Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 694

A Cartesian encoding graph neural network for crystal structure property prediction: application to thermal ellipsoid estimation†

Àlex Solé, [ab] Albert Mosella-Montoro, [a] Joan Cardona, [b] Silvia Gómez-Coca, [*b] Daniel Aravena, [*c] Eliseo Ruiz [*b] and Javier Ruiz-Hidalgo [*a]

In the diffraction resolution of crystal structures, thermal ellipsoids are a critical parameter that is usually more difficult to determine than atomic positions. These ellipsoids are quantified through Anisotropic Displacement Parameters (ADPs), which provide critical insights into atomic vibrations within crystalline structures. ADPs reflect the thermal behaviour and structural properties of crystal structures. However, traditional methods to compute ADPs are computationally intensive. This paper presents CartNet, a novel graph neural network (GNN) architecture designed to predict properties of crystal structures efficiently by encoding the atomic structural geometry to the Cartesian axes and the temperature of the crystal structure. Additionally, CartNet employs a neighbour equalization technique for message passing to help emphasise the covalent and contact interactions and a novel Cholesky-based head to ensure valid ADP predictions. Furthermore, a rotational SO(3) data augmentation technique has been proposed during the training phase to generalize unseen rotations. To corroborate this procedure, an ADP dataset with over 200 000 experimental crystal structures from the Cambridge Structural Database (CSD) has been curated. The model significantly reduces computational costs and outperforms existing previously reported methods for ADP prediction by 10.87%, while demonstrating a 34.77% improvement over the tested theoretical computation methods. Moreover, we have employed CartNet for other already known datasets that included different material properties, such as formation energy, band gap, total energy, energy above the convex hull, bulk moduli, and shear moduli. The proposed architecture outperformed previously reported methods by 7.71% in the JARVIS dataset and 13.16% in the Materials Project dataset, proving CarNet's capability to achieve state-of-the-art results in several tasks. The project website with online demo available at: https://www.ee.ub.edu/cartnet.

Received 31st October 2024
Accepted 22nd January 2025

DOI: 10.1039/d4dd00352g

rsc.li/digitaldiscovery

1 Introduction

Anisotropic Displacement Parameters (ADPs)[1] represent a three-dimensional ellipsoid of atomic displacements within a crystal lattice due to thermal vibrations. These ellipsoids are fundamental for understanding the anisotropic nature of atomic movements and the dynamic behaviour of materials at the atomic scale. They provide critical insights into the structural and thermal properties of materials, influencing the interpretation of experimental data from techniques such as single crystal X-ray diffraction and neutron scattering. Accurate representation of ADPs is essential for constructing precise structural models and understanding the physical behaviour of materials under varying thermal conditions.

ADPs are particularly valuable in crystallography, as they assist in identifying determination issues such as disorder or twinning. Several studies have shown that ADPs can also be utilized to predict thermal motion and translational and vibrational frequencies.[1,2] Furthermore, as demonstrated in previous research, thermal properties such as heat capacity (C_v)[3,4] and vibrational entropy[5,6] are directly linked to ADPs. Additionally, ADPs have been related to the thermal expansion of crystal structures,[7,8] making them especially useful for identifying materials with negative thermal expansion. From an experimental point of view, inconsistencies in atomic positions are often easily spotted using chemical intuition when solving a crystal structure, allowing for straightforward visual identification of errors. However, such intuitive assessments are not as straightforward when it comes to thermal ellipsoids, making visual detection of discrepancies more challenging. As a result, some database structures exhibit ellipsoids with seemingly

[a]*Image Processing Group – Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona, Spain. E-mail: j.ruiz@upc.edu*

[b]*Inorganic and Organic Chemistry Department, Institute of Theoretical and Computational Chemistry, Universitat de Barcelona, Barcelona, Spain. E-mail: silvia.gomez@qi.ub.es; eliseo.ruiz@qi.ub.edu*

[c]*Materials Chemistry Department, Faculty of Chemistry and Biology, Universidad de Santiago de Chile, Santiago, Chile. E-mail: daniel.aravena.p@usach.cl*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00352g

anomalous sizes and orientations. Developing theoretical methods that enable quick evaluation of these ellipsoids can thus serve as valuable tools for facilitating accurate structural determination from diffraction data.

From a theoretical perspective, ADPs can be calculated using periodic electronic structure calculations to obtain vibrational frequencies based on the harmonic approximation.[9,10] This method requires the numerical calculation of forces at displaced geometries for all atomic positions in solid-state systems. Thus, these calculations are computationally expensive and time-consuming, often creating a bottleneck in the process. Our contribution is particularly significant in this context, as deep learning methods based on graph neural networks can dramatically reduce the computation time required, providing a more efficient alternative to traditional approaches.

Fig. 1 shows the full pipeline proposed for this work, named CartNet. Our presented network uses a graph representation of the crystal structure and, through a set of learnable encodings and geometrical operations, predicts atomic or material properties.

Moreover, predicting ADPs presents a unique challenge and opportunity in the field of machine learning for crystal structure property prediction. Unlike most tasks typically explored in the literature, such as formation energy, band gap, or total energy, which are unaffected by rotation, predicting ADPs requires models to be sensitive to rotational orientation.

In this paper, we make the following contributions:

(1) ADP dataset: we present a meticulously curated dataset of ADPs with over 200 000 experimental crystal structures from the Cambridge Structural Database (CSD).[11]

The ADPs are both temperature-dependent and rotation-equivariant, offering a robust foundation for exploring the necessity and impact of rotation-equivariant architectures in crystal modelling.

