to the original dataset, except that 0.05% outliers (chemical shielding < 0 or > 40) in the ^{1}H-NMR training dataset were removed.

Atom Type	Number of Samples			
	Training Dataset		Testing Dataset	
	w/o Augmentation	w/ Augmentation	w/o Augmentation	w/ Augmentation
^{1}H	76,174	609,392	29,913	239,304
^{13}C	58,148	465,184	26,607	212,856
^{15}N	27,814	222,512	2,713	21,704
^{17}O	25,924	207,392	5,404	43,232

Table 6.1: The number of samples in training and testing datasets with and without data augmentation.

Given the limited number of examples in the training dataset, we apply a physically motivated data augmentation method to improve the prediction performance of the MR-3D-DenseNet model. Since the chemical shift is invariant under rotational operations, we augment the data by rotating the Cartesian coordinates of atoms randomly with the Euler angles uniformly distributed between $[-\frac{\pi}{2}, \frac{\pi}{2}]$ along each of x, y and z axis. During the training phase, both the original data and augmented data are included in the training dataset. During the testing phase, we average the prediction results among 8 different rotation configurations. The final number of training and testing examples after this augmentation are given in Table 6.1.

The input data representation to the MR-3D-DenseNet assumes that chemical shifts are sensitive to the electron density distribution of atoms in molecules. Hence a molecule is represented on a 3D grid in which each atom takes on a radial Gaussian density. The 3D image is a bounded box with $16 \times 16 \times 16$ voxels, with the density $D(\boldsymbol{r})$ at each voxel taken as a sum of Gaussian distributions from all of the atoms

$$D(\boldsymbol{r}) = \sum_{\boldsymbol{r}' \in A} \exp(-\frac{||\boldsymbol{r} - \boldsymbol{r}'||^2}{\sigma^2}) \tag{6.1}$$

where the summation runs over atoms of a given atom type A and the $\boldsymbol{r}'$ are the corresponding atomic centers. The coordinate $\boldsymbol{r} = (x, y, z)$ at the center of voxel (with index (i, j, k)) is calculated as

$$\boldsymbol{r} = (x, y, z) = (\frac{(i - \frac{15}{2})d}{15}, \frac{(j - \frac{15}{2})d}{15}, \frac{(k - \frac{15}{2})d}{15}) \tag{6.2}$$

where d is the grid resolution. Unlike the Gaussian smearing method reported in literature [173], we calculate the density at the center of the voxel numerically using 16-bit floating point numbers. We

also considered additional electron density representations including Slater orbitals and calculated from the inverse Fourier transform of the atomic form factor, but found that they performed worse than the Gaussian representation that can be explained by their heavy tails (see Supplementary Information).

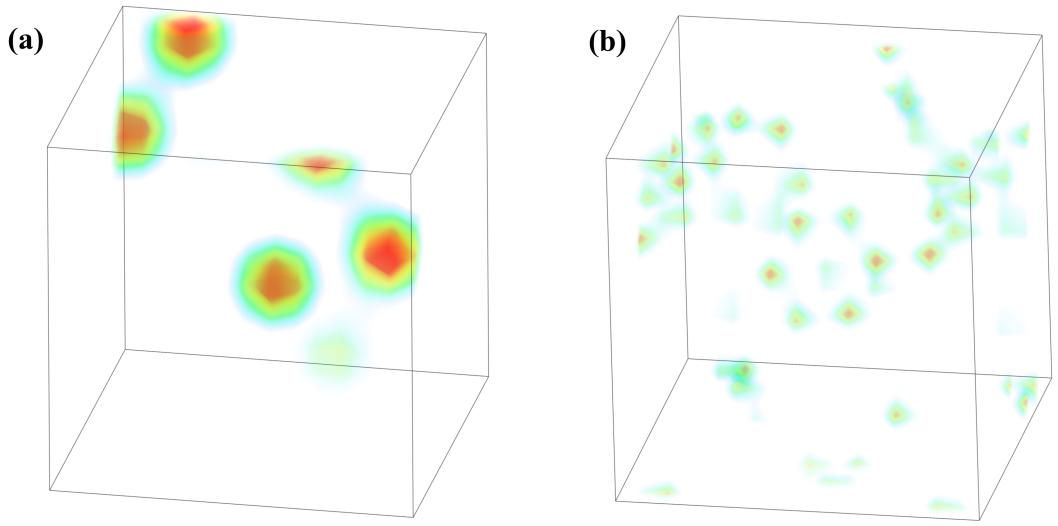


Figure 6.1: Visualization of the Gaussian densities of atoms on different grid sizes. Representative example is shown for carbon channels on (a) 4 Å and (b) 10 Å grid. The densities are visualized through Mayavi package [179].

The atom whose chemical shift is being evaluated is placed at the center of the 3D grid, and its chemical environment is represented by calculating the density under different grid sizes, where d = 4 Å, 6 Å, 8 Å, 10 Å, and 14 Å, each of which is represented by a dedicated channel in the MR-3D-DenseNet model. Under each grid size, we divide the density based on the atom types into 4 different channels for ^{1}H, ^{13}C, ^{15}N, ^{17}O, respectively, resulting in a total of 20 separate channels in the MR-3D-DenseNet network. Figure 6.1 shows a visualization of the carbon channels of the molecule (Z)-2-Hydroxy-3,3',4'-trimethoxystilbene (reported by Stomberg et al. [180]) at two different grid size resolutions.

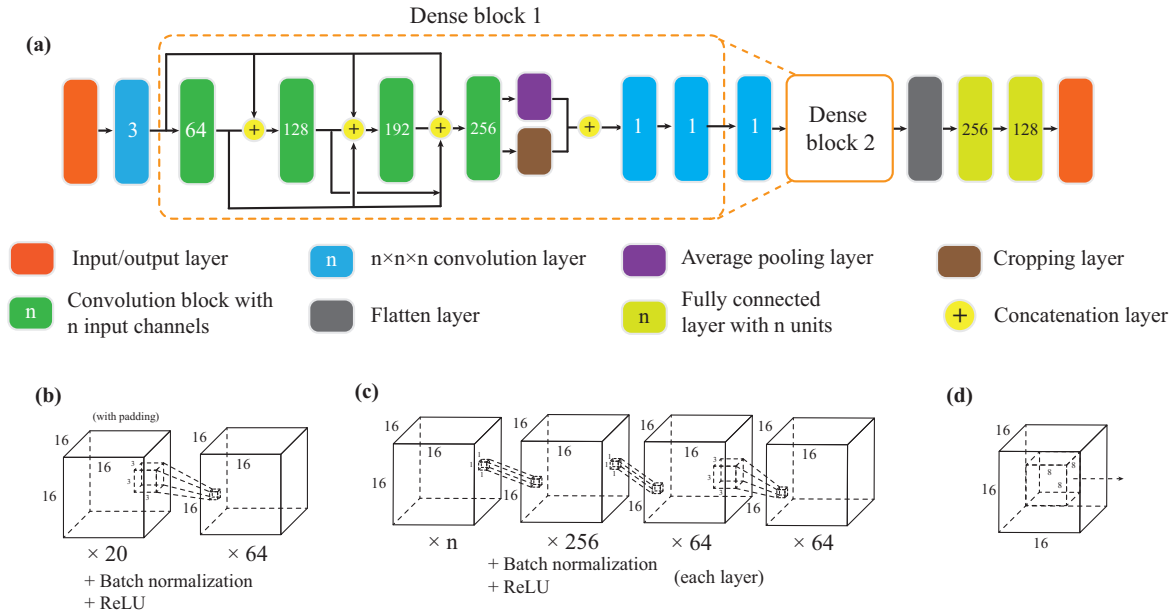


Figure 6.2: Illustration of the overall architecture of the MR-3D-DenseNet model. (a) Flowchart of the network (b) Illustration of $3\times3\times3$ convolution layer prior to the first dense block (c) Illustration of the repeating unit in DenseNet block that contains two $1\times1\times1$ convolution layers followed by a $3\times3\times3$ convolution layer (d) Illustration of the cropping layer from the center of the feature map.

6.3 Machine Learning Models

In this study, we start from a variant of the DenseNet architecture that omits skip connections going directly to the Dense block output. In addition, we designed a modification to the DenseNet that is motivated by the hypothesis that the importance of a given voxel increases as the distance between it and the investigated atom decreases, which is represented by multi-resolution channels. A schematic of the MR-3D-DenseNet architecture is shown in Figure 6.2, and is comprised of a regular $3\times3\times3$ convolutional layer followed by two DenseNet blocks with a $1\times1\times1$ transition convolutional layer in between them. The flattened output from the last DenseNet block is then fully connected to a layer with 256 units which is fully connected to a 128 unit layer, which is then fully connected to the output layer.

Each DenseNet block has four repeating units: each repeating unit has two $1\times1\times1$ bottleneck convolutional layers with 256 and 64 channels followed by a $3\times3\times3$ convolution layer with 64 channels. The MR-3D-DenseNet utilizes cropping and pooling such that at the end of each block, we concatenate the $2\times2\times2$ average pooling layer and the cropping of the center segment of the feature map with the same size ($\frac{l}{2}$, where l is the current feature map size). This retains low and high resolution features throughout the deep layers. After the concatenation of pooling and cropping, there are two $1\times1\times1$ convolutional layers with 256 and 64 channels, respectively, to

Atom Type	MR-3D-DenseNet	R^2
H	0.37 (24%)	0.9856
C	3.3 (23%)	0.9957
N	10.2 (23%)	0.9916
O	15.3 (14%)	0.9933

Table 6.2: Testing RMSEs (ppm) using MR-3D-DenseNet. We also report the improvement of RMSE in percentage compared to KRR [19] and the R^2 values using MR-3D-DenseNet.

process the information and retain the channel size to 64 before entering the next block. Using this network architecture, we describe the detailed training protocol and hyperparameters in Methods.

6.4 Result and Discussion

The performance on chemical shift predictions for all atoms using MR-3D-DenseNet compared to KRR is summarized in Table 6.2. The testing RMSEs of chemical shifts for ^{1}H, ^{13}C, ^{15}N, and ^{17}O using the MR-3D-DenseNet architecture is found to be 0.37 ppm, 3.3 ppm, 10.2 ppm and 15.3 ppm, which are 24%, 23%, 23% and 14% lower than the RMSEs given by a KRR method [19]. Among the four atom types, the error of ^{1}H prediction is comparable to the error between high standard *ab initio* calculations and experiment, i.e. GIPAW/PBE and SCRMP/PBE0 which have chemical shift accuracy RMSEs of 0.43 ppm and 0.33 ppm, respectively.[164, 181] Although the predictions on the other atom types are very good, we attribute their lessened performance with respect to *ab initio* models as a lack of unique data compared to that available for ^{1}H (Table 6.1), a point to which we return to later.

In a separate publication, we will present a full study of different deep learning architectures, but here we contrast the best MR-3D-DenseNet model to the KRR machine learning method for which results are available on the same chemical shift problem [19]. We can attribute the success of the MR-3D-DenseNet approach based on three factors: (1) the greater flexibility in input representation of individual atom types and spatial resolution, and the advantages of concatenation of the pooling and cropping operations in the architecture, (2) the dependence on the size and quality of the training set, and (3) the ability to learn chemical bonding features, all of which are unique to chemical shift prediction using the MR-3D-DenseNet architecture.

