npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01402-7

Deep learning for symmetry classification using sparse 3D electron density data for inorganic compounds

Check for updates

Seonghwan Kim[1,4], Byung Do Lee[1,4], Min Young Cho[1], Myoungho Pyo[2], Young-Kook Lee[3], Woon Bae Park[2] & Kee-Sun Sohn[1]

We report a novel deep learning (DL) method for classifying inorganic compounds using 3D electron density data. We transform Density Functional Theory (DFT)-derived CHGCAR files from the Materials Project (MP) and experimental data from the Inorganic Crystal Structure Database (ICSD) into point clouds and sparse tensors, optimized for use in DL models such as PointNet and Sparse 3D CNN. This approach effectively overcomes the limitations of handling the dense 3D data, a common challenge in DL. Contrasting with traditional 1D or 2D X-ray diffraction (XRD) patterns that necessitate complex reciprocal space analysis, our method utilizes 3D density data for direct interpretation in real lattice space. This shift significantly enhances classification accuracy, outperforming traditional XRD-driven DL methods. We achieve accuracies of 97.28%, 90.77%, and 90.10% for crystal system, extinction group, and space group classifications, respectively. Our 3D electron density-based DL approach not only showcases improved accuracy but also contributes a more intuitive and effective framework for materials discovery.

Recent advancements in the field of deep learning (DL) have catalyzed innovative approaches to materials science[1–12]. In this regard, there have been a number of attempts to achieve DL-driven materials analyses and characterizations using one-dimensional (1D) powder XRD patterns. The 1D powder XRD patterns have long been considered as an archetypal crystal structure data and thereby the DL approach has been also focused on the 1D powder XRD pattern for inorganic materials due to its popularity and familiarity in the materials science society[13–38], although several successful DL approaches to 2D (X-ray or electron) diffraction patterns have also been reported[39–43]. When considering the 1D powder XRD-based approaches, the symmetry classification[13–21,44], the phase identification (demixing)[22–28,45,46], the property regression[21,29,30], and the clustering and XRD generation[21,31–33] have been successful for inorganic materials, along with some previous pioneering machine learning (ML)-based XRD analyses[34–38]. Although the XRD-driven ML/DL approach is yet to be outperforming the traditional knowledge-driven approaches based on well-established, rule-based software packages[47–53], it would be drastically improving soon along with the rapid progress in DL techniques.

The majority of the currently booming XRD-driven ML/DL approaches employ convolutional neural network (CNN)-based methodologies, treating the powder XRD pattern as nothing but a conventional 1D image, with characteristic feature extraction occurring across the CNN layers. While 1D powder XRD data has proven instrumental for symmetry classification and phase identification, it has shown limited efficacy in property regression[20]. In contrast, the crystal graph convolutional neural network (CGCNN) and its subsequent variants have demonstrated aptitude in the property regression domain[11,54–56]. Namely, the former excels in symmetry classification but falters in property regression, while the latter exhibits the opposite trend[20]. This trade-off in performance between XRD-based CNN and descriptor-based CGCNN would not seem to be eliminated.

The CGCNN integrates local structural data, represented through nodes and edges. These nodes and edges encapsulate knowledge-driven descriptors, detailing constituent elements and bond characteristics[54,55]. Furthermore, there has been very recently introduced an improved graph-based model, the Connectivity Optimized Nested Graph Network (coGN)[57], which incorporates all crystal symmetry operations for an asymmetric unit and substantially augments property regression capabilities for inorganic crystals leading to the state-of-the-art (SOTA) record in benchmarks[58], which is presented in Supplementary Fig. 1. Nonetheless, the coGN's

[1]Nanotechnology & Advanced Materials Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul, 143-747, South Korea. [2]Department of Advanced Components and Materials Engineering, Sunchon National University, 291-19 Jungang-ro, Sunchon, Chonnam, 540-742, South Korea. [3]Department of Materials Science and Engineering, Yonsei University, Seoul, 03722, Republic of Korea. [4]These authors contributed equally: Seonghwan Kim, Byung Do Lee.
e-mail: wbpark@scnu.ac.kr; kssohn@sejong.ac.kr

proficiency in symmetry classification still remains notably inferior to the XRD-based CNN.

In this study, our primary emphasis is on exploring alternative DL models for symmetry classification and property regression. We deliberately eschew the use of XRD-based CNN and CGCNN-related approaches, implying that the 1D powder XRD data and graph-type data, typically employed in the previous models, are not the focal points of our investigation. Instead, we adopt the 3D electron density (charge density) in the real lattice as training data for our DL models. To adeptly handle this 3D electron density data, we incorporate DL architectures such as PointNet[59] and Sparse 3D CNN[60–62]. Although conventional 3D CNNs have been utilized for property regression within specific material systems, including several face-centered cubic (fcc) materials[63], Sparse 3D CNNs, which are capable of handling large-scale datasets, have not been extensively focused on in the field of materials science, Sparse 3D CNNs that can handle large scale datasets have not been a focal point in materials science. In contrast, there have been PointNet-based DL approaches driven by 3D crystal structural data. For example, PointNet has been applied in crystal structure-related studies, particularly in the context of molecular dynamics (MD)[64,65]. In recent developments like DeepCrysTet[66], crystal structures have been innovatively modeled using 3D tetrahedral meshes, a process facilitated by Delaunay tetrahedralization. Following this, a new methodology known as Neural Structure Field (NeSF) emerged, addressing the limitations associated with voxelization[67]. However, these previous 3D methodologies considered each atom as a point, forming a point cloud with atomic coordinates and features. The concept of treating electron density at arbitrary lattice locations as a point cloud has not been previously explored. Our electron-density-driven approach marks a significant breakthrough in the realm of DL-driven 3D crystal structure analysis by emphasizing electron density as a pseudo continuous entity, rather than merely considering it as discrete atomic positions within the lattice.

Focusing on 3D electron density data may be more advantageous than depending solely on atomic arrangement data, despite both being derived from theoretical crystal structure solutions (such as atomic coordinates, occupancy, thermal factors, and lattice parameters), originally obtained from experimental data diffracted in reciprocal space. Utilizing 3D electron density data would be particularly practical, especially in light of experimental techniques like electron tomography[68–70]. Our primary goal extends beyond merely predicting material properties from known crystal structure solutions. Instead, we aim to identify unknown materials through their experimental electron density patterns. While the techniques for measuring 3D electron density directly are still in progress and not yet fully established, the DL approach utilizing 3D electron density data remains promising due to the ongoing progress in developing these measurement methodologies.

We compiled two unique datasets of 3D electron density: the first, derived from CHGCAR files based on Density Functional Theory (DFT) calculations from the Materials Project (MP)[71], and the second, obtained from crystal structure solution data from the Inorganic Crystal Structure Database (ICSD)[72], predominantly validated experimentally. These datasets were transformed into point cloud and sparse tensor formats for use with PointNet and Sparse 3D convolutional networks, respectively. Our research conducted a thorough comparative analysis of various DL models including Fully Convolved Neural Network (FCN)[73], CGCNN, coGN, PointNet, and Sparse 3D CNN, with diverse training data types such as 1D powder XRD, graph-type data, and 3D electron density data. This comparative study was systematically detailed, with particular emphasis on the finding that the Sparse 3D CNN model, especially when utilizing ICSD-derived 3D electron density data, demonstrated superior performance compared to other models.

Results

Previous XRD-driven DL approaches

Symmetry classification in crystallography involves identifying the crystal system (seven classes), extinction group (101 classes), and space group (230 classes) by analyzing powder XRD patterns using CNN-based deep learning models. Schuetzke et al.[19] expanded the scope of symmetry classification to encompass structure type classification, incorporating 1D spectral data from XRD, Nuclear Magnetic Resonance (NMR), and Raman scattering. Structure types, or prototype structures, are categorized by the ICSD following guidelines by Allmann and Hinek[74]. Our study, however, focuses on classifying the crystal system, extinction group, and space group. The crystal system is typically identified from powder XRD patterns using commercial indexing tools like ITO[47], TREOR[48], DICVOL[49], McMaille[50], EXPO[51], and X-CELL[52]. Space group determination is more complex than simple indexing, as systematic absences in powder XRD patterns may suggest multiple space group candidates within an extinction group. While powder XRD patterns do not guarantee space group identification for inorganic compounds, single crystal diffraction data usually enable this. Powder XRD data, therefore, are mainly used to ascertain the crystal system, extinction group, and lattice parameter. In parallel to traditional XRD analysis, DL approaches face similar challenges for space group determination. While test accuracy for crystal system classification exceeds 90% across various DL-driven approaches[13–21,44], accuracy for space group identification often falls far below 90%. Despite the difficulty in making direct comparisons due to varying training and testing datasets, the state-of-the-art (SOTA) test accuracy for powder-XRD-based crystal system classification was reported at 94.9%[13]. However, this high accuracy was achieved using a researcher-curated dataset comprising only 150k ICSD entries (the number of inorganic compounds in the training dataset). Since this initial foray into DL for powder XRD analysis, numerous attempts have been made using larger datasets[14–21,44].