This necessitates the use of architectures that can handle properties dependent on direction, opening the door to designing networks capable of processing rotationally dependent features.

See Section 3 for a detailed description of the dataset.

(2) CartNet architecture: we introduce an architecture, CartNet, capable of accurately predicting material properties based on the geometry of a crystal structure.

The key contributions of this model are as follows:

(a) Geometry and temperature encoding: we propose a feature descriptor that efficiently encodes the complete 3D geometry referenced to the Cartesian axis and adaptively fuses other input attributes such as temperature.

Since the geometry is anchored to the Cartesian reference axes, there is no need to encode the unit cell.

Cell-less encoding enables the accurate prediction of the crystal's ADP orientation regardless of cell size.

(b) Neighbour equalization: we present a neighbour equalization technique designed to address the exponential increase in neighbours over distance, thereby enhancing the model's ability to detect various types of bonds and force interactions between atoms.

This technique improves the model's sensitivity to different interaction ranges, ensuring a more precise representation of atomic environments.

(c) Cholesky head: we introduce an output layer based on Cholesky decomposition, which guarantees that the model produces positive definite matrices, an essential mathematical property requirement for valid ADP predictions.

(3) Rotation SO(3) augmentation: we propose a data augmentation technique for both input features and output ADPs to facilitate the creation of SO(3) rotation-equivariant representations.

This technique enables the model to learn rotational equivariance without requiring specific layers to enforce it, thereby simplifying the overall architecture and reducing the number of trainable parameters.

These contributions allowed our model to outperform previously reported methods by 8.85% in the JARVIS dataset[12] and 15.5% in the Materials Project dataset,[13] *vide infra*.

Furthermore, in the ADP dataset examined, CartNet demonstrated a 10.87% improvement over other previously reported methods and a 34.77% improvement over the tested low-level Generalized Gradient Approximation (GGA) Density Functional Theory (DFT) calculations with simple dispersion corrections.

2 Background and related work

2.1 Thermal ellipsoids

Anisotropic Displacement Parameters (ADPs) represent the magnitudes and directions of atomic thermal vibrations within

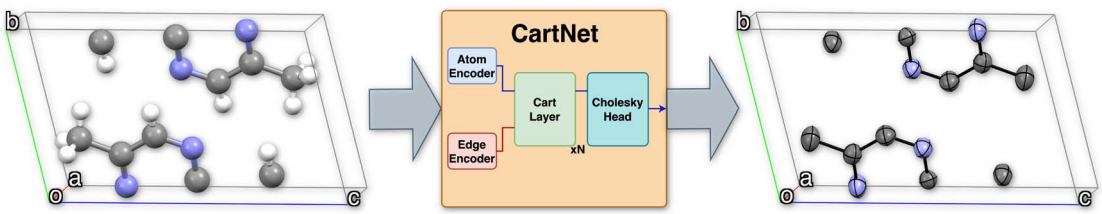

Fig. 1 Schematic of the CartNet graph neural network architecture for a 5,5′-dimethyl-2,2′-bipyrazine crystal structure (CSD refcode: ETIDEQ). The model predicts ADPs for all non-hydrogen atoms based on the positions of atoms within the unit cell. The architecture separately encodes atomic and edge information using dedicated encoders. This information is then aggregated through N iterations of message-passing *via* the CartLayer. Finally, the Cholesky head ensures that the output matrix is symmetric and positive-definite, generating a valid ADP matrix. White, light purple, and grey colours represent hydrogen, nitrogen, and carbon atoms, respectively. The parallelepiped represents the unit cell, and the red, green, and blue lines correspond to the *a*, *b*, and *c* unit cell axes.

crystal structures. ADPs encapsulate the statistical probability distribution of the position of the atoms resulting from thermal vibrations. Typically, ADPs are graphically represented using ellipsoids within the ORTEP visualization[14] that depicts a 50% probability contour indicating the likelihood of finding the atom within the ellipsoid's bounds.

Mathematically, ADPs are represented as a three-dimensional tensor (3×3 matrix), which functions as a covariance matrix for a three-dimensional Gaussian distribution. The covariance matrix comprises variance elements along its diagonal for each X, Y and Z axis and covariance elements off the diagonal, illustrating the relationships between the axes. The covariance matrix $\mathbf{U}$ for a three-dimensional Gaussian distribution can be expressed using eqn (1).

$$\mathbf{U} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) & \mathrm{Cov}(X,Z) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) & \mathrm{Cov}(Y,Z) \\ \mathrm{Cov}(Z,X) & \mathrm{Cov}(Z,Y) & \mathrm{Var}(Z) \end{bmatrix} \quad (1)$$

Covariance matrices exhibit distinctive mathematical properties: they are symmetric, $\mathbf{U} = \mathbf{U}^T$, and positive semidefinite, which entails that all matrix eigenvalues are non-negative.

Fig. 2 shows an example with the graphical representation of the ADPs of a 5,5′-dimethyl-2,2′-bipyrazine unit cell. In the figure, each atom is depicted using the thermal ellipsoid representations of the ADPs obtained experimentally by X-ray diffraction.

ADPs can be calculated theoretically by computing the dynamical matrix of the crystal structure.[15,16] The dynamical matrix is a fundamental concept in solid-state physics used to describe the vibrations of atoms in a crystal lattice, known as phonons. It is a mathematical construct that captures how atoms interact with each other when they are slightly displaced from their equilibrium positions. The dynamical matrix $\mathbf{D}(\mathbf{q})$ is computed at each point $\mathbf{q}$ in the Brillouin zone. The Brillouin zone represents the fundamental region in reciprocal space that contains all the unique wave vectors necessary to describe a crystal structure's physical properties. The eigenvalues of the dynamical matrix represent the phonon's frequencies $\omega_\nu(\mathbf{q})$, also known as phonon modes.