In regards the first point we decompose the MR-3D-DenseNet result based on its multi-resolution input representation with no special concatenation of pooling and cropping operations (MR-NoConcat) versus the network architecture that utilizes concatenation of pooling and cropping but takes in only a single resolution input representation (SR-Concat). It is evident that the input and architecture features trained in isolation of each other offer significant improvements in performance over KRR, with further benefit being realized by their combined used in MR-3D-DenseNet

Atom Type	KRR	SR-Concat	MR-NoConcat	3D-MR-DenseNet	std
H	0.49	0.38 (10 Å)	0.38	0.37	< 0.01
C	4.3	3.5 (6 Å)	3.5	3.3	< 0.1
N	13.3	10.2 (8 Å)	10.3	10.2	0.2
O	17.7	16.3 (6 Å)	15.6	15.3	0.5

Table 6.3: Testing RMSEs (ppm) for KRR and using different features of the MR-3D-DenseNet model for each atom type: SR-Concat, MR-NoConcat, and MR-3D-DenseNet. For the single-resolution input, the SR-Concat model is sensitive to the grid size for a given atom type and an optimized value must be determined (parentheses).

(Table 6.3).

The main limitation of the SR-Concat model is that the 3D-grid size of the single resolution input depends on the atom under consideration, and even when the grid size is optimized it gives only limited improvement for the oxygen chemical shift. Hence the multi-resolution input representation is undoubtedly very important, and the novel feature of our approach is its greater flexibility compared to KRR which must determine the weights for mixed-scale kernels that use different cut-off sizes [19]. Although the concatenation of pooling and cropping operations appears to play a more limited role for some of the atom types (MR-NoConcat), it does improve the prediction for carbon as judged by the standard deviation.

Furthermore, the success of a deep learning network model will be highly dependent on the size, variety and quality of the training dataset. To understand the effect of the training data size, we examine the ^{1}H chemical shift testing RMSE for KRR and the MR-3D-DenseNet model as a function of increasing number of training examples (Figure 6.3a). Without data augmentation, the prediction performance of MR-3D-DenseNet improves over KRR after being presented $\sim$ 1500 training points. However, the MR-3D-DenseNet has the capacity to exploit the augmented data to outperform the KRR model even with only a hundred training examples.

Although it might be argued that the KRR model has no need for the augmented data, since rotational invariance is built directly into the kernel, the data augmentation is clearly doing something more than invoking the rotational invariance feature of the chemical shift (i.e. the performance would be the same otherwise). In addition, augmentation of the testing dataset can be seen as equivalent to an ensemble averaging prediction without the need to retrain many networks to realize the same benefit, lowering the testing RMSE further to realize the best MR-3D-DenseNet performance (Figure 6.3a).

KRR has unfavorable computational scaling for kernel matrix computation and kernel matrix inversion, which limits it capability to exploit data augmentation. By contrast the training time

for the MR-3D-DenseNet model scales linearly with the number of training samples (Figure 6.3b). More importantly, the prediction time of MR-3D-DenseNet with a trained model does not scale with the number of training examples, whereas the testing time for KRR scales linearly because the similarity kernel has to be calculated using all of the training samples.

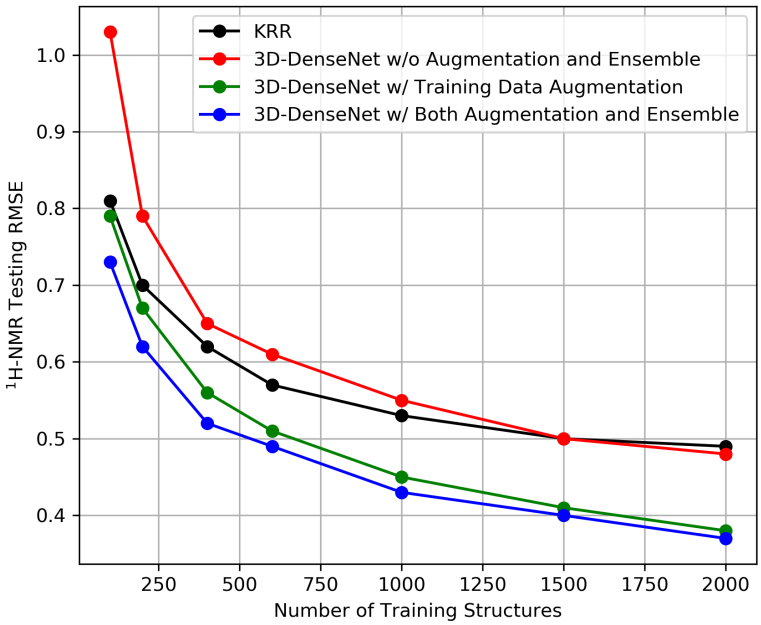


Figure 6.3: Testing RMSEs and timings for ^{1}H chemical shift for different numbers of samples using the MR-3D-DenseNet. (a) using no augmentation (red), with training dataset 8-fold augmentation (green), using both training and testing dataset with 8-fold augmentation (blue), and compared to the testing error reported previously for KRR on the same dataset [19] (black). The models are trained under the same number of batches to obtain a fair comparison; for example, when the data is augmented by 8-fold, the number of training epochs decrease to 1/8. (b) Training (8-fold) time of MR-3D-DenseNet model for the ^{1}H chemical shift under the same network architecture and number of epochs. The testing time (1-fold) of ^{1}H chemical shift is about 4-5 minutes for 500 preprocessed testing structures and is independent on the number of training structures. The training and testing time are benchmarked on Nvidia Tesla P100 GPU.

In totality, the MR-3D-DenseNet architecture with data augmentation yields a much tighter prediction error across the unique data across all atom types relative to KRR as seen in Figure 6.4. We found that further increasing the data augmentation to 16-fold rotations or adding the effects of small vibrational smearing of atom positions had a neutral effect on the prediction performance. Instead Figure 6.4 emphasizes that creating more unique data for the heavy atoms will certainly

improve the MR-3D-DenseNet performance relative to *ab initio* models, as the number of heavy atom samples are limited compared to ^{1}H samples in the current dataset. This may also explain why the prediction performance was limited for ^{15}N and ^{17}O when compared to MR-NoConcat (Table 3) because there was insufficient data to exploit the deep network architecture design.

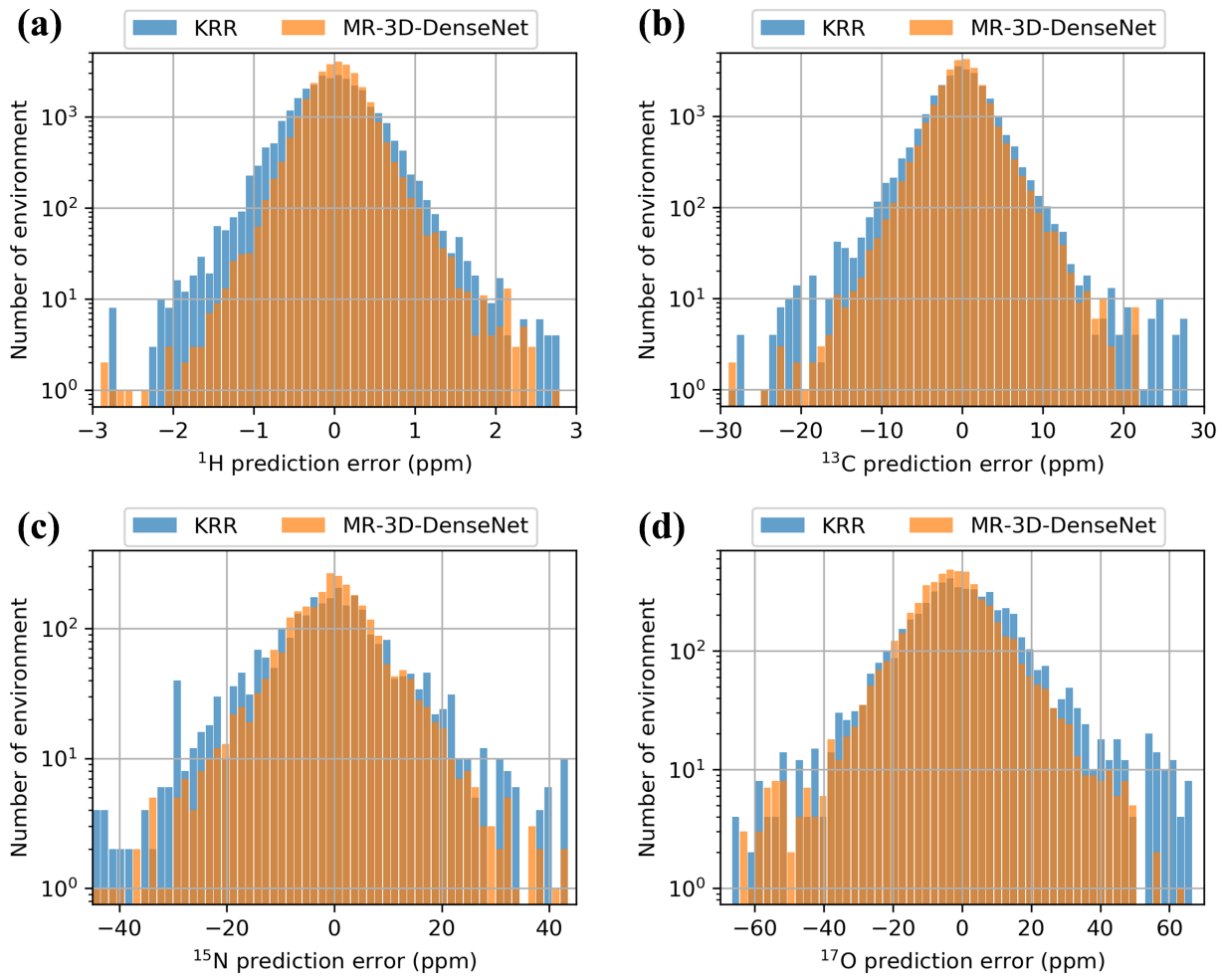


Figure 6.4: Histogram of testing error distribution comparing MR-3D-DenseNet and KRR for (a) ^{1}H, (b) ^{13}C, (c) ^{15}N and (d) ^{17}O.

Finally, we interpret the MR-3D-DenseNet model using PCA to extract chemical bonding and hydrogen-bonding information derived from the transformed data in the last fully connected layer from the ^{1}H chemical shift prediction, and projecting the first 3 principal components into a reduced 3D space as shown in Figure 6.5. Even though no explicit bonding information was provided as input to the MR-3D-DenseNet network, the model is capable of separating the C-H from the N-H and O-H chemical bond clusters (Figure 6.5a), although there is less well-defined separability of the

different types of chemical and hydrogen-bonding environments for the N-H and O-H bonds (Figure 6.5a and Figure 6.5b). The relative lack of clean separability for the N-H and O-H bonding clusters exhibited by the PCA analysis would support the conclusion that there is a lack of unique chemical environments for oxygen and nitrogen, since there are more C-H than N-H or O-H examples in the training dataset. We caution that it is possible that the chemical or hydrogen-bonded clusters for the O-H and N-H data may well be separated in a higher dimensional space, however the three principal components shown here can explain > 85% of the variances of the data (Figure S2 in the Supplementary Information).