Suzuki et al.[16] made a significant observation that most misclassified cases were in lower symmetry groups such as orthorhombic, monoclinic, and triclinic, a finding later reaffirmed by Lee et al.[20]. Interestingly, Park et al.[13] had excluded low symmetry entries from their dataset, unaware of Suzuki et al.'s[16] discovery, which likely inflated the test accuracy for crystal system classification. Therefore, the more realistic SOTA test accuracy for crystal system classification is around 93.06%[72], as determined using a more comprehensive dataset that includes almost all entries (197,131 inorganic compounds) from the ICSD as of 2023. This dataset has been utilized in two subsequent studies[20,21], where FCNs were applied for symmetry classification. FCNs diverge from the conventional CNNs with several fully connected layers commonly used in previous DL-driven XRD analyses. The FCN method proved more effective than the conventional CNN approach, achieving higher performance with even fewer parameters[20]. In these studies, the 1D powder XRD-driven FCN approach achieved state-of-the-art (SOTA) test accuracy of 93.06% for crystal system classification, considered the current upper limit[21]. Interestingly, despite the transformer's known superiority over CNN in certain image-related DL tasks[75,76], it did not surpass FCN in XRD-driven symmetry classification[20].

Table 1 summarizes recent DL-driven results in symmetry classification and phase identification. Generalizability refers to whether the training dataset encompasses the entirety of inorganic compounds or just a small portion. Most phase identification studies have used a narrow range of materials within a limited composition space, leading to low generalizability. For example, Lee et al.'s[22] study focused on the Li–Al–Si–O–N composition system, Oviedo et al.'s[14] on thin-film metal halides limited to seven space-group categories, Szymanski et al.'s[26] on the Li–Mn–Ti–O–F composition space, Massuyeau et al.'s[24] on chlorides, bromides, and iodides in perovskite or non-perovskite structures, and Maffettone et al.'s[27] on specific systems like $BaTiO_3$ phase transition, ADTA crystal structure prediction, and phase mapping of the Ni–Co–Al alloy system. In contrast, DL models for symmetry classification tend to cover a broader range of inorganic materials registered in the ICSD. Studies by Park et al.[13], Vecsei et al.[15], Suzuki et al.[16], Lee et al.[20,21], and Salgado et al.[44] have almost fully encompassed inorganic compounds from the ICSD for training and testing their DL models in symmetry classification.

The perturbation column in Table 1 exhibits various attempts at perturbation to achieve more realistic XRD patterns when simulating XRD patterns. Oviedo et al.[14] considered physics-informed data perturbations

Table 1 | The recent advances in DL-driven crystallographic symmetry classification and phase identification

		Perturbation Incorporation	Data size	Problem size		Test accuracy (%)
Authors	Data Generalizability	Preferred orientation, peak shifting, peak broadening, noise, etc	# of XRD patterns	# of classes	Diffraction Type	Top one Accuracy
Park et al.[13]	Generalizable	Fair	150,000	7/101/230	Conventional 1D	94/84/81
Ziletti et al.[40]	Narrow (eight elemental solids)	N/A	100,000	7	Virtual 2D	100
Oviedo et al.[14]	Narrow (thin-film metahalides)	Good	2000	7	Conventional 1D	99.1
Vecsei et al.[15]	Generalizable	Fair	128,404	7/230	Conventional 1D	85/76
Lee et al.[22]	Narrow (Li-Sr-Al-O)	Fair	188,000	9000	Conventional 1D	99
Suzuki et al.[16]	Generalizable	Fair	199,391	7/230	Conventional 1D	92/80
Maffettone et al.[27]	Narrow (Ni-Co-Al)	Good	600,000	6	Conventional 1D	94.7
Tiong et al.[42]	Pseudo- Generalizable	N/A	108,658	72	Virtual 2D	80.12
Aguiar et al.[43]	Generalizable	N/A	571,340	7	Azimuthal integration profile 1D	85.87
Szymanski et al.[26]	Narrow (Li-Mn-Ti-O-F)	Good	21,000	140	Conventional 1D	94
Wang et al.[28]	Narrow (Metal-Organic Frameworks)	Good	72,864	1012	Conventional 1D	N/A 96.7 (Top-5 Acc.)
Massuyeau et al.[24]	Narrow (Perovskite & non-Perovskite)	N/A	998	2	Conventional 1D	>85
Lee et al.[20]	Generalizable	N/A	189,476	7/101/230	Conventional 1D	92/81/79
Lee et al.[21]	Generalizable	Good	1,974,760	7/101/230	Conventional 1D	93/87/84
Schuetzke et al.[18]	Narrow (Iron ore & cement)	Good	100,000~500,000	28/76	Conventional 1D	95/99
Schuetzke et al.[19]	Pseudo- Generalizable	Good	30,000	500	Conventional 1D	98.9
Salgado et al.[44]	Generalizable	Good (No perturbations)	1,200,000 (171,006)	7/230	Conventional 1D	86/77 (96/94)

This table outlines key attributes of various DL approaches, encompassing data generalizability, perturbation incorporation, data size, problem size, and diffraction type. 'Data generalizability' refers to the diversity range of inorganic material systems used, from narrow to broad. Datasets focused on specific chemical compositions are categorized as having a 'narrow' range. The 'perturbation incorporation' column evaluates the variety of perturbations applied in data simulation to assess the method's robustness. 'Data size' and 'problem size' columns denote the number of data points and labels utilized for training the DL models. The 'diffraction type' column clarifies the fundamental nature of the data employed. Lastly, the 'test accuracy' column displays the outcome accuracies as top-one accuracy.

common in thin film samples, like texture and epitaxial strain. Maffettone et al.[27] used an ensemble of 50 CNN-based classifiers to efficiently handle perturbations without combinatorial explosion. Wang et al.[28] synthesized new training spectra by merging perturbations extracted from experimental data with theoretical spectra. Szymanski et al.[16] employed physics-informed data augmentation, incorporating perturbations into synthetic XRD patterns considering domain size, lattice parameters, and random preferred orientation. In their method, removing identified phases from blended XRD aided in determining all constituent phases[26]. Schuetzke et al.[18] aimed for well-perturbed XRD patterns resembling real-world samples and systematically studied the impact of various perturbation parameters on the F1 score. Additionally, some studies have directly incorporated perturbations into DL models rather than relying on perturbed data augmentation. For instance, Chen et al.[25] introduced a deep reasoning network (DRN) that merges DL with constraint reasoning to include prior scientific knowledge. This approach required only a minimal amount of unlabeled data to address perturbations, as these are integrated into the DRN's latent embedding layer, reducing the need for a large number of synthetic XRD patterns with perturbations. The data size column in Table 1 shows the total number of XRD patterns used in the training dataset, along with other columns exhibiting the number of classes and their corresponding hold-out test accuracies. Most high-accuracy results in phase identification and symmetry classification were achieved using conventional CNN models, including inception nets[77] and residual nets[78], while Vecsei et al.[15] argued for the superiority of a basic ANN with a multi-layer perceptron (MLP) architecture over CNNs. However, Lee et al.[20,21] reported superior performances with FCNs compared to the others.

We must acknowledge that the accuracy levels reported for methods utilizing the ICSD and MP databases might be somewhat overstated, with each approach purportedly achieving test accuracies above 90%. This notably high performance in symmetry classification could stem from the extensive redundancy within the materials dataset. Despite this, addressing this issue is practically challenging, even with the adoption of advanced validation strategies like leave-one-out cross-validation[79]. The only feasible approach appears to be manually identifying and removing highly redundant (duplicated) entries on a subjective basis when partitioning the dataset into training, validation, and test sets. A clear and objective criterion for measuring similarity between entries is absent. Moreover, it is challenging to identify a representative structure among those deemed similar, even if such a criterion were available. However, this overestimation due to redundancy might not significantly impact the evaluation of ML model performance, especially since our primary focus was on comparing the relative effectiveness of different ML models.

3D electron density data-driven DL approach

Traditionally, 1D or 2D diffraction patterns are represented in reciprocal space, an essential approach given that experimental methods such as X-ray, neutron, and electron diffraction rely on interpretation in reciprocal space. While reciprocal interpretation is practical, understanding crystal structures in real space, which avoids the need for forward and backward Fourier transformation, is often more intuitive. The historical preference for reciprocal space in crystal structure analysis tools and methodologies stems from a lack of high-resolution real-space imaging techniques. However, recent advances in experimental methods are enhancing the feasibility of real-space imaging. For example, Scanning Transmission Electron Microscopy (STEM) provides high-resolution real-space imaging, which is vital in materials science and nanotechnology. Contrary to traditional diffraction methods that provide data in reciprocal space, like spatial frequencies or wavevectors, STEM techniques make real-space imaging available, despite their current limitation to 2D. Although this evolving capability does not yet indicate a bright future for real-space 3D imaging techniques, it might be hopefully available by advancing beyond current incomplete 3D technologies such as tomography-based 3D STEM imaging[68–70] and 3D XRD imaging[80] techniques. Following this, the implementation of a 3D-electron-density-driven DL approach promises to be extremely beneficial, even though experimental 3D data have yet to be realized.