ADPs can be calculated by integrating all the phonon modes for all the $\mathbf{q}$ points using eqn (2).

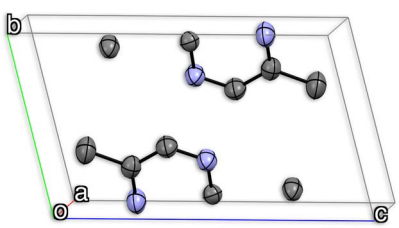

Fig. 2 Thermal ellipsoid ORTEP representations from experimental ADPs of a 5,5′-dimethyl-2,2′-bipyrazine crystal structure (CSD refcode: ETIDEQ). Light purple and grey colours represent nitrogen and carbon atoms, respectively. Hydrogen atoms have been omitted. The parallelepiped represents the unit cell, and the red, green, and blue lines correspond to the *a*, *b*, and *c* unit cell axis.

$$\mathbf{U}(j,T) = \frac{\hbar}{2Nm_j}\sum_{\mathbf{q},\nu}\frac{(1+2n_\nu(\mathbf{q},T))}{\omega_\nu(\mathbf{q})}\mathbf{e}_\nu(j,\mathbf{q})\otimes\mathbf{e}_\nu^*(j,\mathbf{q}) \quad (2)$$

where j is the atom, T is the temperature, $\hbar$ is the reduced Planck constant, N is the number of unit cells, m_j is the atomic mass, $\mathbf{e}_\nu$ are the eigenvectors of the dynamical matrix, $\otimes$ is the outer product, and n_ν is the phonon population. The phonon population can be described by eqn (3).

$$n_\nu(\mathbf{q},T) = \frac{1}{\exp(\hbar\omega_\nu(\mathbf{q})/k_\mathrm{B}T)-1} \quad (3)$$

where k_B is the Boltzmann constant.

This method has the limitation that it is extremely time-consuming since it needs a different Density Functional Theory (DFT) calculation for each of the $\mathbf{q}$ points. High-performance computing (HPC) clusters are often needed since DFT calculations can be computationally demanding, even for a single crystal structure.

2.2 Crystal structure property prediction with graph neural networks

GNNs[17] have been the most widely adopted neural network architecture for modelling data with complex relational structures such as molecules[18–20] or crystal structures.[21–23] Unlike traditional neural networks, which typically operate on fixed-size grid-like structures, GNNs are designed to work on graph-structured data, where entities are represented as nodes and relationships between them are represented as edges.

In the case of molecules, this ability to use graphs to model the system is especially useful, since a molecule can be naturally modelled as a graph of atoms. However, how atoms are interconnected by edges in the graph, denoted as neighbourhood, is not trivial and becomes a fundamental step to achieve good representations using GNNs.

While GNNs and message-passing mechanisms had been introduced previously, Gilmer *et al.*[18] popularized a unified message-passing framework specifically for quantum chemistry predictions, leveraging iterative information exchange between nodes (atoms) and edges (bonds) to capture local molecular interactions. This approach laid the groundwork for subsequent models that further enhance predictive accuracy for molecular properties. DimeNet[24] and its successor, DimeNet++,[25] build a GNN upon this concept by incorporating directional message passing between pairs of interactions, allowing these models to capture angular dependencies and long-range interactions within molecules more effectively than traditional message-passing methods using chemical bonds. This advancement significantly improves the accuracy of predicting molecular properties, such as quantum mechanical characteristics. Notably, both DimeNet and DimeNet++ are rotation-invariant, ensuring that their predictions remain consistent regardless of the molecule's orientation, which is crucial for accurate modelling in diverse molecular environments. MACE[26] proposed an equivariant message-passing approach that incorporates advanced three-dimensional geometric information, ensuring rotational and translational

symmetries for robust generalization across diverse molecular configurations. By enforcing rotational and translational symmetries, MACE provides robust generalization across diverse molecular configurations. Similarly, TensorNet[27] utilizes tensor-factorization techniques to incorporate higher-order interactions efficiently, capturing subtle quantum effects and long-range correlations. These approaches represent a significant evolution in molecular modeling, surpassing earlier methods in both accuracy and computational efficiency.

On the other hand, crystal structures consist of an assembly of atoms, ions or molecules that are ordered in a symmetric way. The formed symmetric pattern is repeated in the three spatial dimensions. We represent these structures using their smallest repeating unit, the unit cell, defined by a 3×3 lattice matrix encoding the three vectors that describe its geometry. Similar GNN models have been extensively applied to crystal property prediction, showcasing their adaptability to the unique challenges of materials science. One critical adaptation in this domain involves modifying the neighbourhood definition to account for periodic boundary conditions (PBCs). PBCs are essential for modelling the infinite nature of crystalline materials. They treat the border of the simulation cell as if they were connected to the opposite border, thus creating a continuous, repeating structure. This modification ensures that the model accurately captures interactions across the boundaries of the material.

Matformer[21] introduces a transformer-like GNN architecture designed explicitly for material property prediction. It proposes adding self-loops for each atom to encode the cell dimensions and the PBC radius neighbourhood, thereby enhancing the model's ability to represent the material's structure accurately. PotNet[22] advances this approach by presenting a GNN with a dual-neighbourhood strategy. In addition to the PBC radius neighbourhood, PotNet approximates the infinite summation of interactions for every pair of atoms in the crystal, providing a more comprehensive representation of the material's properties. Yan *et al.*[23] presented two GNN architectures tailored for this task, the iConformer and eConformer. The iConformer refines the cell to create a unique representation and uses the angle between the cell's axis and the Cartesian vector between atom pairs in the PBC radius neighbourhood to produce an invariant representation. In contrast, the eConformer introduces a rotation-equivariant model using tensor products,[28] relying solely on Cartesian information to maintain consistency across different orientations.