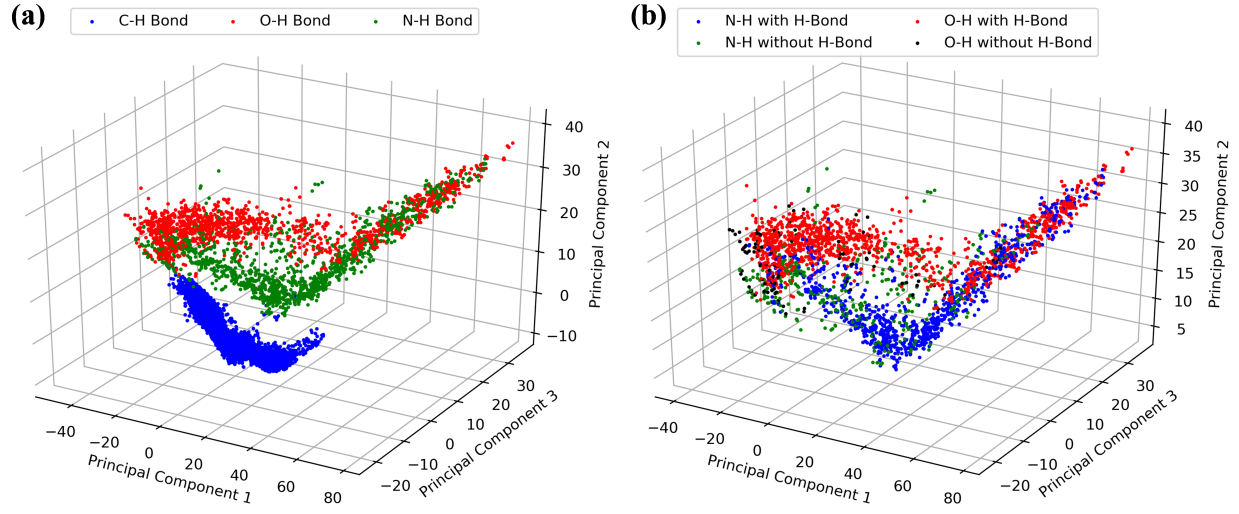


Figure 6.5: Visualization of the data in the last fully connected layer by projecting the data into 3D space using principal component analysis (PCA). It shows the clustering of different (a) chemical bonds and (b) hydrogen bonds.

6.5 Conclusion

We have presented a 3D-DenseNet deep learning model that exploits a multi-resolution input representation for NMR chemical shift prediction of atoms in molecular crystals. A unique feature of our deep learning model is the use of multi-resolution spatial input data organized into individualized channels for each atom type, that provides a learning framework for different tasks which are sensitive to the density representation with different cut-off sizes. Furthermore, the multi-resolution architecture combines the benefits of both pooling and the high-resolution feature map close to the investigated atom, which can potentially be applied to the prediction of other chemical properties sensitive to different length scales.

In addition to its greater flexibility in representing spatial density distributions, the MR-3D-DenseNet can more efficiently handle much larger data sets. In comparison to the KRR approach,

Atom Type	Number of Training Epoch	Decay Rate α
^{1}H	12	0.6
^{13}C	15	0.5
^{15}N	24	0.25
^{17}O	24	0.25

Table 6.4: Number of epochs and learning rate decay.

the MR-3D-DenseNet method has the capacity and favorable scaling characteristics that allowed us to increase the training data by an order of magnitude through rotation of the input samples to predict chemical shift values based in part on the rotational invariance property of chemical shifts. As a result, the totality of our deep learning approach can predict the chemical shifts more accurately, especially for ^{1}H chemical shifts.The accurate chemical shift prediction of ^{1}H is important for the structure characterization of many solid-state chemistry and biological systems as NMR crystallography is one of the most powerful techniques to study the structure and dynamics of hydrogen atoms in solid-state under natural abundance.

6.6 Methods

The neural network is implemented using Keras [182] with Tensorflow [183] as the backend. The density generation step is accelerated using PyTorch [184] with GPU implementation. The neural network architecture and hyperparameters are optimized through a 4-fold cross-validation on the training dataset. The training and testing are performed on Tesla P100 GPU. We trained four dedicated models for the chemical shift prediction of ^{1}H, ^{13}C, ^{15}N and ^{17}O-NMR separately. To speed up the data preprocessing process, we calculate the density using 320 nearest neighbor atoms. To accelerate the convergence, we subtract the mean of the chemical shielding values and divide them by $1, 10, 30, 40$ for ^{1}H, ^{13}C, ^{15}N and ^{17}O-NMR respectively during the training phase. The mean and scaling factors are applied back during the testing phase. Each layer is followed by a batch normalization (BatchNorm [185]) layer, and a rectified linear units (ReLU [186]) layer. There are two dropout layers added after each fully connected layers with rate 0.1. The L_2 regularizer with $\lambda = 3\times10^{-5}$ is applied to all the weights in the neural network. The training epochs used are 12, 15, 24 and 24, and the decay rates α are 0.6, 0.5, 0.25 and 0.25 for ^{1}H, ^{13}C, ^{15}N and ^{17}O, respectively. The batch size is fixed to 128. The learning rate starts with 10^{-3} and decays exponentially. In epoch i, the learning rate is decayed to $10^{-3} \times \exp(-i\alpha)$. The hyperparameter details are summarized in Table 6.4. The testing RMSEs are reported by averaging the results from at least three experiments in which the models are initialized with different random seed and the training data are shuffled in random order in each training epoch. The implementation, trained models and detailed instructions for reproducibility are available online: https://thglab.berkeley.edu/software-and-data/

6.7 Future Directions

Although our chemical shift prediction for ^{13}C, ^{15}N and ^{17}O were significantly improved with respect to KRR, it does not reach the same level of accuracy as compared to high quality *ab initio* methods. This is almost certainly due to the more limited amount of training examples available for these atom types that prevents us from exploiting the capacity of the MR-3D-DenseNet, and highlights the importance of the size, diversity, and uniqueness of the training datasets. Finally, when interpreting the MR-3D-DenseNet with PCA, we found that we can extract relevant chemical intuition for its performance on chemical shift predictions, similar to deep network interpretation for structural properties in proteins [165–169], while also yielding insight into characterizing data sufficiency that can guide future improvements in chemical shift prediction.

Currently, we only have preliminary understandings of the multiresolution architecture (crop-pool-concat). We are exploring this approach in other neural network architectures (such as 3D-CNN and 3D-ResNet). Moreover, we propose to extend this approach to other physical chemistry tasks which are sensitive to the distance.

6.8 Acknowledgments

We thank the National Institutes of Health for support under Grant No. 5U01GM121667. Partial work on the development of machine learning algorithms was supported by the Center of Advanced Mathematics for Energy Research Applications (CAMERA) through the Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Early Career Program. This research used the computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

6.9 Supplementary Information

Electron density representation

Our MR-3D-DenseNet can represent different electron density models straightforwardly by calculating the numerical density on the 3D grid. Here, we investigate three different densities: Gaussian density with different variance (σ^2), Slater orbital density, and the electron density calculated from the inverse Fourier transform of the atomic form factor. The Slater orbital is formulated as

$$R(r) = Nr^{n-1}e^{-\zeta r} \tag{6.3}$$

where n is the principal quantum number, N is the normalizing constant, r is the distance in the physical space, and ζ is the effective charge of nucleus estimated by the Slater's rules. The ζ for different atom types are available in literature [187]. The atomic form factor can be approximated by a sum of Gaussian distribution

$$f(q) = \sum_{i=1}^{4} a_i \exp(-b_i(\frac{q}{4\pi})^2) + c \tag{6.4}$$

where a, b, c are coefficients available in literature [188]. By calculating the inverse Fourier transform, the electron density can be approximated by

$$D(x) = \sum_{i=1}^{4} \frac{2\sqrt{2}\pi a_i}{\sqrt{b_i}} \exp(-\frac{1}{b_i}(2\pi x)^2) + \sqrt{2\pi}c\delta(x) \tag{6.5}$$

where δ is Dirac delta function, which is not represented in the numerical density calculation.

The testing RMSEs of ^{1}H-NMR chemical shift predictions are summarized in Table 6.5. In comparison to the Slater orbital density and the electron density calculated from atomic form factor, the Gaussian density gives the best prediction performance. We also investigated the Gaussian density with different variances (σ^2). The smaller variance leads to a narrow Gaussian distribution over the 3D grid. The nearest voxel representation is to approximate the condition when $\sigma \to 0$. Under small σ, the prediction performance is unsatisfactory because the information on the 3D grid is too sparse. A similar trend was also observed in by Kuzminykh et al. in literature [173]. Under large σ, the Gaussian distribution is flat with the heavy tail, which also leads to unsatisfactory prediction performance. Empirically, the Gaussian density with variance $\sigma^2 = \frac{1}{3}$ provides the best prediction performance.

Density	Testing RMSE
Nearest Voxel	0.45
Gaussian $\sigma^2 = \frac{1}{10}$	0.42
Gaussian $\sigma^2 = \frac{1}{3}$	0.37
Gaussian $\sigma^2 = 1$	0.39
Slater Orbital Density	0.48
Electron Density from Atomic Form Factor	0.39

Table 6.5: Testing RMSEs (ppms) of ^{1}H-NMR chemical shift predictions using MR-3D-DenseNet model with different densities with data augmentation.

In comparison to Gaussian densities, the Slater orbital density and the electron density calculated from atomic form factors both lead to worse performance, presumably because of the heavy tails in these distributions (similar to the Gaussian distribution with large variance). All the density distributions are plotted in Figure 6.6. In comparison to the single $\exp(-x^2)$ decay of Gaussian distribution, the Slater orbital density has a longer tail $\exp(-x)$. Similarly, in the atomic form factor, the components with large b contribute to high variance components in the electron density distribution, which also leads to the heavy tail issue.