Ziletti et al.[40] were the first to use 3D crystal data as a descriptor, projecting it onto 2D reciprocal space to create a 'two-dimensional diffraction fingerprint (DF)', which served as an effective pseudo-3D descriptor for deep learning. However, DF remains within the realm of reciprocal space, akin to conventional 2D diffraction, and requires expensive analytical interpretation to reproduce the actual crystal structure from 2D DF data, despite its benefits such as size invariance and defect accommodation[40]. Building on the groundwork laid by Ziletti et al.[40], Tsuruta et al.[65] and Chiba et al.[66] have also employed 3D crystal data as descriptors for representing crystal structures, specifically for application in DL approaches. In contrast to these previous works, which are based on discrete atomic positions, our approach focuses on treating electron density as a pseudo continuous field. Sparse 3D electron density in real space was marked as the first real-space crystal structure descriptor for scaled DL. Even though Zhao et al.[63] and Saha et al.[81] have previously introduced small-scale 3D electron density data for DL applications, they utilized conventional 3D CNNs without incorporating sparsity. The electron density data were sourced through two pathways: one from DFT-calculated CHGCAR files in VASP[82–84], obtained from the Materials Project (MP) database[71], and the other from ICSD structure solutions[72] combined with atomic scattering factors[85]. The ICSD solution data-based electron density was computed using our in-house Python code and the FullProf[86] and VESTA[87] software API.

The optimal DL model for effectively handling the 3D electron density descriptor in real space is the 3D convolutional neural network (CNN)[88]. However, using unit cells or supercells from every ICSD and MP entry as input for 3D CNNs introduces two major challenges. The first is the varying cell shapes across different crystal systems, and the second is the disparity in unit cell size between small and large periodicities. One solution is to use a large, uniformly shaped input volume unit, such as extracting a 20 × 20 × 20 Å^3 cube from each ICSD and MP entry. This approach ensures consistency in volume and voxel count, and thereby a requirement for traditional dense 3D CNN models can be met. However, this method significantly increases computational overhead. Additional issue would arise such that a standardized input volume might encompass multiple unit cells for entries with smaller cells, while only capturing a fraction of a cell for entries with larger cells. This discrepancy can disrupt the seamless integration of lattice periodicity information during the DL model training process. Recognizing the limitations of traditional dense 3D CNNs, we have adopted PointNet[59] and Sparse 3D Convolutional Neural Networks (Sparse 3D CNNs)[60–62] as more practical alternatives. The point cloud and sparse tensor data formats, optimized for use in PointNet and Sparse 3D CNN models, adeptly tackle two major issues: the issue of varying shapes and sizes of unit cells and the computational challenge. Derived from downsampling the originally dense electron density data, these data formats inherently become sparser, which leads to more efficient management of computational load.

Figure 1 schematically illustrates a summary of the proposed 3D-electron-density-driven DL approach, and the schematic representations of both PointNet and Sparse 3D CNN are depicted in Fig. 2, with their detailed architectures illustrated in Supplementary Fig. 2. PointNet is adept at analyzing point cloud data, which captures 3D objects as discrete points in space, extracting features and patterns crucial for tasks like object recognition, classification, and segmentation. Its utility is especially notable in computer vision and 3D scene understanding, allowing for direct interaction with 3D spatial data. A key advantage of PointNet is its size invariance; it can process point clouds irrespective of the points' order or quantity. The size invariance is achieved through a symmetric function that assesses each point independently, facilitating efficient handling of point clouds of varying sizes. This characteristic is particularly beneficial for tasks like symmetry classification in crystal structure analysis, as it allows PointNet to robustly adapt to the varying number of points typically seen in electron density data

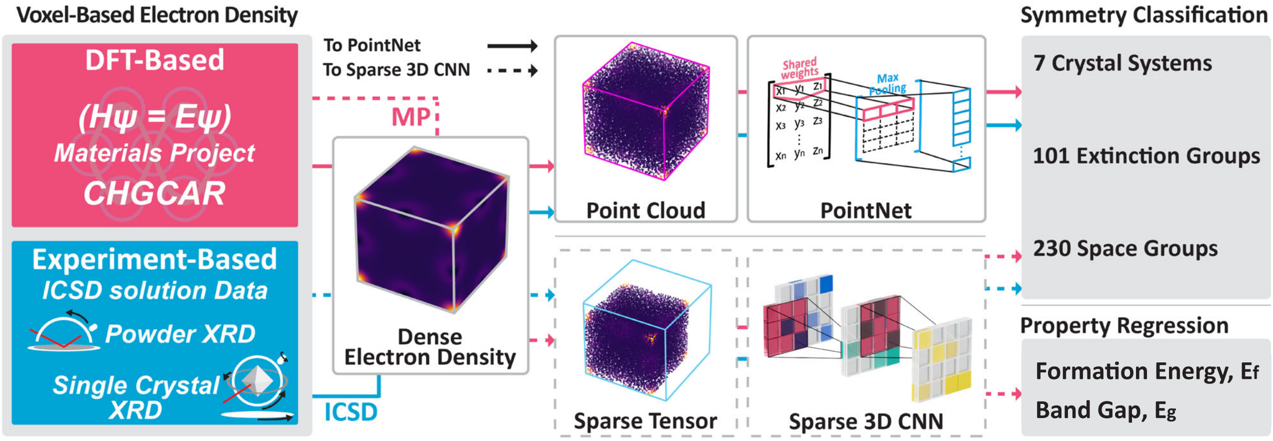


Fig. 1 | Schematic illustration of symmetry classification and property regression using 3D electron Density. The two leftmost cells in the image illustrate the methods of generating 3D electron density. The red section represents 3D electron density from the MP database, derived from ab-initio DFT calculations, while the blue sections depict 3D electron density from the ICSD, generated using structure factors determined from experimental structure analyses. The middle gray cells illustrate typical examples of point cloud and sparse tensor formats, which are transformed from dense voxel-based electron density data and utilized in PointNet and Sparse 3D CNN, respectively. The DL model illustration demonstrates the use of either point cloud or sparse tensor data for symmetry classification, aiming to predict 7 crystal systems, 101 extinction groups, and 230 space groups. Notably, Sparse 3D CNN is exclusively used for property regression, specifically to predict formation energy (E_f) and band gap (E_g). The red and blue arrows in the image indicate the source of the 3D electron density data, with red for MP and blue for ICSD. Dashed and solid lines represent the PointNet and Sparse 3D CNN models, respectively.

for ICSD and MP entries. Thus, PointNet can handle variously sized unit cells from all ICSD and MP entries without needing the standardized input volume unit required by traditional dense 3D CNN models.

Sparse 3D CNNs excel in analyzing 3D data, especially when much of it is sparse or empty[60–62]. These networks are adept at processing and extracting features from occupied segments of 3D volumetric grids (voxels), which makes them ideal for tasks like 3D object detection, segmentation, and scene understanding. By disregarding empty spaces and utilizing sparse tensors, Sparse 3D CNNs minimize computational overhead, proving efficient in scenarios where not all parts of the 3D space contain relevant information. To comprehend the benefits of Sparse 3D CNNs in analyzing 3D crystal structures, it is essential to first understand the downsampling methods. We sampled data to create sparse tensors from the original, densely voxelized 3D data with a grid size of 0.06 Å (CHGCAR data from MP and electron density data from ICSD), which are considered as being pseudo continuous.

The posterior random sampling process results in a significant number of empty voxels in the resultant sparse tensor, which is clearly illustrated as ‘SparseTensor$_{\text{In}}$‘ in Fig. 2b, wherein only a non-zero voxel (p1) is present. Another aspect of this approach is the introduction of a sufficiently large virtual spatial frame (sparse tensor dimension), as shown in Supplementary Fig. 4. This conceptual frame can accommodate the largest unit cell found in the ICSD and MP databases, allowing for a range of unit cell sizes. The virtual frame is not fixed but rather imaginary. The size of the virtual frame is not a parameter required for Sparse 3D CNN execution; it was introduced to emphasize the size independence of the Sparse 3D CNN execution. In reality, the size of input 3D images has no limitation, as the index (v_{in} and v_{out} in Fig. 2b) can make convolution available only where the electron density is non-zero. The primary focus is on the non-zero elements indexed by their coordinates, ensuring that the overall size of the input grid does not significantly impact the processing. This confirms that Sparse 3D CNNs enable the management of inputs with different dimensions without requiring uniform resizing or padding, and thereby efficiently handle diverse crystal periodicities (various unit cell sizes).