Regarding the specific problem of ADPs, current state-of-the-art methods exhibit several limitations. One of them is that most approaches rely solely on the distance between atoms to create a rotationally invariant representation. Since ADPs are expressed relative to the system's Cartesian axes (XYZ), using only distance does not provide the necessary references. Models that depend exclusively on distance are unable to differentiate between the variances along the axes (Var(X), Var(Y), or Var(Z)), often resulting in spherical ellipsoids that fail to capture the true anisotropy of the ADPs.

Another limitation of current state-of-the-art methods arises from the requirement for the lattice matrix. This is particularly problematic because multiple lattice matrices can represent the same crystal structure, as discussed in iComformer.[23] Although iComformer proposes a solution by creating a unique representation for each cell, several challenges remain when using this approach. The unique representation proposed by iComformer faces a border case when all three axes of the lattice matrix have the same length ($a = b = c$). Since iComformer's approach is based on the assumption that $a < b < c$, when the lattice matrix is in this degenerate case, the model is unable to differentiate between Var(X), Var(Y), or Var(Z), once again resulting in spherical ellipsoids.

Our proposed solution is a cell-less architecture that directly encodes the complete 3D geometry referenced to the Cartesian axes instead of only Euclidean distance-based approaches. This approach avoids the issues of border case scenarios and cell-based overfitting, as it relies entirely on the Cartesian space. Furthermore, it allows the inclusion of PBCs, providing a more robust and accurate representation of anisotropy in ADPs without the need for a lattice matrix.

3 Dataset

The ADP dataset of crystal structures has been created to facilitate the study of atomic thermal vibrations. This dataset is derived from the Cambridge Structural Database (CSD).[11]

The ADP dataset was meticulously curated from the CSD's built-in ADP subset, applying several filtering criteria to ensure its quality and reliability. First, only structures possessing 3D coordinates for all atoms and anisotropic thermal displacements for all non-hydrogen atoms were selected. Only non-polymeric crystal structures with only one type of molecule in the unit cell were considered, to avoid dispersion and errors due to solvent molecules and counterions. Structures with an R-factor less than 5% ($R < 5\%$), free from errors and disorders, and site occupancy of 1 for all atoms were included to maintain the dataset's integrity. Structures reported at non-standard pressures were also discarded.

Additionally, temperature/pressure data in the CSD can sometimes be incomplete or missing. For pressure, it was observed that in many cases the pressure was recorded only in the remarks field rather than in the dedicated pressure field, so structures containing any kind of remarks were discarded to avoid extensive text parsing. For temperature, if the CSD Python API did not provide a value, the corresponding CIF files and experimental notes were cross-checked to verify consistency. Structures lacking reliable temperature information were excluded from the dataset.

Several additional filtering criteria were applied to ensure the quality and reliability of the ADPs. While the 3D positions of the hydrogen are saved, their thermal ellipsoids are not included, as these are often isotropic or undefined in the CSD. Structures containing any non-hydrogen atom with negative or zero eigenvalues were excluded. Structures with any ellipsoids with eigenvalue (λ) ratios $\lambda_{max}/\lambda_{min} < 8$ were excluded to avoid poorly defined or flat ellipsoids. The ratio between the ellipsoid volume and the volume of a sphere with the covalent radius,[29] Vol/Vol$_{cov}$, was also used as a filtering criterion.

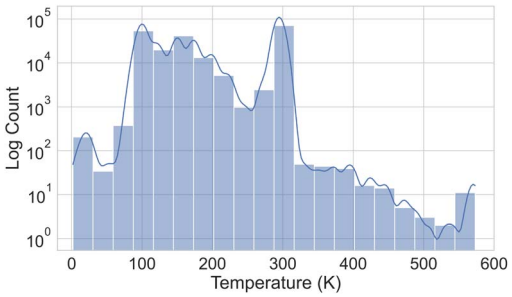

Fig. 3 Histogram showing the number of crystal structures within each temperature range in the ADP dataset, displayed on a logarithmic scale on the *y* axis.

Structures were discarded if any ellipsoid had a volume greater than 1.25 Å^3 or if Vol/Vol$_{cov}$ > 0.35. Furthermore, structures with any ellipsoid exhibiting a volume ratio Vol/Vol$_{cov}$ < 10^{-4} at temperatures above 150 K were also excluded to remove ellipsoids that were either too small or insufficiently defined.

At the end, 208, 042 crystal structures from the CSD met the criteria outlined, resulting in an average number of 194.2 atoms and 105.95 ADPs per crystal structure. Since the ADPs of hydrogen atoms are not considered, the dataset has more atoms than ADPs per crystal.

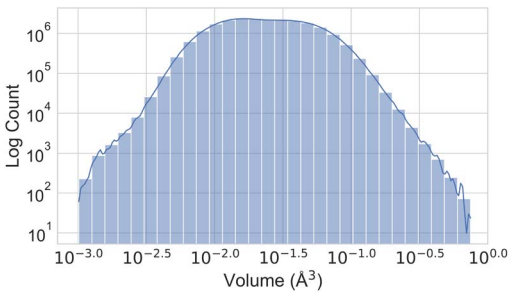

Fig. 4 Histogram illustrating the number of atoms within each ADP volume range in the ADP dataset, presented on a logarithmic scale on both axes.