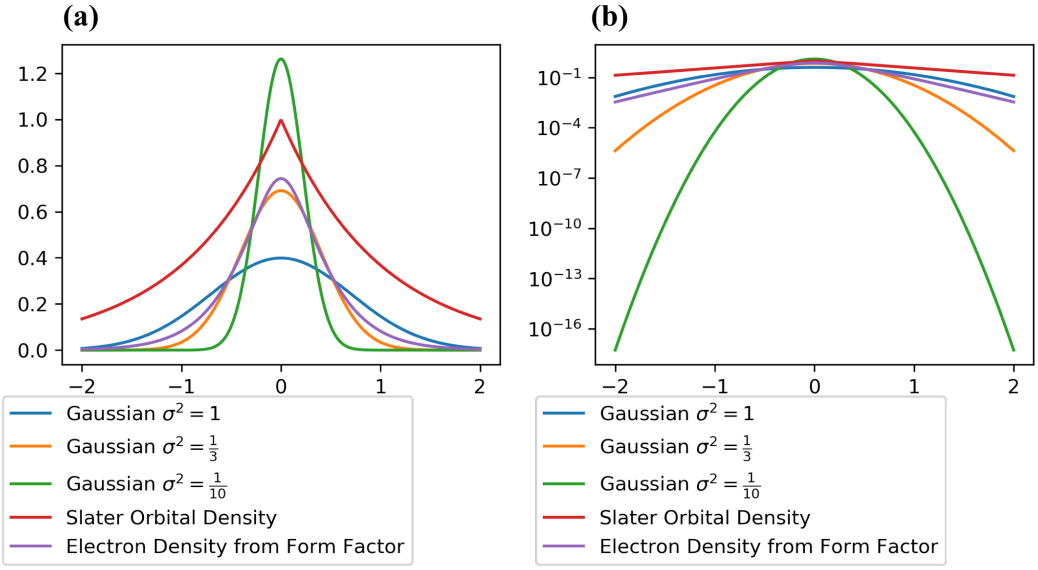


Figure 6.6: The plots of (a) different densities and (b) in log scale.

PCA Result

To visualize the data, we performed Principal Component Analysis (PCA) on the last fully connected layer of the CNN, which is a projection of feature map to low dimensional space.

Mathematically, we can measure the information carried in principal components using explained variance ratio. The explained variance ratio of component k is defined as

$$\frac{\sigma_k^2}{\sum_{i=1}^{d} \sigma_i^2} \tag{6.6}$$

which is plotted in Figure 6.7.

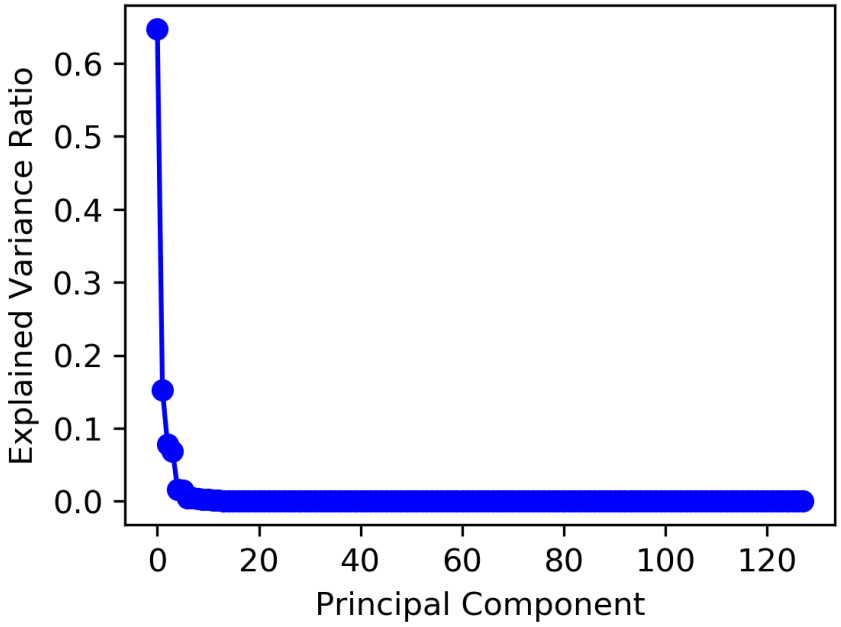


Figure 6.7: Explained ratio as a function of the number of principal components.

Chapter 7

Data Driven Approach for Facility Optimization: A Case Study at ALS

In this chapter, I apply machine learning methods to characterization facility optimization, which is the prerequisite to collect high quality characterization data. We present the first attempt to apply machine learning model to stabilize the light source, where the current physics model cannot provide perfect predictions and corrections. This chapter consists of a first proof-of-concept example during applied physics time and the discussions of future improvement suggestions on the long-term running in daily operations.

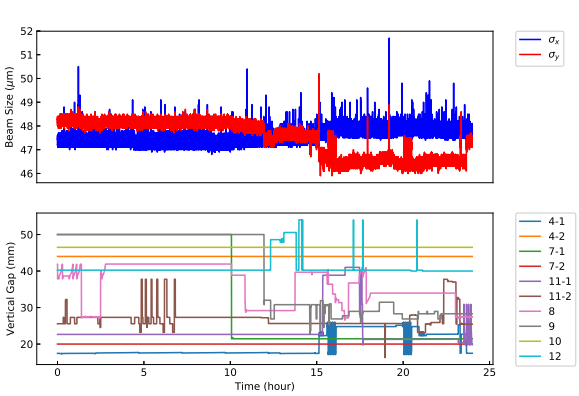
This chapter is a collaborative work joint with the Accelerator Group at the Advanced Light Source (ALS). My contribution has been to build the machine learning models and related data and prediction analysis, and my collaborators (Dr. Simon Leemann, Dr. Hiroshi Nishimura and Dr. David Shapiro) conducted the data collection and built the interface to communicate with the accelerator.

Part of this chapter is adapted from the manuscript "Demonstration of Machine Learning-Based Model-Independent Stabilization of Source Properties in Synchrotron Light Sources" with the permission.

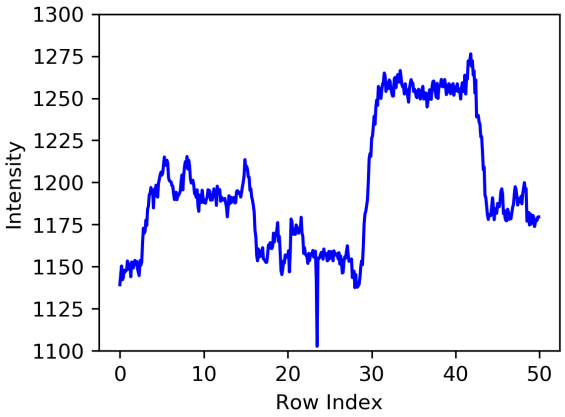
7.1 Introduction

A Motivating Example

Synchrotron light source is one of the most powerful techniques for the characterization of chemistry, materials or biological systems. The quality of characterization data is sensitive to many instrument factors. One of them is the beam size variation from the radiation source due to the insertion device (ID) movement. This beam size variation influences the data quality for many different experiments, for example, the Scanning Transmission X-ray Microscopy (STXM), which has been applied to many polymer systems to understand their structure-property relationship. Figure 7.1 provides an example of beam size variation and how this variation perturbs the STXM data. In this chapter, I will provide a proof-of-concept demonstration of predicting and stabilizing the the synchrotron radiation source using neural networks.



(a) Beam size variation with ID movement



(b) STXM intensity fluctuation

Figure 7.1: Beam size variation and the induced intensity instability in STXM intensity. The measurement of STXM is collected with the help from Dr. David Shapiro at ALS beamline 5.3.2.2.

Introduction to the Accelerator

This subsection is a short version of the introduction adapted from manuscript Liu et al., "Demonstration of Machine Learning-Based Model-Independent Stabilization of Source Properties in Synchrotron Light Sources" with permission.

Synchrotron radiation sources have been widely applied to understanding chemical, materials and biological systems, which significantly advanced these fields. One of the advantage of these radiation source is that they can provide stable radiation with broad spectrum and high brightness. Since the high demand of the synchrotron radiation facility, the new generation light source (4GLS) has been proposed and will be widely applied in the next decade [189]. These sources will increase

average brightness by 2–3 orders of magnitude while also delivering high degrees of transverse coherence, for the first time even for X-rays.

These success relies on constant radiation intensity delivered by the source. However, the stability of new radiation source starts to be compromised by the perturbation of source size control. Currently, many stabilization efforts are based on the linear approximation and superpositions (for example, based on linear optics [190–192]) using a pre-collected look-up table on each insertion device (ID) individually. However, practically, these linear approximations sometimes cannot provide satisfactory stabilization. Also, the look-up collection process is time-consuming, however, it has to be constantly refreshed due to the drifting with different parameters. On the other hand, traditional feedback corrections attempt to counteract such drift, but often do not offer sufficient closed-loop bandwidth to remove perturbations over the entire desired range. In all, the stabilization of next generation radiation source requires a more advanced technique.

Our Approach

Recently, data driven methods have been applied to many different research areas. Specifically, neural networks (NNs) have proved to be one of the most effective approaches for nonlinear function fitting, both theoretically and empirically [193]. Here, we propose a NN approach to predicting electron beam size as a function of arbitrary ID gap/phase configurations and employing this prediction to correct for perturbations thereby suppressing source size fluctuations. Control of the electron beam size exploits the fact that commonly 3GLSs use skew quadrupoles to correct betatron coupling and spurious vertical dispersion first, and then to excite a vertical dispersion wave which improves beam lifetime within the boundaries of the diffraction limit [194]. In this chapter, I demonstrate for the first time an alternative approach to stabilizing source sizes through use of machine learning relying only on previously existing instrumentation.

7.2 Models and Data

Data Source Description

As discussed in the previous sections, we aim to predict and stabilize the beam size on the vertical direction. Table 7.1 list the IDs that affect the vertical beam size. We performed a preliminary study on the beam size prediction during user operations with 7,000,000 data points by shuffling and then splitting the dataset into training and testing data with ratio of 4 to 1. In the beam size stabilization section, beam size (as measured at a diagnostic beamline) along with all relevant beam parameters and ID settings have to be captured at high data rates. At ALS, we have so far chosen an acquisition rate of 10 Hz (similar scale to beam size measurement update rates and typical ID gap/phase motion) at which we collect data for roughly 23 independent channels. Moreover, the DWP needs to be included in stabilization section as the extra parameter.

Index	Parameters	Descriptions
1	SR04U.1.V	Vertical Gap of SR04U.1
2	SR04U.1.A	Horizontal Shift of SR04U.1 (A)
3	SR04U.1.B	Horizontal Shift of SR04U.1 (B)
4	SR04U.2.V*	Vertical Gap of SR04U.2
5	SR04U.2.A*	Horizontal Shift of SR04U.2 (A)
6	SR04U.2.B*	Horizontal Shift of SR04U.2 (B)
7	SR07U.1.V	Vertical Gap of SR07U.1
8	SR07U.1.A	Horizontal Shift of SR07U.1 (A)
9	SR07U.1.B	Horizontal Shift of SR07U.1 (B)
10	SR07U.2.V	Vertical Gap of SR07U.2
11	SR07U.2.A	Horizontal Shift of SR07U.2 (A)
12	SR07U.2.B	Horizontal Shift of SR07U.2 (B)
13	SR011U.1.V	Vertical Gap of SR11U.1
14	SR011U.1.A	Horizontal Shift of SR11U.1 (A)
15	SR011U.1.B	Horizontal Shift of SR11U.1 (B)
16	SR011U.2.V*	Vertical Gap of SR11U.2
17	SR011U.2.A*	Horizontal Shift of SR11U.2 (A)
18	SR011U.2.B*	Horizontal Shift of SR11U.2 (B)
19	SR06U	Vertical Gap of SR06U
20	SR08U	Vertical Gap of SR08U
21	SR09U	Vertical Gap of SR09U
22	SR10U*	Vertical Gap of SR10U
23	SR12U	Vertical Gap of SR12U
24	DWP	Dispersion Wave Parameter

Table 7.1: Input of the NN model. The DWP is included in stabilization section (section 7.3). The IDs with $*$ are only included in some experiments due to limit amount of data collection time and/or accelerator instability.