In contrast to the fixed input volume of traditional dense 3D CNNs, this approach significantly reduces computational costs. Supplementary Fig. 4 showcases typical point cloud and sparse tensor data for PointNet and Sparse 3D CNN, featuring four different inorganic compounds with varying unit cell shapes and sizes. While Supplementary Fig. 4a and e presents a large cubic unit cell, Supplementary Figs. 4b–d and f–h depict smaller unit cells with isotropic, rod-shaped, and plate-shaped structures. Besides the downsampling methods described above, we have incorporated another method to clarify the sampling effect on the final DL performance. Grid-based sampling outperformed random sampling, such that the test accuracies were slightly enhanced to 97.57%, 91.09%, and 90.55% for crystal system, extinction group, and space group classifications, respectively. Supplementary Figs. 5 and 6 and Supplementary Note 1 describes details on both the downsampling methods (random and grid-based samplings) that we adopted in the study.

DL model performances

Figure 3 and Supplementary Table 1 present the results of symmetry classification for both the MP and the ICSD datasets. The MP dataset, referred to as 120k_MP, comprises 122,689 entries (both real and virtual inorganic compounds), validated through DFT calculations. For the purpose of model training, the 120k_MP dataset was formatted into four distinct types to accommodate various models: coGN, FCN, PointNet, and Sparse 3D CNN. Specifically, the dataset was structured as graph-structured data for coGN, a 1D X-ray diffraction (XRD) pattern for FCN, a point cloud for PointNet, and a sparse tensor for Sparse 3D CNN. In terms of hold-out test accuracy for symmetry classification, the models were ranked as follows: Sparse 3D CNN > PointNet > FCN > coGN. The F1 scores mirrored this ranking, with Sparse 3D CNN being the highest, followed by PointNet, FCN, and coGN. Additionally, the ICSD dataset, labeled as 190k_ICSD that comprises 195,300 entries (most of them are experimentally realized inorganic compounds), was evaluated with a primary focus on its impressive test accuracies and F1 scores, which surpassed those of the 120k_MP dataset. The model comparison for the 190k_ICSD dataset revealed a similar trend to that observed with the 120k_MP dataset, namely, Sparse 3D CNN > PointNet > FCN for test accuracy and F1 score.

Notably, coGN was not utilized for symmetry classification in the 190k_ICSD dataset for convenience. The ICSD comprises a significant number of real-world inorganic compounds, with a large portion featuring partially occupied (disordered) structures; approximately 120,000 of the around 200,000 ICSD entries exhibit this characteristic[72]. The node features in coGN encompass various atomic properties, including atomic number,

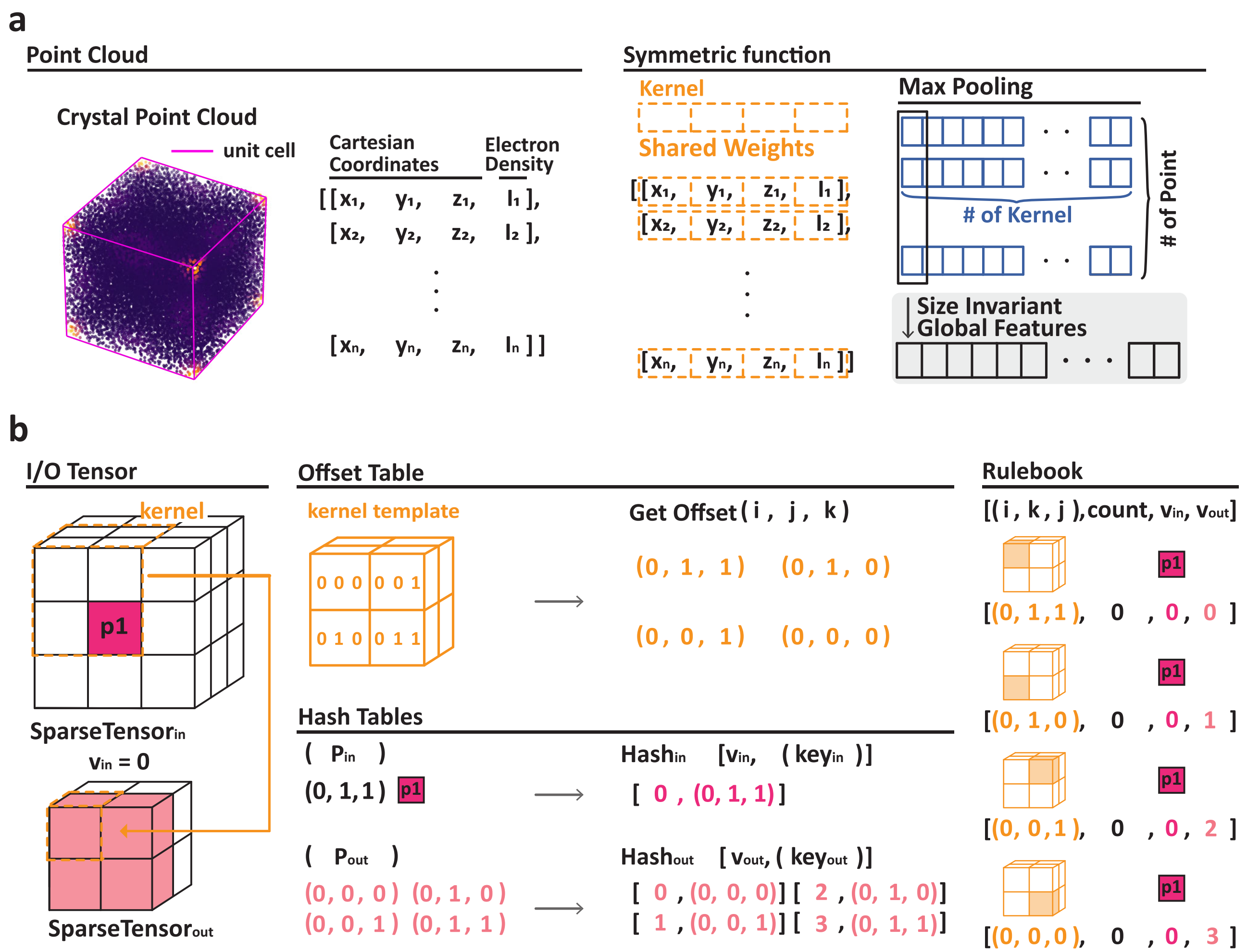


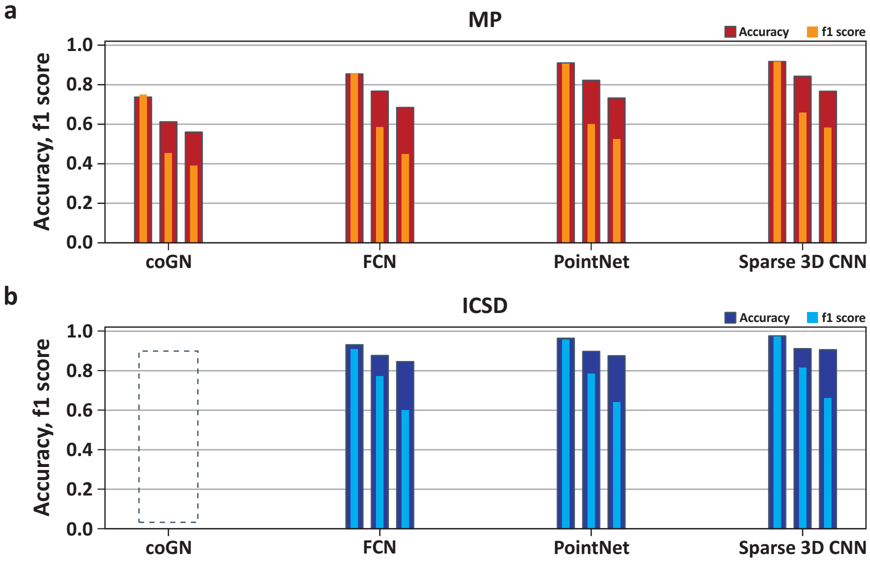
Fig. 2 | Overview of PointNet and 3D Sparse CNN architectures. a PointNet: Visualization of the PointNet architecture, focusing on the symmetric function. **b** Sparse 3D CNN Process: This illustration delineates the sparse convolutional process, beginning with rulebook creation and execution of sparse convolution. The visualization of the PointNet and 3D Sparse CNN architectures in Supplementary Fig. 2, and more detailed descriptions of the PointNet and 3D Sparse CNN functioning are presented in Supplementary Fig. 3.

atomic mass, atomic radius, electronegativity, ionization energy, and oxidation state. When processing a partially occupied structure, coGN necessitates the assignment of a weighted average for each attribute to every shared atomic position (node) in the graph. However, this approach of averaging node attributes may seem inadequate, despite the pioneering efforts of Chen et al.[56], who addressed disorder in their multi-fidelity graph network by weighted-averaging learned elemental embeddings, albeit for a limited set (278) of disordered inorganic materials. More critically, in terms of symmetry classification, addressing the disorder issue may not be beneficial, given that coGN's performance is inferior to other 3D electron density-driven DL models. Since coGN is specifically tailored for the MP dataset, which exclusively contains fully occupied structures[67], applying the ICSD dataset to coGN would be unsuitable.