Fig. 3 illustrates the temperature distribution across the curated dataset. Even though the dataset spans a wide range of temperatures, from 2 K to 573 K, most structures rely on the range from 100 K to 300 K since this range is the most commonly studied by diffractometry. Fig. 4 displays the distribution of the ADP's volumes used for this dataset. The same behaviour applies to the volumes, since the charts' extremes are under-represented in our data. Section 5.2.3 discusses the impact of this imbalance of the data on the prediction performance of the proposed model. Fig. 5 presents a heatmap of the atomic numbers included in this dataset. The dataset encompasses a wide range of atomic numbers, reflecting a diverse set of elements. It excludes some noble gases and most of the radioactive elements, except for some radioactive actinoids. This diversity is crucial for ensuring that the dataset can support the development of generalised models capable of predicting ADPs across a broad spectrum of chemical compositions.

In this study, the dataset was randomly split into 162, 270 training, 22, 219 validation, and 23, 553 test crystal structures. To ensure the integrity of these splits, we verified that all atom types, temperature ranges, and volume ranges present in the validation and test sets are also represented in the training set. This precaution was taken to prevent the model from encountering unseen data during validation and testing, which could otherwise lead to an inaccurate assessment of its performance.

Additionally, we ensured that repeated crystal structures with different temperatures or distinct CSD entries were kept together within the same split. This restriction was made to avoid any situation in which the model might be exposed to test or validation samples during the training phase, which could compromise the evaluation by introducing data leakage. By maintaining these strict controls on the dataset splits, we ensured that the training, validation, and test sets were independent. The splits used for this work are also publicly available to facilitate reproducibility.

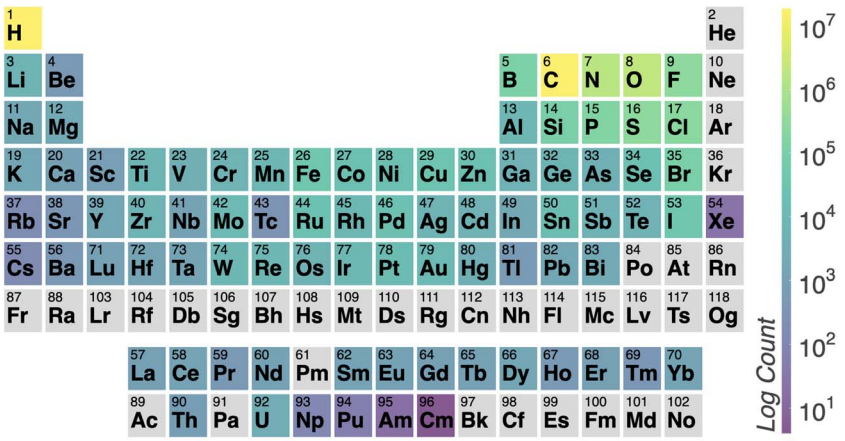

Fig. 5 Heatmap illustrating the number of atoms per element in the ADP dataset. Lighter colours represent higher counts, while darker colours indicate lower counts. The colour scale is logarithmic.

4 Methodology

4.1 Model architecture

Our proposed architecture, CartNet, efficiently encodes the geometry and any other relevant information of the crystal structures. The geometrical structure is encoded in the edges *via* Cartesian unit directions, and the input information is encoded along the nodes. In the case of ADPs, the input information is the temperature and the atomic number of each atom in the structure. The atom and edge information are then iteratively aggregated through the CartLayers, a process that is repeated four times to create a final vector representation for each atom. This final output is then processed through a specific head to predict the final properties. In the case of ADPs, the Cholesky head is used to produce mathematically valid ADPs. The complete architecture of CartNet is depicted in Fig. 1.

To construct the graph representation, we employ a radius-based neighbourhood approach using PBCs for all the atoms in the unit cell. Fig. 6 shows an example of how the graph is created for a single atom. Specifically, a cutoff radius ($r_c = 5$ Å) is defined around each central atom, and all atoms within this radius are considered part of the local neighbourhood. The choice of 5 Å is based on the analysis of the intermolecular interaction distances described in previous studies.[30] The graph obtained by this neighbourhood captures both short-range, usually covalent bonds, and relevant weaker intermolecular interactions, such as hydrogen bonds, π–π stacking, halogen bonds, cation–π and anion–π, or van der Waals interactions. By including all neighbour atoms within this defined radius, the graph effectively represents each atom's local environment, which is crucial for accurately modelling the system's behaviour.

4.1.1 Atom encoder. The atom encoder is responsible for encoding the input information of each atom in the graph. In the case of ADPs, this corresponds, for each atom, to its atomic number and the global temperature of the crystal structure. Fig. 7 shows a schematic of the atom encoder used in CartNet.

The atom type is encoded using an embedding[31] layer, which generates a feature vector corresponding to each distinct atom type. The temperature is standardised using the training temperature statistics to achieve zero mean and unitary standard deviation. The standardised temperature is passed through a linear layer to ensure dimensional compatibility with the atom-type feature vector. The resulting temperature and atom-type feature vectors are combined, passed through a SiLU[32] activation function, followed by another linear layer and another SiLU. Eqn (4) describes the encoding of the atom.