Neural Networks

The NNs are implemented using the Keras [182] with the Tensorflow [183] backend. The loss function is mean squared error (MSE), which is a common metric for regression problems. The models are trained using the back-propagation method employing the Adam optimizer for 40 epochs. The learning rate is set to 10^{-3} with a decay multiplier of 0.8 after each epoch. By optimizing the NN architecture and the parameters, we choose the NN containing three hidden layers with sizes 128, 64, 32, respectively, with activation function ReLU. A small L_2 regularization with $\lambda = 10^{-4}$ is added to each layer to mitigate overfitting. The training takes 20 minutes on a single desktop-class CPU. We obtained good validation error (less than 0.3 μm, the training/validation splitting is discussed in last subsection). An important note is that this validation process only proves the

feasibility of fitting the vertical beam size using neural network. However, this does not guarantee that the NN can be generalized to *any* ID parameter space with this validation error. The data distribution in this dataset may not be diverse enough even it has 7 million unique measurement data points. However, the generalizability is application dependent. To test the performance of neural network approach for the accelerator, we conduct several preliminary studies in the next section to illustrate the first attempts.

7.3 Beam Size Stabilization

A Proof-of-Concept Demonstration During Applied Physics Time

In this section, we add DWP during the data collection phase in dedicated applied physics time and train the NN. Such a pre-trained NN can then be employed for beam size stabilization in a FF fashion by screening DWP during the data collection. Given a target beam size and the current combination of ID settings, we pre-screened 100 possible DWPs between -0.06 to 0.06 uniformly using NN[1]. Evaluating 100 DWPs only takes milliseconds on a single CPU, which enables us to implement this control at > 10 Hz. We select the DWP which renders the beam size closest to the target. The selected DWP value is used in a FF manner to adjust the skew quadrupole excitation pattern that generates the vertical dispersion wave[2]. The experimental result is shown in Figure 7.2.

[1] At the ALS, scanning the DWP over a range of ± 0.06 corresponds to roughly ± 5 μm around 48 μm vertical beam size as measured at diagnostic beamline 3.1. The source point of this beamline is in the first bend magnet of the triple-bend achromat cell resulting in roughly equal transverse beam sizes.

[2] We have so far chosen 3 Hz as the update rate of the FF to match roughly the update rate of the beamline 3.1 beam size measurement. This measurement can be refreshed at much higher rate if a region of interest is chosen in the camera. This will allow us to increase the update rate of the FF in the near future.

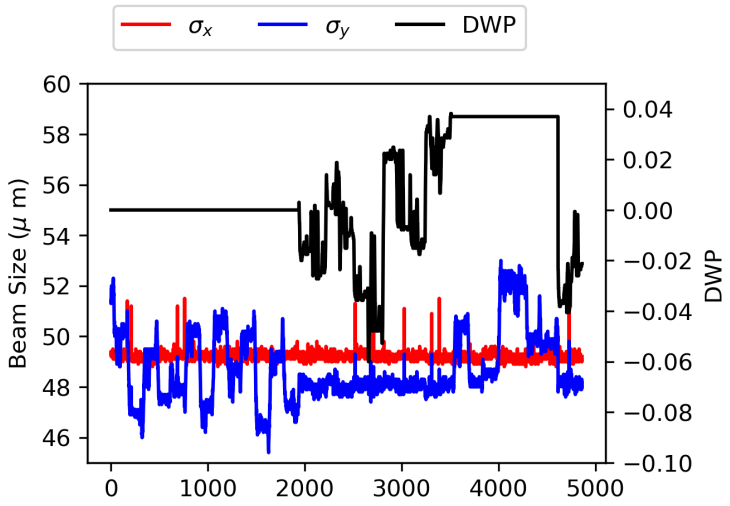
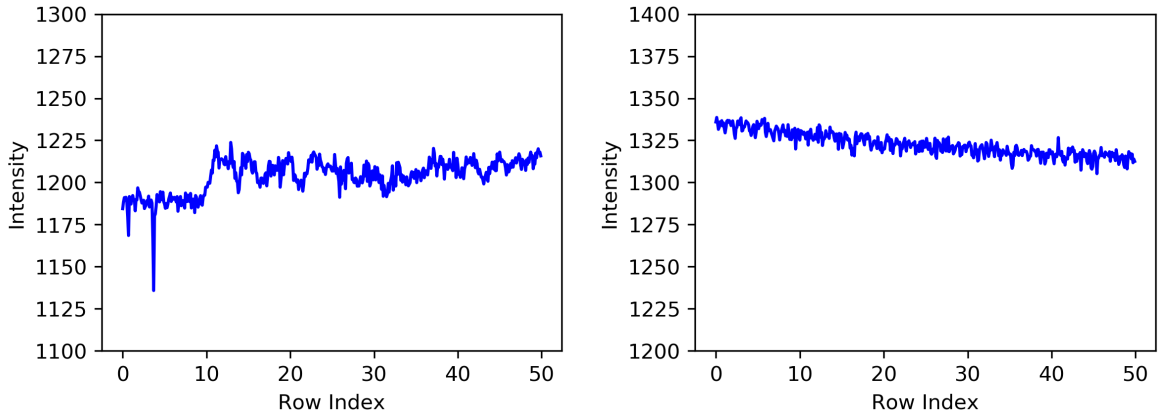


Figure 7.2: Beam sizes as measured at the ALS diagnostic beamline 3.1 along with various ID vertical gap settings over several hours. The NN-based FF loop was opened and closed repeatedly.

We turned FF control on and off repeatedly to verify the effectiveness of our beam size stabilization approach. In this example, when the FF is on, the variation of vertical beam size as measured at the diagnostic beamline significantly decreases in rms and peak-to-peak variance. For comparison with the NN-based FF, we also implemented a simple FB loop relying solely on beam size measurement as an input and returning a DWP requested for beam size correction. During calm periods with only very slow ID configuration changes, the FB performance was capable of delivering similar rms stabilization as the NN-based FF. However, as soon as ID configurations changed at rates typically observed during experiments (e.g. 4 mm/s vertical gap motion and 16.7 mm/s horizontal shifts), the FB failed. Depending on PID tuning it was either not capable of stabilizing against transients (leading to 3 μm peak-to-peak vertical beam size variation, i.e. 6%) or it became unstable. The NN-based FF approach outperforms the FB method primarily for two reasons. First, the FF approach is agnostic to the current beam size. Implementing this FF does not require beam size as an input, hence adjusting beam size ahead of the measurement and avoiding measurement delay. Second, the NN-based FF does not have to adjust the DWP in a continuous fashion employing a series of small steps. It can instantaneously adjust the DWP by any large amount required to maintain stable beam size without overshoot.

So far, these experiments have revealed that the NN-based FF can stabilize the vertical beam size at the diagnostic beamline. It is, however, a priori not at all evident that stabilizing the source size at one point in the storage ring is equivalent to stabilizing the beam at the relevant source points. We originally chose to act on the beam size by means of the vertical dispersion wave, since it adjusts the vertical emittance, a global and conserved property, and we can therefore expect it to stabilize globally in spite of training the NN using a localized measurement. In order to demonstrate that this interpretation is correct, we conducted experiments at ALS beamline 5.3.2.2, which is



(a) With NN-based Model (b) Without ID motion

Figure 7.3: STXM intensity from ALS beamline 5.3.2.2 at 390 eV.

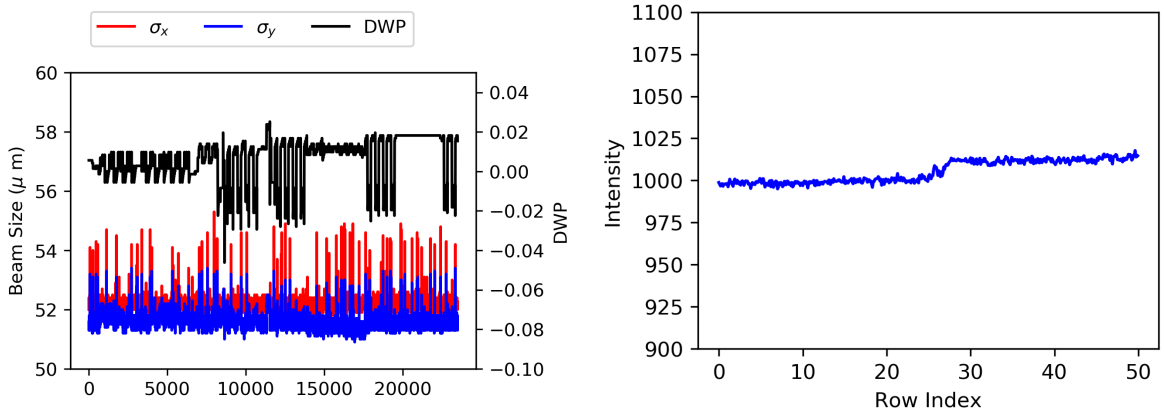
the most sensitive ALS beamline in terms of vertical beam size. Figure 7.3 shows STXM scan data taken at 5.3.2.2 while ID configurations in the rest of the ALS were continuously changing. The measurement data reveals that the stabilization observed at the diagnostic beamline can indeed also be observed in the STXM scans at this sensitive beamline. In fact, the reduction of intensity variation detected at the STXM beamline (compare Figure 7.3 left to Figure 7.1 right) corresponds almost exactly to the peak-to-peak reduction in beam size variation noted at the diagnostic beamline when opening and closing the NN-based FF loop (cf. above). These STXM measurements also reveal that this stabilization of low-frequency perturbations does not occur at the expense of exciting additional high-frequency noise. As shown in Figure 7.3, with the reduction in noise observed during the STXM scan, the residual noise from ID configuration changes now lies only 60% above the noise floor of the experiment. We expect to reduce this residual by increasing the beam size measurement refresh rate and consequently the NN-based FF update rate.

This proof-of-concept study is important because it is the first attempt to use machine learning technique to stabilize the radiation source. However, this study also has its limitation. The ID settings of training data and online testing are unique but from similar sampling distribution (with the knowledge from domain experts to simulate data distribution during user operations). More experiments are needed to evaluate the generalizability of the current approach under the ID settings with a significantly different data distribution. Practically, for beam size stabilization during user operations, we need to better understand the ID setting distribution of user operations and test it on the user operations, which will be illustrated in the next subsection.

Preliminary Studies During Regular Experimental Operations

In this subsection, we present a preliminary study of the beam size stabilization during user operations. We fit the NN by combining the data collected during the dedicated applied physics

time with the data during the user operations to better capture the distribution of ID parameter space. The user operation data was randomly down-sampled to 1/15 of its original size to balance sample sizes. Retraining the NN using both data sets requires just 15 minutes on a desktop-class CPU. An example of the beam size measurement under this scheme is shown in Figure 7.4.



(a) Beam size variation w/ NN-based model

(b) STXM data with NN-based model

Figure 7.4: Beam size variation and STXM data during user operations with NN-based model during user operations.