Sparse 3D CNN, when trained on the 190k_ICSD dataset, achieved a SOTA test accuracy of 97.28% for crystal system classification. This result is significant as it was obtained from the largest dataset, encompassing a broad range of real-world inorganic materials, without any artificial pruning. This level of test accuracy, especially considering the dataset's size and generalizability, is more notable than the previous SOTA record of 94.99% reported by Park et al.[13], which was achieved after the removal of many low symmetry entries. Moreover, the Sparse 3D CNN achieved SOTA test accuracies in both extinction and space group classifications, surpassing 90%. This marks a significant milestone, as neither extinction nor space group classification has previously attained approximately 90% accuracy using any XRD-driven DL models. Specifically, the Sparse 3D CNN reached an accuracy of 90.77% in extinction group classification and 90.10% in space group classification.

With the symmetry classification complete, the focus shifts to a deeper interpretation of the results. One key aspect is the comparison between the 190k_ICSD and 120k_MP datasets. Contrary to expectations, every model trained on the 190k_ICSD dataset outperformed those on the 120k_MP dataset. This disparity might be mainly due to the different symmetry distribution of the datasets. The 120k_MP dataset has a distribution skewed more towards lower symmetry than the 190k_ICSD dataset, as depicted in Supplementary Fig. 7. In addition, the theoretical (virtual) structure entries in the 120k_MP dataset could have impacted test accuracy negatively. The 190k_ICSD dataset contains a significantly smaller proportion of theoretical entries (for instance, 922 virtual entries for ICSD collection codes 1~100,000) compared to the MP dataset, which has over 86,974 virtual entries out of 122,689 in total. Most entries in the 190k_ICSD are experimentally realized, adding to their reliability. According to the 2023 ICSD data[72], 91,099 entries are validated by single crystal XRD and 114,150 by powder XRD among the experimental structures in ICSD. This suggests that nearly half of the ICSD data are confirmed by highly reliable single crystal XRD, devoid of peak overlap complications. More importantly, the 120k_MP dataset gives electron density for primitive cells but the 190k_ICSD for Bravais lattice that contains more structural information.

Fig. 3 | Comparative performance metrics for crystallographic symmetry classification using DL Models. This figure presents bar plots that compare the performance of coGN, FCN, PointNet, and Sparse 3D CNN models in three symmetry classification categories: crystal system (left), extinction group (middle), and space group (right). For each category and model, the broader bars represent top-1 hold-out test accuracy, while the overlaid lighter bars indicate the F1 score, a measure of test precision. **a** Displays hold-out test accuracy results obtained using the 120k_MP dataset. **b** Shows hold-out test accuracy outcomes from the 190k_ICSD dataset. Note: The results for the coGN model with the 190k_ICSD dataset are excluded due to compatibility issues between the model and the dataset.



These considerations help rationalize the better performance of the ICSD dataset over the MP dataset.

In the 120k_MP dataset, all theoretical structures are assumed to be fully occupied. However, this assumption might be slightly bold when considering the real inorganic structures cataloged in the ICSD. The majority of the 86,974 virtual entries from the 120k_MP dataset feature somewhat unrealistic, fully-occupied, ordered structures that are not found in the 190k_ICSD dataset. Theoretically, these structures, if realized in the real world, could form partially occupied structures, which typically exhibit higher symmetry than their fully occupied counterparts. This potential discrepancy could explain why the 120k_MP dataset appears skewed toward lower symmetry compared to the ICSD dataset, as illustrated in Supplementary Fig. 7. Supporting this observation, the XRD-driven DL approach by Suzuki et al.[16] and the subsequent reconfirmation by Lee et al.[20,21] indicate that misclassified entries predominantly fall into lower symmetries, such as orthorhombic, monoclinic, and triclinic categories.

The second issue concerns the superiority of PointNet and Sparse 3D CNN over FCN in symmetry classification. Normally, DL models benefit more from training with 3D electron density data in real space compared to relying on 1D powder XRD-based classification. This is because 3D electron density inherently provides more information about the crystal structure. The enhanced performance of PointNet and Sparse 3D CNN is understandable when considering that 1D powder XRD is derived from 3D electron density, and inevitably, some information is lost during the transformation from 3D to 1D data (a process of contraction gives rise to peak overlaps). A well-known challenge in generating 1D powder XRD data is the issue of peak overlap, which can complicate the data analysis process. This issue likely contributes to a reduction in FCN's accuracy.

The third issue to address is the relatively poor performance of coGN compared to both PointNet and Sparse 3D CNN, where it also lags far behind FCN. coGN, an advanced iteration of CGCNN[55], was designed to better represent crystal structures. While CGCNN accounts for the periodicity of the unit cell graph, emphasizing translational symmetry alone, coGN employs an asymmetric unit graph representation that considers all symmetries of a crystal structure. This approach placed coGN at the top rank in the property regression benchmark, as shown in Supplementary Fig. 1. Previous comparative analyses between CGCNN and FCN revealed that CGCNN excelled over FCN in property regression but fell short in symmetry classification[74]. This pattern persists in comparisons between coGN and 3D electron density-based DL models like PointNet and Sparse 3D CNN. Despite significant enhancements in coGN's ability to analyze symmetry, thereby improving its property regression capabilities, it still cannot outperform FCN, PointNet, and Sparse 3D CNN in symmetry classification. The reasons behind coGN's shortcomings in symmetry classification remain unclear, in spite of its substantial advancements in symmetry consideration.

The regression capabilities of coGN were impressive, largely due to the traditional descriptors used for node and edge features in graph-type networks. These descriptors include atomic number, atomic mass, atomic radius, electronegativity, ionization energy, oxidation state, bond length, and others. However, coGN was less effective in capturing symmetry. We believe that the success of graph-based DL models in property regression is primarily attributed to the incorporation of such human-selected features, specifically traditional descriptors that represent atomic and bonding characteristics for node and edge features. This is a key factor in property regression, rather than the graph representation of crystalline structure. Evidence supporting this view is that even a simple sparse composition vector-driven MLP demonstrated nearly equivalent performance to CGCNN in band gap regression[20]. While coGN's performance has significantly improved, it appears that much of its enhanced capability can be credited to the selection of traditional material descriptors by domain experts playing a major role in property regression, rather than the graph structure itself.

The fourth issue addresses the marginally inferior performance of PointNet in symmetry classification compared to Sparse 3D CNN, despite its significant advantage over FCN and coGN. PointNet typically processes point cloud data derived from original dense voxel-type data through random downsampling. These point clouds consist of 5000 to 30,000 points, the number of which varies based on the lattice size. The point represents 3D coordinates and their associated feature values, with electron density being the primary feature. The size of the point cloud, dictated by the number of points, might be insufficient for capturing detailed local electron density, potentially leading to suboptimal performance of PointNet. Evidence suggests that increasing the sample size could enhance PointNet's performance, as indicated by the plot in Supplementary Fig. 8, showing PointNet performance against point cloud size. However, pursuing larger point clouds may be impractical, considering that Sparse 3D CNN achieves superior results at the same point cloud size with lower computational demands. Therefore, opting for Sparse 3D CNN, instead of a larger-sample-based PointNet, appears to be a more efficient approach for achieving improved test accuracy in symmetry classification.

Supplementary Table 2 presents the property regression results for band gap (E_g) and formation energy (E_f) using only the 120k_MP dataset, as the 190k_ICSD dataset absences labels for E_g and E_f. In these regressions,