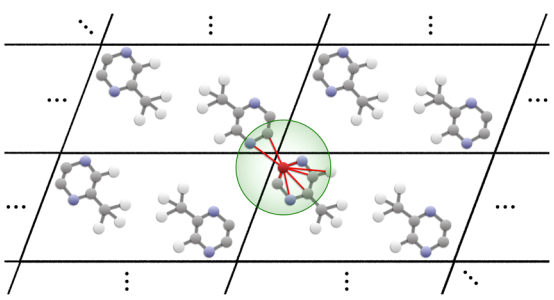

Fig. 6 Representation of the graph construction process for the atom highlighted in red colour. Covalent bonds are ignored, and the radius around the atom (depicted in green) is defined. Any atom within this radius is considered a neighbour of the red atom and is connected in the graph (depicted using red lines). Periodic boundary conditions are employed to replicate the infinite nature of the crystal.

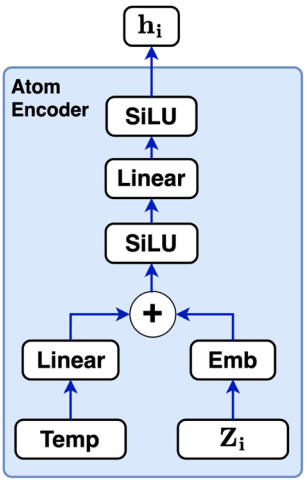

Fig. 7 Schematic of the atom encoder used in CartNet. The encoder processes each atomic number (Z_i) using an embedding layer and its temperature using a linear layer. The resulting encoded features are summed and passed through a SiLU activation function, followed by an additional linear layer and another SiLU. The final vector h_i serves as the initial node features for the CartLayer.

$$\mathbf{h_i} = \mathrm{SiLU}(\mathbf{W}_2(\mathrm{SiLU}(\mathrm{Emb}(Z_i) + \mathbf{W_1}(T) + \mathbf{b_1})) + \mathbf{b_2}) \in \mathbb{R}^{\mathrm{dim}} \quad (4)$$

where Z_i represents the atomic number, Emb is an embedding layer $\in \mathbb{R}^{2\mathrm{din}}$, T is the standardized temperature in Kelvin, $\mathbf{W_1} \in \mathbb{R}^{1\times 2\mathrm{dim}}$, $\mathbf{W_2} \in \mathbb{R}^{2\mathrm{dim}\times \mathrm{dim}}$, $\mathbf{b_1} \in \mathbb{R}^{2\mathrm{dim}}$, $\mathbf{b_2} \in \mathbb{R}^{\mathrm{dim}}$, and dim is the number of dimensions of the latent vector.

4.1.2 Edge encoder. The edge encoder is responsible for encoding the geometric relationships between atoms in the system. Fig. 8 shows a schematic of the edge encoder used in CartNet.

The edge is defined as the connection between the receiving atom i and the sender atom j. From each edge, the Euclidean distance (d_{ij}) and the direction vector (v_{ij}) are calculated based on their positions $\mathbf{p}$. Eqn (5) and (6) define the distance and direction vector, respectively.

$$d_{ij} = \|\mathbf{p_j} - \mathbf{p_i}\| \quad (5)$$

$$\hat{v}_{ij} = \frac{\mathbf{p_j} - \mathbf{p_i}}{d_{ij}} \quad (6)$$

where $\mathbf{p_i}$ and $\mathbf{p_j}$ are the receiver and sender atom positions. The distance is encoded through a Radial Basis Function (RBF) of K elements, transforming the scalar distances into a higher-dimensional space, allowing for a more nuanced representation of geometric relationships, as proposed by previous studies in molecular systems.[27] Eqn (7) defines the RBF at each k element.

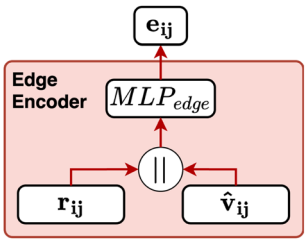

Fig. 8 Schematic of the edge encoder used in CartNet. The $\mathbf{e_{ij}}$ and the $\hat{\mathbf{v}}_{\mathbf{ij}}$ are concatenated and then processed by a $\mathrm{MLP_{edge}}$ to generate the initial edge features utilized by the CartLayer.

$$r_k(d_{ij}) = \exp(-\beta(\exp(-d_{ij}) - \mu_k)^2) \tag{7}$$

where β and μ_k are fixed values that determine the centre and width of the k-th radial basis function and K is the number of RBFs. The μ_k values are equally spaced between $\exp(-r_c)$ and 1, while the β value is equal to $[2K^{-1}(1 - \exp(-r_c))]^{-2}$ for all k. Here, r_c represents the cutoff radius distance used to define the neighbourhood.

Eqn (8) formalizes the concatenation of all r_k values into vector $\mathbf{r}_{ij}$ to encode distances d_{ij} in a higher dimensional space.

$$\mathbf{r_{ij}} = [r_0, r_1, \ldots, r_{K-1}] \in \mathbb{R}^K \tag{8}$$

The director vector, $\mathbf{v}_{ij}$, and the RBF-transformed distances, $\mathbf{r}_{ij}$, are concatenated and passed through a Multi-Layer Perceptron (MLP) to produce the edge feature vectors. The $\mathrm{MLP_{edge}}$ consists of one first linear layer that doubles the existing dimension, a SiLU, another linear layer that returns to the original dimension, and a final SiLU. This distance and direction information combination ensures that the edge encoding captures the geometric relationships necessary for accurate modelling of the geometric structure. Eqn (9) describes the edge encoding process mathematically.

$$\mathbf{e_{ij}} = \mathrm{MLP_{edge}}\left(\mathbf{r_{ij}}(d_{ij}) \parallel \hat{v}_{ij}\right) \in \mathbb{R}^{\dim} \tag{9}$$

4.1.3 CartLayer. The CartLayers are responsible for aggregating the information between nodes through message passing. They consist of two key components: the gating mechanism and the message-passing mechanism. Fig. 9 illustrates a schematic of the CartLayer.