From Figure 7.4, we observe that during user operations, the DWP can be tuned given the ID settings. The beam size variation stays low in the 6-hour window. We also have the statistics of the beam size variation for days with and without NN-based control scheme (shown in Table 7.2).

Date	**Beam Size Standard Deviation**	
	w/o NN	**w/ NN**
04/25	0.496	0.275
04/26	0.216	0.393
04/27	0.901	0.268
04/28	0.723	N/A
Machine Physics Time and Maintenance		
04/30	0.616	0.479
05/01	0.531	0.932
05/02	0.347	0.446

Table 7.2: The standard deviation of beam size from 04/25 to 05/02. The model was trained using the data up to 04/24. We turn on the NN control several hours each day to calculate the standard deviation of the beam size with and without control of NN, respectively.

During the first three days after the NN was trained, the average beam size variation over 3 days (04/25 to 04/28) are 0.312 μm and 0.538 μm with and without NN-based FF control. This indicates that the model can be potentially applied to the user operations and stabilize the beam size. However, its long-term effectiveness has to be more carefully evaluated because (1) we do observe that the standard deviation of beam size w/o NN on 04/26 is smaller than w/ NN, (2) the day-to-day beam size variance is large. Moreover, after the machine physics time and maintenance, the average beam size over three days are 0.619 μm and 0.507 μm (over 04/30 to 05/02), respectively. Therefore, the NN based method needs to be more carefully examined. One hypothetical reason is that the accelerator configuration changes as a function of time or during the reset after the machine physics time. Another possible reason is that the distribution of ID settings during user operation is much more diverse than the parameter space scanned during the applied physics time (even together with the sub-sampled user operation data). There are several possible improvements and suggestions:

1. Attempt to maintain the configuration stable, consistent and standardized manually at each time.
2. Cover a more diverse ID distribution (or closer to the distribution in user operation hours/settings) during the data collection in applied physics time.
3. Extend the time and variety of data during the user operations and include them into the model retraining by better sampling strategy.

7.4 Conclusion

We have demonstrated that machine learning can be employed to render NNs that can predict the vertical source size at storage ring light sources. Moreover, we performed the first proof-of-concept study during the applied physics time on stabilizing the beam size without requiring any new instrumentation. We also conducted some preliminary studies and analysis during the user operations and provided suggestions for future improvements. For example, the training data distribution needs to be more diverse and consistent with the user operation distributions. In addition, standardizing the accelerator configuration is essential but not trivial. In all, the demonstrated technique can be potentially applied for future accelerator development with more demanding brightness and transverse coherence.

7.5 Future Outlooks

This chapter presents a first proof-of-concept example of controlling the stability of light source using neural network. However, the long-term influence and maintenance need more testing and engineering efforts. Moreover, in the future, we plan to investigate if a NN-based FF can replace model-based FFs entirely, thus freeing up on the order of one hundred hours of dedicated machine time a year, which are nowadays still required to re-record look-up tables. In addition, when more control parameters are involved, reinforcement learning methods (e.g., policy gradient or deep

Q-learning) will be investigated systematically. Moreover, we hope that the analysis of the trained NN may provide some insights on understanding the machine physics at ALS.

7.6 Acknowledgments

We would like to thank Greg Penn, Thorsten Hellert, and the ALS Operators for their support during measurement shifts. We would like to acknowledge David Shapiro for many helpful discussions and his generous support at beamline 5.3.2.2. We are most grateful to Fernando Sannibale, Marco Venturini, and Andreas Scholl for many interesting discussions and their valuable advice. Finally, extensive support is also acknowledged from Daniela Ushizima. The research performed here is funded by the US Department of Energy (BES & ASCR Programs), and supported by the Director of the Office of Science of the US Department of Energy under Contract No. DEAC02-05CH11231.

Bibliography

(1) Cargill, J. F.; Lebl, M. *Current opinion in chemical biology* **1997**, *1*, 67–71.

(2) Peplow, M. *Nature News* **2014**, *512*, 20.

(3) Kündig, P. *Science* **2006**, *314*, 430–431.

(4) Roe, R.-J. *Oxford University Press* **2000**, *9*, 10–12.

(5) Maslen, E.; Fox, A.; O'Keefe, M. **2006**.

(6) Tolan, M., *X-ray scattering from soft-matter thin films: materials science and basic research*; Springer: 1999.

(7) Harris, R. K.; Wasylishen, R. E.; Duer, M. J., *NMR crystallography*; John Wiley & Sons: 2010.

(8) Brünger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J.-S.; Kuszewski, J.; Nilges, M.; Pannu, N. S., et al. *Acta Crystallographica Section D: Biological Crystallography* **1998**, *54*, 905–921.

(9) Harris, R. K. *Solid state sciences* **2004**, *6*, 1025–1037.

(10) Walenta, E. *Acta Polymerica* **1985**, *36*, 296–296.

(11) Roe, R.-J. *Oxford University Press* **2000**, *9*, 10–12.

(12) Chu, B.; Hsiao, B. S. *Chemical Reviews* **2001**, *101*, 1727–1762.

(13) Bawendi, M.; Kortan, A.; Steigerwald, M.; Brus, L. *The Journal of chemical physics* **1989**, *91*, 7282–7290.

(14) Pietsch, U.; Holy, V.; Baumbach, T., *High-resolution X-ray scattering: from thin films to lateral nanostructures*; Springer Science & Business Media: 2013.

(15) Fewster, P. F., *X-Ray scattering from semiconductors and other materials*; World Scientific: 2015.

(16) Svergun, D. I.; Petoukhov, M. V.; Koch, M. H. *Biophysical journal* **2001**, *80*, 2946–2953.

(17) Bouwstra, J. A.; Gooris, G. S.; van der Spek, J. A.; Bras, W. *Journal of Investigative Dermatology* **1991**, *97*, 1005–1012.

(18) Bagno, A.; Rastrelli, F.; Saielli, G. *The Journal of Physical Chemistry A* **2003**, *107*, 9964–9973.

(19) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. *Nature communications* **2018**, *9*, 4501.

(20) Sarje, A.; Li, X. S.; Hexemer, A. In *2014 43rd International Conference on Parallel Processing*, 2014, pp 201–210.

(21) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. *Nature* **2016**, *533*, 73.

(22) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Computers & chemistry* **2001**, *26*, 5–14.

(23) Schütt, K.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K.; Gross, E. *Physical Review B* **2014**, *89*, 205118.

(24) Kutz, J. N. *Journal of Fluid Mechanics* **2017**, *814*, 1–4.

(25) Janet, J. P.; Chan, L.; Kulik, H. J. *The journal of physical chemistry letters* **2018**, *9*, 1064–1071.

(26) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. *Scientific reports* **2013**, *3*, 2810.

(27) Hexemer, A.; Müller-Buschbaum, P. *IUCrJ* **2015**, *2*, 106–125.

(28) Liu, S.; Melton, C. N.; Venkatakrishnan, S.; Pandolfi, R. J.; Freychet, G.; Kumar, D.; Tang, H.; Hexemer, A.; Ushizima, D. M. *MRS Communications* **2019**, 1–7.

(29) Vineyard, G. H. *Physical Review B* **1982**, *26*, 4146.

(30) Sinha, S.; Sirota, E.; Garoff; S; Stanley, H. *Physical Review B* **1988**, *38*, 2297.

(31) Rauscher, M.; Paniago, R.; Metzger, H.; Kovats, Z.; Domke, J.; Peisl, J.; Pfannes, H.-D.; Schulze, J.; Eisele, I. *Journal of Applied Physics* **1999**, *86*, 6763–6769.

(32) Hexemer, A.; Müller-Buschbaum, P. *IUCrJ* **2015**, *2*, 106–125.

(33) Nogales, A.; Hsiao, B. S.; Somani, R. H.; Srinivas, S.; Tsou, A. H.; Balta-Calleja, F. J.; Ezquerra, T. A. *Polymer* **2001**, *42*, 5247–5256.

(34) Liu, Y.; Zhao, J.; Li, Z.; Mu, C.; Ma, W.; Hu, H.; Jiang, K.; Lin, H.; Ade, H.; Yan, H. *Nature communications* **2014**, *5*, 5293.

(35) Park, S.; Lee, D. H.; Xu, J.; Kim, B.; Hong, S. W.; Jeong, U.; Xu, T.; Russell, T. P. *Science* **2009**, *323*, 1030–1033.

(36) Loo, Y.-L.; Register, R. A.; Ryan, A. J. *Macromolecules* **2002**, *35*, 2365–2374.

(37) Rancatore, B. J.; Mauldin, C. E.; Tung, S.-H.; Wang, C.; Hexemer, A.; Strzalka, J.; Fréchet, J. M.; Xu, T. *ACS nano* **2010**, *4*, 2721–2729.

(38) Bai, P.; Kim, M. I.; Xu, T. *Macromolecules* **2013**, *46*, 5531–5537.

(39) Paik, M. Y.; Bosworth, J. K.; Smilges, D.-M.; Schwartz, E. L.; Andre, X.; Ober, C. K. *Macromolecules* **2010**, *43*, 4253–4260.

(40) Agzenai, Y.; Lindman, B.; Alfredsson, V.; Topgaard, D.; Renamayor, C. S.; Pacios, I. E. *The Journal of Physical Chemistry B* **2014**, *118*, 1159–1167.

(41) Polte, J.; Erler, R.; Thunemann, A. F.; Sokolov, S.; Ahner, T. T.; Rademann, K.; Emmerling, F.; Kraehnert, R. *ACS nano* **2010**, *4*, 1076–1082.

(42) Kwon, S. G.; Krylova, G.; Phillips, P. J.; Klie, R. F.; Chattopadhyay, S.; Shibata, T.; Bunel, E. E.; Liu, Y.; Prakapenka, V. B.; Lee, B., et al. *Nature materials* **2015**, *14*, 215.

(43) Krycka, K. L.; Borchers, J. A.; Salazar-Alvarez, G.; López-Ortega, A.; Estrader, M.; Estrade, S.; Winkler, E.; Zysler, R. D.; Sort, J.; Peiro, F., et al. *ACS nano* **2013**, *7*, 921–931.

(44) Lin, Y.; Böker, A.; He, J.; Sill, K.; Xiang, H.; Abetz, C.; Li, X.; Wang, J.; Emrick, T.; Long, S., et al. *Nature* **2005**, *434*, 55.

(45) Ye, X.; Zhu, C.; Ercius, P.; Raja, S. N.; He, B.; Jones, M. R.; Hauwiller, M. R.; Liu, Y.; Xu, T.; Alivisatos, A. P. *Nature communications* **2015**, *6*, 10052.

(46) Macfarlane, R. J.; Lee, B.; Jones, M. R.; Harris, N.; Schatz, G. C.; Mirkin, C. A. *science* **2011**, *334*, 204–208.

(47) Cui, J.; Olmsted, D. L.; Mehta, A. K.; Asta, M.; Hayes, S. *Angewandte Chemie International Edition* **2019**.

(48) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2016**, *72*, 171–179.