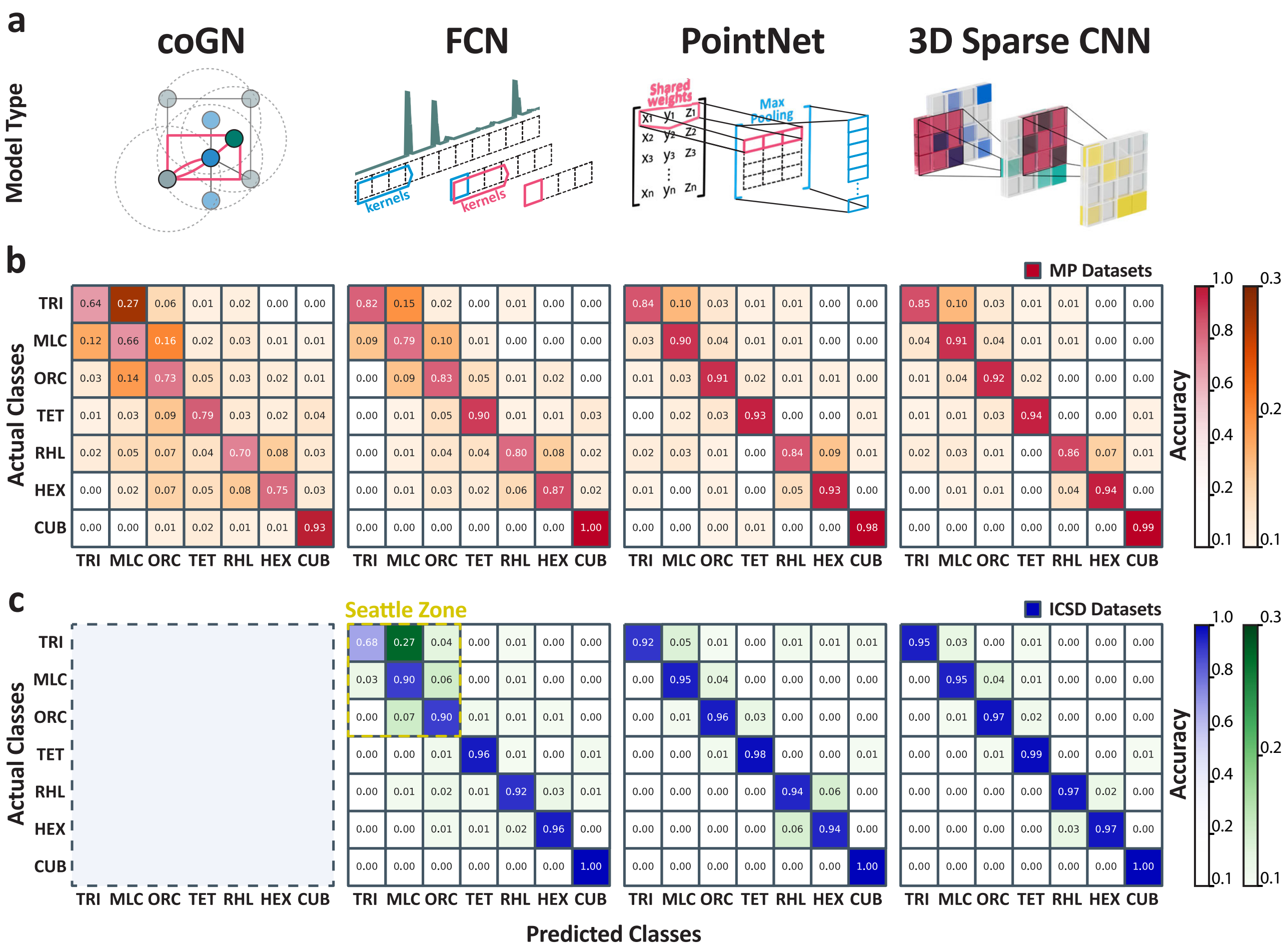


Fig. 4 | Classification accuracy of crystal systems by DL models represented through confusion matrices. a This section displays representations of four distinct DL models used in symmetry classification: coGN, FCN, PointNet, and 3D Sparse CNN. Each model is symbolized by a central concept illustration. **b** Confusion matrices showing the classification of 7 crystal systems using the 120k_MP dataset, corresponding to the DL models presented. **c** Confusion matrices for crystal system classification utilizing the 190k_ICSD dataset. The 'Seattle zone' is highlighted with a yellow square in the confusion matrix for the FCN model trained on the 190k_ICSD dataset. Confusion matrices for extinction group and space group classifications are included in Supplementary Fig. 9. A dashed outline indicates the absence of data for the coGN model with the 190k_ICSD dataset, due to an incompatibility issue between the model and the dataset.

coGN demonstrated superior performance, while Sparse 3D CNN exhibited a higher mean absolute error (MAE). Notably, if Sparse 3D CNN had used only electron density as a feature, the MAE for E_g regression was even higher. The MAE values for E_g and E_f regression were reduced considerably when using a four-feature sparse tensor, which included total electron density, positive spin density, negative spin density, and the difference between the two spin densities (magnetization density), as evidenced in Supplementary Table 3. The inclusion of these spin-related features significantly enhanced the regression results, underscoring the importance of spin considerations in relation to E_g and E_f. However, coGN outperformed Sparse 3D CNN despite the use of a four-feature sparse tensor. It is also noted that the MAE values for coGN-based regression in Supplementary Table 2 are higher (worse) than those in the benchmark[58], which can be attributed to differences between the 120k_MP dataset and the dataset used in the benchmark.

The 3D-electron-density-driven DL approach, despite its theoretical advantages, faces practical challenges. Obtaining 3D electron density data requires high-cost computations and is not directly measurable through experimental methods; it relies on crystal structure solution data, which includes atomic positions, occupancies, thermal vibration tensors, as well as exact symmetry (space group) and lattice parameters. Typically, the crystal structure solution for inorganic materials is derived from powder or single crystal XRDs. This practicality issue also applies to other 3D-data-driven approaches like those by Tsuruta et al.[66] and Ziletti et al.[40], which necessitate known crystal structure solutions to produce training data to be fed to their DL models. Unless the exact crystal structure solution is fully known, the input data for DL models cannot be obtained, as there is currently no experimental means to directly measure them.

In the current landscape, XRD data-driven DL approaches, such as FCN, hold more practical value in experimental materials science. FCN can predict the exact symmetry of an unknown material using its experimental XRD pattern, which is a task easily achievable in standard materials laboratories. Despite its slightly inferior performance compared to Sparse 3D CNN, FCN's role in materials analysis and discovery remains crucial. The excellent performance of Sparse 3D CNN serves as a potential upper limit for what FCN might achieve in the future. Nevertheless, there is optimism that direct measurement of 3D electron density in experimental settings will be feasible sometime in the future. This anticipated advancement would underscore the practical advantages of our 3D-electron-density-driven DL approach over other methods that rely solely on structure solution data.

Failure analysis for the symmetry classification

A failure analysis using confusion tables, as depicted in Fig. 4, was conducted. Focusing on the 190k_ICSD dataset (Fig. 4c), the confusion table (or confusion matrix) for FCN reveals significant insights into the challenges

DL models face in classifying low symmetry structures. Suzuki et al.[16] initially highlighted that DL model predictability deteriorates in low symmetry cases, a finding later corroborated by Lee et al.[20,21]. However, the underlying cause of this deterioration remains unclear. Figure 4c demonstrates that FCN struggles with accuracy and exhibits high mishit rates in the off-diagonal areas, particularly in the low symmetry region (upper-left zone of the confusion table) of the confusion table (Triclinic, Monoclinic, and Orthorhombic), a phenomenon termed the 'Seattle zone'[21]. Notably, both PointNet and Sparse 3D CNN do not exhibit this Seattle zone, as evidenced in Fig. 4c. The mishit rates for Sparse 3D CNN-driven symmetry classification are consistently low, ranging from 0 to 0.04, regardless of symmetry, in stark contrast to FCN's high off-diagonal mishit rates in the 'Seattle zone'.

Although the Seattle zone appears less pronounced in the 120k_MP dataset compared to the 190k_ICSD dataset, this observation is somewhat misleading. The overall accuracy for the 120k_MP dataset is significantly lower than that for the 190k_ICSD dataset, indicating that results from the MP dataset are equally subpar, irrespective of symmetry. Interestingly, coGN also exhibits the Seattle zone, albeit with an overall decrease in accuracy. The Seattle zone, or the observed accuracy deterioration in low symmetry regions, seems not to be a result of DL model limitations but rather an issue stemming from the data itself. Crystals can be represented as either nuclei arrangements or electron clouds in real space, but contemporary crystallographic analyses predominantly rely on diffractions (electron, neutron, or X-ray) projected onto 2D and 1D reciprocal spaces. Powder XRD, a 1D diffracted projection, inherently suffers from significant information loss, notably the peak-overlap complication. The original 3D electron density data in real space offers a more comprehensive representation of crystals and is thus ideal for DL approaches. The absence of the Seattle zone in Sparse 3D CNN highlights the robustness of 3D electron density data, confirming it as the superior choice for symmetry classification training data. The limitations of 1D powder XRD and graph-type data in accurately representing the structure of inorganic materials become evident as both FCN and coGN exhibit the Seattle zone. The Seattle zone issue, it should be emphasized, stems from data incompleteness, not from any inherent shortcomings in the DL models themselves.

Discussion

The use of 3D electron density data in real space proved to be more effective for DL-driven symmetry classification of inorganic materials compared to traditional crystallographic structural data, such as 1D XRD and graph-type data. However, traditional 3D CNNs struggled with training on immensely dense 3D charge density data due to computational limitations. To alleviate such computational burdens, we introduced sparsity to the 3D electron density data through downsampling, resulting in sparse formats like point clouds and sparse tensors. Consequently, we utilized PointNet and Sparse 3D CNN for symmetry classification, leveraging these sparse 3D electron density data representations.