Our gating mechanism considers the sender and receiver atoms and the edge attributes connecting them and processes them through an $\mathrm{MLP_{gate}}$. The $\mathrm{MLP_{gate}}$ consists of a linear layer that reduces the dimensions from 3dim to the dim, followed by a SiLU, and another linear layer that does not modify the dimensions. The gating mechanism has a dual purpose: determining the weight of the message and updating the edge attributes. Additionally, our gating mechanism incorporates an envelope function inspired by previous work in molecular systems.[27] Eqn (10) describes the gating mechanism.

$$\mathbf{gate_{ij}} = \mathrm{Sigmoid}\left(\mathrm{BN}\left(\mathrm{MLP_{gate}}\left(\mathbf{h_i} \parallel \mathbf{e_{ij}} \parallel \mathbf{h_j}\right)\right)\right) \odot \mathrm{Env}(d_{ij}) \in \mathbb{R}^{\dim} \tag{10}$$

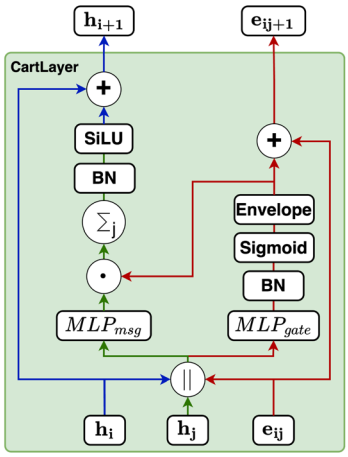

Fig. 9 Schematic of the CartLayer used in CartNet. This layer aggregates information between two neighbouring atoms, the sender atom $\mathbf{h}_i^l$ and the receiver atom $\mathbf{h}_j^l$, and their corresponding edge, $\mathbf{e}_{ij}^l$. It utilizes this aggregated information to update the receiving node's features and the edge's latent vector. The updated receiving node, $\mathbf{h}_i^{l+1}$ and the respective edge, $\mathbf{e}_{ij}^{l+1}$, representations are then propagated to subsequent layers for further processing.

$$\mathrm{Env}(d_{ij}) = \frac{1}{2}\left(\cos\left(\frac{\pi d_{ij}}{r_c}\right) + 1\right) \tag{11}$$

Here, $\mathbf{h_i}$ and $\mathbf{h_j}$ represent the hidden feature vectors of the receiver and sender atoms, respectively, while $\mathbf{e_{ij}}$ represents the hidden feature vector of the edge. BN stands for Batch Normalization. As described by eqn (11), the envelope function applies a cosine decay over distance.

The envelope function equalizes the influence of edges based on distance. This equalization helps the model to detect the peaks from the distribution, making it easier to identify the different interatomic interactions from our dataset.

Moreover, the envelope function also softens the influence of neighbours near the cutoff distance, which is particularly valuable in noisy situations. It ensures that atoms near the cutoff radius gradually lose influence, preventing them from being considered neighbours based strictly on a hard cutoff radius. Since our dataset is derived from experimental data, noise is expected, and the envelope function provides a robust mechanism to handle such noise effectively. ESI Section S1† provides further information about the envelope.

Our message-passing mechanism constructs messages by processing the concatenated sender and receiver atom information and edge attributes through an $\mathrm{MLP_{msg}}$. The $\mathrm{MLP_{msg}}$ has the same configuration as that of the $\mathrm{MLP_{gate}}$. Once the message is created, it is weighted using the gate function, which determines the relative importance of each feature within the message vector. The weighted messages are aggregated at the receiving node and passed through a batch normalization layer, followed by a SiLU non-linearity. The gate mechanism also updates the edge features, ensuring that the edge attributes remain consistent with the evolving node representations. To mitigate the issue of gradient vanishing, a skip connection is incorporated into both the node and edge updates, preserving

the flow of information through the network layers. Eqn (12)–(14) mathematically describe the message used for message passing and how the atoms and the edges are updated.

$$\mathbf{msg_{ij}} = \mathrm{MLP_{msg}}(\mathbf{h_i} \parallel \mathbf{e_{ij}} \parallel \mathbf{h_j}) \odot \mathbf{gate_{ij}} \in \mathbb{R}^{\dim} \quad (12)$$

$$\mathbf{h_i}^{l+1} = \mathbf{h_i}^{l} + \mathrm{SiLU}\left(\mathrm{BN}\left(\sum_{j \in \mathcal{N}_i} \mathbf{msg_{ij}}\right)\right) \in \mathbb{R}^{\dim} \quad (13)$$

$$\mathbf{e_{ij}}^{l+1} = \mathbf{e_{ij}}^{l} + \mathbf{gate_{ij}} \in \mathbb{R}^{\dim} \quad (14)$$

Here, $\mathbf{h_{i+1}}$ and $\mathbf{e_{ij+1}}$ represent the updated node and edge features, respectively, $\mathbf{msg_{ij}}$ denotes the message vector created between atoms i and j, and $\mathcal{N}_i$ is the neighbourhood from the receiving atom.

4.1.4 Cholesky head. The head of our model is designed using Cholesky decomposition to ensure that all output matrices are symmetric and positive-definite, which is a critical requirement for ADPs. Fig. 10 shows a schematic of the Cholesky head used in CartNet.

The Cholesky decomposition says that any symmetric positive-definite matrix, such as ADPs, can be uniquely decomposed into the product of a lower triangular matrix and its transpose. This decomposition can be expressed with eqn (15).

$$\mathbf{U} = \mathbf{L}\mathbf{L}^{\mathbf{T}} \in \mathbb{R}^{3\times3} \quad (15)$$

where $\mathbf{L}$ is a lower triangular matrix, described by eqn (16).