(49) Yee, A. A.; Savchenko, A.; Ignachenko, A.; Lukin, J.; Xu, X.; Skarina, T.; Evdokimova, E.; Liu, C. S.; Semesi, A.; Guido, V., et al. *Journal of the American Chemical Society* **2005**, *127*, 16512–16517.

(50) Elena, B.; Pintacuda, G.; Mifsud, N.; Emsley, L. *Journal of the American Chemical Society* **2006**, *128*, 9555–9560.

(51) Pickard, C. J.; Mauri, F. *Physical Review B* **2001**, *63*, 245101.

(52) Yates, J. R.; Pickard, C. J.; Mauri, F. *Physical Review B* **2007**, *76*, 024401.

(53) Blöchl, P. E. *Physical review B* **1994**, *50*, 17953.

(54) Elena, B.; Emsley, L. *Journal of the American Chemical Society* **2005**, *127*, 9140–9146.

(55) Elena, B.; Pintacuda, G.; Mifsud, N.; Emsley, L. *Journal of the American Chemical Society* **2006**, *128*, 9555–9560.

(56) Abraham, A.; Apperley, D. C.; Byard, S. J.; Ilott, A. J.; Robbins, A. J.; Zorin, V.; Harris, R. K.; Hodgkinson, P. *CrystEngComm* **2016**, *18*, 1054–1063.

(57) Skotnicki, M.; Apperley, D. C.; Aguilar, J. A.; Milanowski, B.; Pyda, M.; Hodgkinson, P. *Molecular pharmaceutics* **2015**, *13*, 211–222.

(58) Zhou, Z.; Zare, R. N. *Analytical chemistry* **2017**, *89*, 1369–1372.

(59) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186–190.

(60) Segler, M. H.; Preuss, M.; Waller, M. P. *Nature* **2018**, *555*, 604.

(61) Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. *Scientific reports* **2013**, *3*, 2810.

(62) Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. *Nature materials* **2006**, *5*, 641.

(63) De Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M.; Gamst, A. *Scientific reports* **2016**, *6*, 34256.

(64) Rossouw, D.; Burdet, P.; de la Penfffdfffda, F.; Ducati, C.; Knappett, B. R.; Wheatley, A. E.; Midgley, P. A. *Nano letters* **2015**, *15*, 2716–2720.

(65) Park, W. B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S. P.; Pyo, M.; Shin, N.; Sohn, K.-S. *IUCrJ* **2017**, *4*, 486–494.

(66) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. *Nature communications* **2018**, *9*, 2775.

(67) Wang, B.; Yager, K.; Yu, D.; Hoai, M. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp 697–704.

(68) Cuny, J.; Xie, Y.; Pickard, C. J.; Hassanali, A. A. *Journal of chemical theory and computation* **2016**, *12*, 765–773.

(69) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. *Nature communications* **2018**, *9*, 4501.

(70) Liu, S.; Xu, T. *Macromolecules* **2016**, *49*, 6075–6083.

(71) Armand, M.; Endres, F.; MacFarlane, D. R.; Ohno, H.; Scrosati, B. In *Materials For Sustainable Energy: A Collection of Peer-Reviewed Research and Review Articles from Nature Publishing Group*; World Scientific: 2011, pp 129–137.

(72) Rogers, R. D.; Seddon, K. R. *Science* **2003**, *302*, 792–793.

(73) Wakai, C.; Oleinikova, A.; Ott, M.; Weingärtner, H. *The Journal of Physical Chemistry B* **2005**, *109*, 17028–17030.

(74) Kim, S. H.; Hong, K.; Xie, W.; Lee, K. H.; Zhang, S.; Lodge, T. P.; Frisbie, C. D. *Advanced Materials* **2013**, *25*, 1822–1846.

(75) Zhang, X.; Zhang, X.; Dong, H.; Zhao, Z.; Zhang, S.; Huang, Y. *Energy & Environmental Science* **2012**, *5*, 6668–6681.

(76) Hasib-ur-Rahman, M.; Siaj, M.; Larachi, F. *Chemical Engineering and Processing: Process Intensification* **2010**, *49*, 313–322.

(77) Maiti, A. *ChemSusChem: Chemistry & Sustainability Energy & Materials* **2009**, *2*, 628–631.

(78) Lozano, L.; Godínez, C.; De Los Rios, A.; Hernández-Fernández, F.; Sánchez-Segado, S.; Alguacil, F. J. *Journal of Membrane Science* **2011**, *376*, 1–14.

(79) Guo, M.; Fang, J.; Xu, H.; Li, W.; Lu, X.; Lan, C.; Li, K. *Journal of membrane science* **2010**, *362*, 97–104.

(80) Wang, P.; Wenger, B.; Humphry-Baker, R.; Moser, J.-E.; Teuscher, J.; Kantlehner, W.; Mezger, J.; Stoyanov, E. V.; Zakeeruddin, S. M.; Grätzel, M. *Journal of the American Chemical Society* **2005**, *127*, 6850–6856.

(81) Rosen, B. A.; Salehi-Khojin, A.; Thorson, M. R.; Zhu, W.; Whipple, D. T.; Kenis, P. J.; Masel, R. I. *Science* **2011**, 1209786.

(82) Tokuda, H.; Hayamizu, K.; Ishii, K.; Susan, M. A. B. H.; Watanabe, M. *The Journal of Physical Chemistry B* **2004**, *108*, 16593–16600.

(83) Tokuda, H.; Hayamizu, K.; Ishii, K.; Susan, M. A. B. H.; Watanabe, M. *The Journal of Physical Chemistry B* **2005**, *109*, 6103–6110.

(84) Tokuda, H.; Ishii, K.; Susan, M. A. B. H.; Tsuzuki, S.; Hayamizu, K.; Watanabe, M. *The Journal of Physical Chemistry B* **2006**, *110*, 2833–2839.

(85) He, Y.; Lodge, T. P. *Macromolecules* **2008**, *41*, 167–174.

(86) Gu, Y.; Zhang, S.; Martinetti, L.; Lee, K. H.; McIntosh, L. D.; Frisbie, C. D.; Lodge, T. P. *Journal of the American Chemical Society* **2013**, *135*, 9652–9655.

(87) He, Y.; Boswell, P. G.; Bühlmann, P.; Lodge, T. P. *The Journal of Physical Chemistry B* **2007**, *111*, 4645–4652.

(88) Zhang, S.; Lee, K. H.; Frisbie, C. D.; Lodge, T. P. *Macromolecules* **2011**, *44*, 940–949.

(89) Zhang, S.; Lee, K. H.; Sun, J.; Frisbie, C. D.; Lodge, T. P. *Macromolecules* **2011**, *44*, 8981–8989.

(90) Cho, J. H.; Lee, J.; Xia, Y.; Kim, B.; He, Y.; Renn, M. J.; Lodge, T. P.; Frisbie, C. D. *Nature materials* **2008**, *7*, 900.

(91) Lee, J.; Panzer, M. J.; He, Y.; Lodge, T. P.; Frisbie, C. D. *Journal of the American Chemical Society* **2007**, *129*, 4532–4533.

(92) Cho, J. H.; Lee, J.; He, Y.; Kim, B.; Lodge, T. P.; Frisbie, C. D. *Advanced materials* **2008**, *20*, 686–690.

(93) Gu, Y.; Lodge, T. P. *Macromolecules* **2011**, *44*, 1732–1736.

(94) Simmons, M. R.; Patrickios, C. S. *Macromolecules* **1998**, *31*, 9075–9077.

(95) Yuan, J.; Antonietti, M. *Polymer* **2011**, *52*, 1469–1482.

(96) Mecerreyes, D. *Progress in Polymer Science* **2011**, *36*, 1629–1648.

(97) Ruokolainen, J.; Mäkinen, R.; Torkkeli, M.; Mäkelä, T.; Serimaa, R.; ten Brinke, G.; Ikkala, O. *Science* **1998**, *280*, 557–560.

(98) Ruokolainen, J.; Saariaho, M.; Ikkala, O.; Ten Brinke, G.; Thomas, E.; Torkkeli, M.; Serimaa, R. *Macromolecules* **1999**, *32*, 1152–1158.

(99) Ruokolainen, J.; Tanner, J.; Ikkala, O.; Ten Brinke, G.; Thomas, E. *Macromolecules* **1998**, *31*, 3532–3536.

(100) Ikkala, O.; ten Brinke, G. *science* **2002**, *295*, 2407–2409.

(101) Fyfe, M. C.; Stoddart, J. F. *Accounts of chemical research* **1997**, *30*, 393–401.

(102) Kato, T.; Frechet, J. M. *Macromolecules* **1989**, *22*, 3818–3819.

(103) Ruokolainen, J.; Ten Brinke, G.; Ikkala, O.; Torkkeli, M.; Serimaa, R. *Macromolecules* **1996**, *29*, 3409–3415.

(104) Bai, P.; Kim, M. I.; Xu, T. *Macromolecules* **2013**, *46*, 5531–5537.

(105) Zhao, Y.; Thorkelsson, K.; Mastroianni, A. J.; Schilling, T.; Luther, J. M.; Rancatore, B. J.; Matsunaga, K.; Jinnai, H.; Wu, Y.; Poulsen, D., et al. *Nature materials* **2009**, *8*, 979.

(106) Rancatore, B. J.; Mauldin, C. E.; Tung, S.-H.; Wang, C.; Hexemer, A.; Strzalka, J.; Fréchet, J. M.; Xu, T. *ACS nano* **2010**, *4*, 2721–2729.

(107) Rancatore, B. J.; Mauldin, C. E.; Freffdfffdchet, J. M.; Xu, T. *Macromolecules* **2012**, *45*, 8292–8299.

(108) Chen, H.; Choi, J.-H.; Salas-de la Cruz, D.; Winey, K. I.; Elabd, Y. A. *Macromolecules* **2009**, *42*, 4809–4816.

(109) Rollet, A.-L.; Porion, P.; Vaultier, M.; Billard, I.; Deschamps, M.; Bessada, C.; Jouvensal, L. *The Journal of Physical Chemistry B* **2007**, *111*, 11888–11891.

(110) Huddleston, J. G.; Visser, A. E.; Reichert, W. M.; Willauer, H. D.; Broker, G. A.; Rogers, R. D. *Green chemistry* **2001**, *3*, 156–164.

(111) Zhang, S.; Sun, N.; He, X.; Lu, X.; Zhang, X. *Journal of physical and chemical reference data* **2006**, *35*, 1475–1517.

(112) Lee, M.; Choi, U. H.; Wi, S.; Slebodnick, C.; Colby, R. H.; Gibson, H. W. *Journal of Materials Chemistry* **2011**, *21*, 12280–12287.

(113) Tromp, R. A.; van Ameijde, S.; Pütz, C.; Sundermann, C.; Sundermann, B.; von Frijtag Drabbe Künzel, J. K.; IJzerman, A. P. *Journal of medicinal chemistry* **2004**, *47*, 5441–5450.

(114) Giessen, B. C.; Gordon, G. E. *Science* **1968**, *159*, 973–975.

(115) Hashimoto, T.; Suehiro, S.; Shibayama, M.; Sauo, K.; Kawai, H. *Polymer Journal* **1981**, *13*, 501.