The combination of the sparse 3D electron density dataset with the Sparse 3D CNN model led to the highest hold-out test accuracy for symmetry classification, achieving 97.28% for crystal system classification, 90.77% for extinction, and 90.10% for space group classification. These results set SOTA records and mark the first time accuracies have surpassed 90% for 230-space-group classification. The PointNet performed comparably to the Sparse 3D CNN. Following them, the FCN, trained on 1D XRD data, and the coGN, trained on graph-type data, demonstrated successive effectiveness.

A comparative analysis of the 190k_ICSD and 120k_MP datasets showed that DL models trained on the 190k_ICSD dataset consistently surpassed those trained on the 120k_MP dataset in performance. This difference could be attributed to the 120k_MP dataset's traits such as skewed data distribution favoring lower symmetry and a greater prevalence of unrealistic, fully-occupied entries.

In the confusion matrix for earlier 1D XRD- and graph-based DL models, a region known as the "Seattle zone," characterized by high misclassification rates, was particularly evident in the low symmetry area. However, this was not the case with the 3D electron density-based DL approach. This indicates that difficulties in classifying materials with low symmetry are more a result of data limitations than inherent flaws in the DL models. Traditional structural data, such as 1D powder XRD and graph-type representations, often struggle to fully capture the complex feature of crystal structures. This shortcoming is primarily due to the inherent loss of structural information, a consequence that arises inevitably when original 3D crystal data is condensed into these simpler formats. Nonetheless, graph-based DL models such as CGCNN and coGN are highly effective in property regression, likely benefiting from the inclusion of human-selected descriptors traditionally employed in materials science.

Methods

Electron density data preparation

We utilized databases from MP[71] and ICSD[72], preparing three different types of data, 1D powder XRD data, graph-type data, and 3D electron density data. The MP database comprises around 150,000 material entries, but charge density data are available for only 122,689 of these, which are the focus of this study. MP data includes structure solution along with bandgap energy (E_g) and formation energy (E_f), facilitating both classification and regression modeling. However, the ICSD database, devoid of material property data, was used solely for symmetry classification modeling.

The 3D electron density data from the MP database were derived from charge density data obtained via DFT calculations using VASP[82–84]. This data was transformed into voxel format using the materials' lattice parameters and Cartesian coordinates (x,y,z), along with the total charge density (spin up + spin down) data (feature), creating a point cloud in the dimension of $N \times (x, y, z, \text{features})$ required by the 3D deep learning models in this study. For property regression modeling with the 3D electron density data from MP database, a single-feature data using only total charge density proved ineffective. In contrast, a four-feature data incorporating total charge density, magnetization density (spin up - spin down), spin up, and spin down data was introduced. For the ICSD data, where 3D charge density is absent and only structural solution data is available, unlike in the MP, we used VESTA[86] software to generate electron density voxel data for each material. This data was then converted into 3D point cloud data, similar to the approach for MP data. Lattice parameters for each database were selected based on the primitive cell for MP data and conventional cell (Bravais lattice) parameters for ICSD data. Further details on the 3D electron density data preparation are available in Supplementary Note 2, 3, and 4.

For 1D XRD pattern data in the FCN approach, CIFs from each database were converted into XRD peak data using Fullprof Suite[87], following the methodology of Park et al.[13]. Graph-type data preparation was limited to fully occupied crystal structures, thus only MP data were used. For this, both the CGCNN[55], the inaugural graph-type DL approach for materials, and coGN[57], the latest crystal graph neural network technique, were referenced.

DL model training

For 1D powder XRD data, we employed the FCN architecture with 1D convolution (Supplementary Fig. 10), as implemented by Lee et al.[20,21], for classification and regression training across both MP and ICSD databases. For graph-type crystal data, the training for coGN and CGCNN models was limited to MP data for classification and regression purposes. The architectures of crystal graph-based DL models (CGCNN and coGN) were adapted from previous literature[55,57]. For 3D electron density data, we focused on two models capable of processing this data type: PointNet, which serves as a baseline for initial 3D point cloud modeling, and Sparse 3D CNN, designed to address the limitations of PointNet. We trained PointNet and Sparse 3D CNN models on MP data for both classification and regression, allowing for a performance comparison across all DL models, including CGCNN, coGN, FCN, PointNet, and Sparse 3D CNN. For ICSD data, which lacks material property data, training with PointNet and Sparse 3D CNN was exclusively focused on symmetry classification, allowing for a performance comparison with the FCN model trained on 1D XRD peak data. The detailed architectures and hyperparameters of the PointNet and Sparse 3D CNN models are presented in Supplementary Fig. 2. Mesh enumeration was used for hyperparameter optimization, with details available in Supplementary Table 4. Data were split into 80% for training, 10% for validation, and 10% for testing across all DL models. The evolution of training and validation performances for PointNet and Sparse 3D CNN, detailed as a function of epoch, are depicted in Supplementary Fig. 11.

Data availability

The datasets employed in this study are openly accessible at https://github.com/socoolblue/3D_Crystal_DL. The source data for all figures presented in this work are also available to readers for further analysis and verification.

Code availability

The source code associated with our research is available to the public and can be accessed at the following URL: https://github.com/socoolblue/3D_Crystal_DL.

Received: 7 February 2024; Accepted: 20 August 2024;
Published online: 09 September 2024

References

1. López, C. Artificial intelligence and advanced materials. *Adv. Mater.* **35**, 2208683 (2023).
2. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **8**, 59 (2022).
3. Szymanski, N. J. et al. Toward autonomous design and synthesis of novel inorganic materials. *Mater. Horiz.* **8**, 2169–2198 (2021).
4. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B.* **89**, 094104 (2014).
5. Ren, F. et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **4**, eaaq1566 (2018).
6. Ludwig, A. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Comput. Mater.* **5**, 70 (2019).
7. Velasco, L. Phase–Property Diagrams for Multicomponent Oxide Systems toward Materials Libraries. *Adv. Mater.* **33**, 2102301 (2021).
8. Stein, H. S. & Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **10**, 9640–9649 (2019).
9. Zhang, Z. et al. Finding the Next Superhard Material through Ensemble Learning. *Adv. Mater.* **33**, 2005112 (2021).
10. Ryan, K., Lengyel, J. & Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
11. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Matter.* **4**, 063801 (2020).
12. Butler, K. T. et al. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
13. Park, W. B. et al. Classification of crystal structure using a convolutional neural network. *IUCrJ* **4**, 486–494 (2017).
14. Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **5**, 60 (2019).
15. Vecsei, P. M., Choo, K., Chang, J. & Neupert, T. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Phys. Rev. B* **99**, 245120 (2019).
16. Suzuki, Y. et al. Symmetry prediction and knowledge discovery from X-ray diffraction patterns using an interpretable machine learning approach. *Sci. Rep.* **10**, 21790 (2020).
17. Suzuki, Y. et al. Machine Learning-based Crystal Structure Prediction for X-Ray Microdiffraction. *Microsc. Microanal* **24**, 142 (2018).
18. Schuetzke, J., Benedix, A., Mikut, R. & Reischl, M. Enhancing deep-learning training for phase identification in powder X-ray diffractograms. *IUCrJ* **8**, 408–420 (2021).
19. Schuetzke, J., Szymanski, N. J. & Reischl, M. Validating neural networks for spectroscopic classification on a universal synthetic dataset. *npj Comput. Mater.* **9**, 100 (2023).
20. Lee, B. D. et al. Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction. *Adv. Intell. Syst.* **4**, 2200042 (2022).
21. Lee, B. D. et al. A Deep Learning Approach to Powder X-Ray Diffraction Pattern Analysis: Addressing Generalizability and Perturbation Issues Simultaneously. *Adv. Intell. Syst.* **5**, 2300140 (2023).
22. Lee, J.-W. et al. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **11**, 86 (2020).
23. Lee, J.-W. et al. A data-driven XRD analysis protocol for phase identification and phase-fraction prediction of multiphase inorganic compounds. *Inorg. Chem. Front.* **8**, 2492–2504 (2021).
24. Massuyeau, F. et al. Perovskite or Not Perovskite? A Deep-Learning Approach to Automatically Identify New Hybrid Perovskites from X-ray Diffraction Patterns. *Adv. Mater.* **34**, 2203879 (2022).
25. Chen, D. et al. Automating Crystal-Structure Phase Mapping: Combining Deep Learning with Constraint Reasoning. *Nat. Mach. Intell.* **3**, 812–822 (2021).
26. Szymanski, N. J. et al. Probabilistic Deep Learning Approach to Automate the Interpretation of Multi-phase Diffraction Spectra. *Chem. Mater.* **33**, 4204–4215 (2021).
27. Maffettone, P. M. et al. Crystallography companion agent for high-throughput materials discovery. *Nat. Comput. Sci.* **1**, 290–297 (2021).
28. Wang, H. et al. Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *J. Chem. Inf. Model.* **60**, 2004–2011 (2020).
29. Dong, H. et al. A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *npj Comput. Mater.* **7**, 74 (2021).
30. Chitturi, S. R. et al. Automated prediction of lattice parameters from X-ray powder diffraction patterns. *J. Appl. Cryst.* **54**, 1799–1810 (2021).
31. Banko, L. et al. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. *npj Comput. Mater.* **7**, 104 (2021).
32. Stanev, V. et al. Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Comput. Mater.* **4**, 43 (2018).
33. Dong, R. et al. DeepXRD, a Deep Learning Model for Predicting XRD spectrum from Material Composition. *ACS Appl. Mater. Interfaces* **14**, 40102–40115 (2022).
34. Bunn, J. K. et al. Semi-supervised approach to phase identification from combinatorial sample diffraction patters. *JOM* **68**, 2116–2125 (2016).
35. Xiong, Z., He, Y., Hattrick-Simpers, J. R. & Hu, J. Automated Phase Segmentation for Large-Scale X-ray Diffraction Data Using a Graph-Based Phase Segmentation (GPhase) Algorithm. *ACS Comb. Sci.* **19**, 137–144 (2017).
36. Hattrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge. *APL Mater* **4**, 053211 (2016).

37. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
38. Iwasaki, Y., Kusne, A. G. & Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **3**, 4 (2017).
39. Kunka, C. et al. Decoding defect statistics from diffractograms via machine learning. *npj Comput. Mater.* **7**, 67 (2021).
40. Ziletti, A., Kumar, D., Scheffler, M. & Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **9**, 2775 (2018).
41. Starostin, V. et al. Tracking perovskite crystallization via deep learning-based feature detection on 2D X-ray scattering data. *npj Comput. Mater.* **8**, 101 (2022).
42. Tiong, L. C. O., Kim, J.-R., Han, S. S. & Kim, D.-H. Identification of crystal symmetry from noisy diffraction patterns by a shape analysis and deep learning. *npj Comput. Mater.* **6**, 196 (2020).
43. Aguiar, J. A. et al. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **5**, 10 (2019).
44. Salgado, J. E. et al. Automated classification of big X-ray diffraction data using deep learning models. *npj Comput. Mater.* **9**, 214 (2023).
45. Szymanski, N. J. et al. Integrated analysis of X-ray diffraction patterns and pair distribution functions for machine-learned phase identification. *npj Comput. Mater.* **10**, 45 (2024).
46. Szymanski, N. J. et al. Adaptively driven X-ray diffraction guided by machine learning for autonomous phase identification. *npj Comput. Mater.* **9**, 31 (2024).
47. Visser, J. W. A fully automatic program for finding the unit cell from powder data. *J. Appl. Cryst.* **2**, 89–95 (1969).
48. Werner, P.-E., Eriksson, L., Westdahl, M. & TREOR a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Cryst.* **18**, 367 (1985).
49. Boultif, A. & Louër, D. Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method. *J. Appl. Cryst.* **24**, 987–993 (1991).
50. Bail, A. L. Monte Carlo indexing with McMaille. *Powder Diffr* **19**, 249–254 (2004).
51. Altomare, A. et al. EXPO2009: structure solution by powder data in direct and reciprocal space. *J. Appl. Cryst.* **42**, 1197–1202 (2009).
52. Neumann, M. A. X-Cell: a novel indexing algorithm for routine tasks and difficult cases. *J. Appl. Cryst.* **36**, 356–365 (2003).
53. Rodriguez-Carvajal, J. Recent advances in magnetic structure determination by neutron powder diffraction. *Physica B* **192**, 55 (1993).
54. Battaglia, P. W. et al. Relational inductive biases, deep learning, and graph networks. Preprint at https://arxiv.org/abs/1806.01261 (2018).
55. Xie, T. & Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
56. Chen, C. et al. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
57. Ruff, R. et al. Connectivity Optimized Nested Graph Networks for Crystal Structures. *Digit. Dicov.* **3**, 594–601 (2023).
58. Dunn, A. et al. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).
59. Charles, R. Q. et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Preprint at https://arxiv.org/abs/1612.00593 (2017).
60. Graham, B. Spatially-sparse convolutional neural networks. Preprint at https://arxiv.org/abs/1409.6070 (2014).
61. Liu, B. et al. Sparse Convolutional Neural Networks. In *IEEE Xplore*, 806–814 (IEEE, 2015).
62. Wang, P.-S. et al. O-CNN: Octree-based Convolutional Neural Networks for 3D Shape Analysis. Preprint at https://arxiv.org/abs/1712.01537v1 (2017).
63. Zhao, Y. et al. Predicting elastic properties of materials from electronic charge density using 3D deep convolutional neural networks. *J. Phys. Chem. C* **124**, 17262–17273 (2020).
64. DeFever, R. S. et al. A generalized deep learning approach for local structure identification in molecular simulations. *Chem. Sci.* **10**, 7503–7515 (2019).
65. Zhengyuan, S., Sun, Y., Lodge, T. P. & Siepmann, J. I. Development of a PointNet for Detecting Morphologies of Self-Assembled Block Oligomers in Atomistic Simulations. *J. Phys. Chem. B* **125**, 5275–5284 (2021).
66. Tsuruta, H., Katsura, Y. & Kumagai, M. DeepCrysTet: A Deep Learning Approach Using Tetrahedral Mesh for Predicting Properties of Crystalline Materials. Preprint at https://arxiv.org/abs/2310.06852v1 (2023).
67. Chiba, N. et al. Neural structure fields with application to crystal structure autoencoders. *Commun. Mater.* **4**, 106 (2023).
68. Midgley, P. A. & Weyland, M. 3D electron microscopy in the physical sciences: the development of Z-contrast and EFTEM tomography. *Ultramicroscopy* **96**, 413–431 (2003).
69. Xin, H. L. & Muller, D. A. Aberration-corrected ADF-STEM depth sectioning and prospects for reliable 3D imaging in S/TEM. *Microscopy* **58**, 157–165 (2009).
70. Scott, M. C. et al. Electron tomography at 2.4-ångström resolution. *Nature* **483**, 444–447 (2012).
71. Jain, A. et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater* **1**, 011002 (2013).
72. Inorganic Crystal Structure Database (ICSD release 2023.1) https://icsd.products.fiz-karlsruhe.de/ (accessed: September 2023).
73. Long, J., Shelhamer, E. & Darrell, T. Fully Convolutional Networks for Semantic Segmentation. Preprint at https://arxiv.org/abs/1411.4038v2 (2015).
74. Allmann, R. & Hinek, R. The introduction of structure types into the Inorganic Crystal Structure Database ICSD. *Acta Cryst. A* **63**, 412–417 (2007).
75. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. Preprint at https://arxiv.org/abs/2106.08254 (2021).
76. Dosovitskiy, A. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at https://arxiv.org/abs/2010.11929 (2021).
77. Szegedy, C. et al. Going Deeper with Convolutions. Preprint at https://arxiv.org/abs/1409.4842 (2014).
78. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. Preprint at https://arxiv.org/abs/1512.03385 (2015).
79. Wong, T.-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit* **48**, 2839–2846 (2015).
80. Miao, J. et al. High resolution 3D x-ray diffraction microscopy. *Phys. Rev. Lett.* **89**, 088303 (2002).
81. Saha, P. & Nguyen, M. T. Electron density mapping of boron clusters via convolutional neural networks to augment structure prediction algorithms. *RSC Adv.* **13**, 30743–30752 (2023).
82. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558 (1993).
83. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B Condens. Matter Mater. Phys.* **49**, 14251–14269 (1994).
84. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).

85. Prince, E. *International Tables for Crystallography*, vol. C (Wiley, 2004). http://lampx.tugraz.at/~hadley/ss1/crystaldiffraction/atomicformfactors/formfactors.php
86. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *J. Appl. Crystallogr.* **44**, 1272–1276 (2011).
87. Rodriguez-Carvajal, J. Recent developments of the program FULLPROF, commission on powder diffraction. *IUCr Newsl*. **26**, 12–19 (2001).
88. Tran, D. et al. Learning Spatiotemporal Features with 3D Convolutional Networks. Preprint at https://arxiv.org/abs/1412.0767 (2014).

Acknowledgements

This research was supported by the Alchemist Project (20012196) funded by MOTIE, Korea and partly by the National Research Foundation of Korea (NRF) (2021R1A2C1009144 and RS-2024-00446825).

Author contributions

K.-S.S. conceived the concept for the entire process and directed the computational process. S.K. prepared the simulated 3D electron density data. S.K. and B.D.L. took part in coding, trained and tested all the DL models. W.-B. P. and M.P. carried out the powder XRD-driven DL approach and the crystallographic analyses. B.D.L., Y.-K.L. and M.C. prepared the simulated 1D powder XRD data. B.D.L. prepared figures and tables. K.-S.S. and S.K. wrote the paper. All authors discussed the results and commented on the manuscript. S.K. and B.D.L. contributed equally to this work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01402-7.

Correspondence and requests for materials should be addressed to Woon Bae Park or Kee-Sun Sohn.