(116) Henrich, B.; Bergamaschi, A.; Broennimann, C.; Dinapoli, R.; Eikenberry, E.; Johnson, I.; Kobas, M.; Kraft, P.; Mozzanica, A.; Schmitt, B. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **2009**, *607*, 247–249.

(117) Dinapoli, R.; Bergamaschi, A.; Henrich, B.; Horisberger, R.; Johnson, I.; Mozzanica, A.; Schmid, E.; Schmitt, B.; Schreiber, A.; Shi, X., et al. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **2011**, *650*, 79–83.

(118) Koch, M.; Bordas, J. *Nuclear Instruments and Methods in Physics Research* **1983**, *208*, 461–469.

(119) Amenitsch, H.; Bernstorff, S.; Laggner, P. *Review of scientific instruments* **1995**, *66*, 1624–1626.

(120) Feigin, L.; Svergun, D. I., et al., *Structure analysis by small-angle X-ray and neutron scattering*; Springer: 1987; Vol. 1.

(121) Prescher, C.; Prakapenka, V. B. *High Pressure Research* **2015**, *35*, 223–230.

(122) Rodriguez-Navarro, A. B. *Journal of Applied Crystallography* **2006**, *39*, 905–909.

(123) Kiapour, M. H.; Yager, K.; Berg, A. C.; Berg, T. L. In *IEEE Winter Conference on Applications of Computer Vision*, 2014, pp 933–940.

(124) Silla, C. N.; Freitas, A. A. *Data Mining and Knowledge Discovery* **2011**, *22*, 31–72.

(125) Schriber, E. A.; Popple, D. C.; Yeung, M.; Brady, M. A.; Corlett, S.; Hohman, J. N. *ACS Applied Nano Materials* **2018**.

(126) Roisnel, T.; Rodríquez-Carvajal, J. In *Materials Science Forum*, 2001; Vol. 378, pp 118–123.

(127) Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. *arXiv preprint arXiv:1902.10811* **2019**.

(128) Girshick, R. In *Proceedings of the IEEE international conference on computer vision*, 2015, pp 1440–1448.

(129) Ren, S.; He, K.; Girshick, R.; Sun, J. In *Advances in neural information processing systems*, 2015, pp 91–99.

(130) He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. In *Proceedings of the IEEE international conference on computer vision*, 2017, pp 2961–2969.

(131) Dai, J.; Li, Y.; He, K.; Sun, J. In *Advances in neural information processing systems*, 2016, pp 379–387.

(132) Rasmussen, C. E. In *Advanced lectures on machine learning*; Springer: 2004, pp 63–71.

(133) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. *Proceedings of the IEEE* **1998**, *86*, 2278–2324.

(134) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. In *Advances in Neural Information Processing Systems 25*, Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc.: 2012, pp 1097–1105.

(135) Snoek, J.; Larochelle, H.; Adams, R. P. In *Advances in neural information processing systems*, 2012, pp 2951–2959.

(136) Ling, J.; Hutchinson, M.; Antono, E.; DeCost, B.; Holm, E. A.; Meredig, B. *Materials Discovery* **2017**, *10*, 19–28.

(137) Pelt, D. M.; Sethian, J. A. *Proceedings of the National Academy of Sciences* **2018**, *115*, 254–259.

(138) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. *Chemistry of Materials* **2016**, *28*, 7324–7331.

(139) Ushizima, D.; Yang, C.; Venkatakrishnan, S.; Araujo, F.; Silva, R.; Tang, H.; Mascarenhas, J. V.; Hexemer, A.; Parkinson, D.; Sethian, J. In *Applied Imagery Pattern Recognition Workshop (AIPR), 2016 IEEE*, 2016, pp 1–12.

(140) Zhang, L.; Shum, H. P.; Shao, L. *IEEE Transactions on Image Processing* **2017**, *26*, 969–981.

(141) Ekeberg, T.; Engblom, S.; Liu, J. *The International Journal of High Performance Computing Applications* **2015**, *29*, 233–243.

(142) Yoneda, Y. *Phys. Rev.* **1963**, *131*, 2010–2013.

(143) Douarre, C.; Schielein, R.; Frindel, C.; Gerth, S.; Rousseau, D. *Journal of Imaging* **2018**, *4*, 65.

(144) Deyhle, H.; White, S.; Botta, L.; Liebi, M.; Guizar-Sicairos, M.; Bunk, O.; Müller, B. *Journal of Imaging* **2018**, *4*, 81.

(145) Kiapour, M. H.; Yager, K.; Berg, A. C.; Berg, T. L. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp 933–940.

(146) Wang, B.; Yager, K.; Yu, D.; Hoai, M. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, 2017, pp 697–704.

(147) Li, Y.; Cheng, W.; Yu, L. H.; Rainer, R. *Phys. Rev. Accel. Beams* **2018**, *21*, 054601.

(148) Franke, D.; Jeffries, C. M.; Svergun, D. I. *Biophysical journal* **2018**, *114*, 2485–2492.

(149) Rossouw, D.; Burdet, P.; de la Penfffdfffda, F.; Ducati, C.; Knappett, B. R.; Wheatley, A. E.; Midgley, P. A. *Nano letters* **2015**, *15*, 2716–2720.

(150) Laanait, N.; Zhang, Z.; Schlepütz, C. M. *Nanotechnology* **2016**, *27*, 374002.

(151) Timoshenko, J.; Lu, D.; Lin, Y.; Frenkel, A. I. *The journal of physical chemistry letters* **2017**, *8 20*, 5091–5098.

(152) Araujo, F. H.; Silva, R. R.; Medeiros, F. N.; Parkinson, D. D.; Hexemer, A.; Carneiro, C. M.; Ushizima, D. M. *Expert Systems with Applications* **2018**, *109*, 35–48.

(153) Chourou, S. T.; Sarje, A.; Li, X. S.; Chan, E. R.; Hexemer, A. *Journal of Applied Crystallography* **2013**, *46*, 1781–1795.

(154) Bradley, A. P. *Pattern recognition* **1997**, *30*, 1145–1159.

(155) Goodfellow, I.; Bengio, Y.; Courville, A., *Deep Learning*; MIT Press: 2016.

(156) Wold, S.; Esbensen, K.; Geladi, P. *Chemometrics and intelligent laboratory systems* **1987**, *2*, 37–52.

(157) Liu, S.; Li, J.; Bennett, K. C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T. *J. Phys. Chem. Lett.* **2019**, 4558–4565.

(158) Martineau, C. *Solid state nuclear magnetic resonance* **2014**, *63*, 1–12.

(159) Bryce, D. L. *IUCrJ* **2017**, *4*, 350–359.

(160) Shahid, S. A.; Bardiaux, B.; Franks, W. T.; Krabben, L.; Habeck, M.; van Rossum, B.-J.; Linke, D. *Nature methods* **2012**, *9*, 1212.

(161) Macholl, S.; Tietze, D.; Buntkowsky, G. *CrystEngComm* **2013**, *15*, 8627–8638.

(162) Baias, M.; Widdifield, C. M.; Dumez, J.-N.; Thompson, H. P.; Cooper, T. G.; Salager, E.; Bassil, S.; Stein, R. S.; Lesage, A.; Day, G. M., et al. *Physical Chemistry Chemical Physics* **2013**, *15*, 8069–8080.

(163) Pickard, C. J.; Mauri, F. *Physical Review B* **2001**, *63*, 245101.

(164) Hartman, J. D.; Balaji, A.; Beran, G. J. O. *Journal of Chemical Theory and Computation* **2017**, *13*, PMID: 29139294, 6043–6051.

(165) Shen, Y.; Bax, A. *Journal of biomolecular NMR* **2007**, *38*, 289–302.

(166) Shen, Y.; Bax, A. *Journal of biomolecular NMR* **2010**, *48*, 13–22.

(167) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *Journal of biomolecular NMR* **2003**, *26*, 215–240.

(168) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. *Journal of biomolecular NMR* **2011**, *50*, 43.

(169) Kohlhoff, K. J.; Robustelli, P.; Cavalli, A.; Salvatella, X.; Vendruscolo, M. *Journal of the American Chemical Society* **2009**, *131*, 13894–13895.

(170) Leclaire, J.; Poisson, G.; Ziarelli, F.; Pepe, G.; Fotiadu, F.; Paruzzo, F. M.; Rossini, A. J.; Dumez, J.-N.; Elena-Herrmann, B.; Emsley, L. *Chemical science* **2016**, *7*, 4379–4390.

(171) Cuny, J.; Xie, Y.; Pickard, C. J.; Hassanali, A. A. *Journal of chemical theory and computation* **2016**, *12*, 765–773.

(172) Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharaki, E. I. *PeerJ* **2018**, *6*, e4750.

(173) Kuzminykh, D.; Polykovskiy, D.; Kadurin, A.; Zhebrak, A.; Baskov, I.; Nikolenko, S.; Shayakhmetov, R.; Zhavoronkov, A. *Molecular pharmaceutics* **2018**, *15*, 4378–4385.

(174) Torng, W.; Altman, R. B. *BMC bioinformatics* **2017**, *18*, 302.

(175) Ryczko, K.; Mills, K.; Luchak, I.; Homenick, C.; Tamblyn, I. *Computational Materials Science* **2018**, *149*, 134–142.

(176) He, K.; Zhang, X.; Ren, S.; Sun, J. In *European conference on computer vision*, 2016, pp 630–645.

(177) Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K. Q. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp 4700–4708.

(178) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. *Acta Crystallographica Section B* **2016**, *72*, 171–179.

(179) Ramachandran, P.; Varoquaux, G. *Computing in Science & Engineering* **2011**, *13*, 40–51.

(180) Stomberg, R.; Li, S.; Lundquist, K.; Albinsson, B. *Acta Crystallographica Section C: Crystal Structure Communications* **1998**, *54*, 1929–1934.

(181) Hartman, J. D.; Kudla, R. A.; Day, G. M.; Mueller, L. J.; Beran, G. J. O. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21686–21709.

(182) Chollet, F. Keras., `https://keras.io`, 2015.

(183) Martín Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems., Software available from tensorflow.org, 2015.

(184) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. **2017**.

(185) Ioffe, S.; Szegedy, C. *arXiv preprint arXiv:1502.03167* **2015**.

(186) Nair, V.; Hinton, G. E. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp 807–814.

(187) Clementi, E.; Raimondi, D.-L. *The Journal of Chemical Physics* **1963**, *38*, 2686–2689.

(188) Brown, P.; Fox, A.; Maslen, E.; OfffdfffdfffdKeefe, M.; Willis, B. *International tables for crystallography* **2006**.

(189) Leemann, S. *Nucl. Instr. and Meth. A* **2018**, 883.

(190) Chrin, J.; Schmidt, T.; Streun, A.; Zimoch, D. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **2008**, *592*, 141–153.

(191) Wall, E. *Proceedings of 2011 Particle Accelerator Conference* **2011**, 1262.

(192) Leemann, S. *6th International Particle Accelerator Conference, IPAC2015* **2011**, 1262.

(193) LeCun, Y.; Bengio, Y.; Hinton, G. *Nature* **2015**, *521*, 436.

(194) Steier, C. *Proceedings of the 2003 Particle Accelerator Conference* **2003**, 3213.