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \in \mathbb{R}^{3\times3} \quad (16)$$

In this matrix, the diagonal elements l_{11}, l_{22}, and l_{33} are always positive, ensuring that the resulting matrix $\mathbf{U}$ is both symmetric and positive-definite.

Based on this mathematical foundation, the feature vector of each node from the final aggregation layer is processed through a $\mathrm{MLP_{head}}$ to produce a feature vector o_i with $i = 1, \ldots, 6$. The $\mathrm{MLP_{head}}$ consists of a linear layer that reduces the dimensions from dim to dim/2, followed by a SiLU, and another linear layer that reduces the dimensions from dim/2 to 6. The first three elements are activated using the softplus function.[33] In this context, the softplus activation ensures that the diagonal elements of the matrix $\mathbf{L}$ are strictly positive, which is essential for maintaining the positive-definite property of the output matrix. The remaining three elements of the feature vector are used as the lower off-diagonal elements of the matrix $\mathbf{L}$. The construction of the matrix $\mathbf{L}$ is as follows, where the first three elements of the feature vector are placed on the diagonal and the remaining elements are placed in the lower triangular part, as can be seen in eqn (17).

$$\mathbf{L} = \begin{bmatrix} \mathrm{Softplus}(o_1) & 0 & 0 \\ o_4 & \mathrm{Softplus}(o_2) & 0 \\ o_6 & o_5 & \mathrm{Softplus}(o_3) \end{bmatrix} \in \mathbb{R}^{3\times3} \quad (17)$$

Finally, the ellipsoid matrix $\mathbf{U^{pred}}$ is obtained by performing a matrix multiplication between $\mathbf{L}$ and its transpose, as can be seen in eqn (18).

$$\mathbf{U^{pred}} = \mathbf{L}\mathbf{L}^{\mathbf{T}} \in \mathbb{R}^{3\times3} \quad (18)$$

This construction ensures that the predicted ellipsoid matrix $\mathbf{U_i^{pred}}$ is always symmetric and positive-definite, which is essential for accurate modelling of ellipsoid matrices.

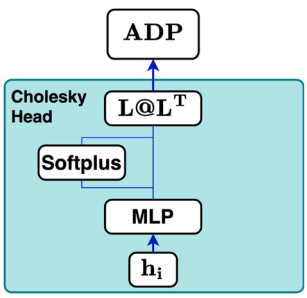

Fig. 10 Schematic of the Cholesky head used in CartNet. This layer enforces the creation of symmetric and positive-definite output matrices, which are necessary conditions for ADP matrices. The final hidden state $\mathbf{h_i}$ is processed by an $\mathrm{MLP_{head}}$ that outputs a vector of six elements. The first three elements are activated using a softplus function, while the remaining three remain unchanged. These six elements are utilized to construct the lower-triangular matrix L. The final ADP representation is obtained by multiplying L with its transpose, resulting in $\mathrm{LL^T}$.

4.2 Rotation SO(3) augmentation

Rotation SO(3) augmentation was applied to promote the model to generalize to unseen rotations. This augmentation was implemented by multiplying a random three-dimensional rotation matrix with the direction vector between two atoms, as described by eqn (19).

$$\hat{\mathbf{v}}_{\mathbf{ij}}^{\mathbf{aug}} = \mathbf{R}\hat{\mathbf{v}}_{\mathbf{ij}} \in \mathbb{R}^{1\times3} \quad (19)$$

In this equation, $\mathbf{R}$ is a random rotation matrix, where $\mathbf{R} \in \mathbb{R}^{3\times3}$, and $\hat{\mathbf{v}}_{\mathbf{ij}}$ represents the direction vector between two atoms. Our approach ensures that the model is regularly exposed to diverse rotational configurations during training, helping it learn features that generalize to previously unseen orientations. Although this effectively increases the complexity of the learning problem, online data augmentation is used to keep the additional overhead manageable. By adopting this well-established deep learning technique, we can rely on simpler model architectures without strictly enforcing equivariance, thereby reducing the computational cost per prediction and enhancing overall efficiency.

The situation is particularly nuanced for the ADP dataset because ellipsoids are inherently rotationally equivariant. Therefore, the ellipsoids must rotate consistently with the input data during augmentation. Eqn (20) describes how the rotation is applied to the original $\mathbf{U_i}$ to rotate it. The proof of eqn (20) can be found in the ESI Section S2.†

$$\mathbf{U}_i^{aug} = \mathbf{R}\mathbf{U}_i\mathbf{R}^T \in \mathbb{R}^{3\times 3} \quad (20)$$

5 Experiments and results

5.1 Computational details

The computational setup for all our experiments consisted of an NVIDIA RTX 3090 GPU with 24GB of memory and a system powered by 2× AMD EPYC 7313 16-core CPUs. All code implementations used PyTorch v1.13.1 (ref. 34) and PyTorch Geometric v2.3.[35] Theoretical calculations used the Vienna *Ab initio* Simulation package (VASP) c6.4.3 (ref. 36–38) and Phonopy v2.19.1[39,40] and were computed on the MareNostrum 5 HPC from the Barcelona Supercomputing Centre (BSC). The code is publicly available in the Github repository: **https://github.com/imatge-upc/CartNet**.

5.2 Results

To evaluate our model, we first tested its prediction performance on two well-known public datasets (JARVIS dataset[12] and the Materials Project dataset[41]) and our proposed ADP dataset. By applying our method to both proprietary and public datasets, we aimed to demonstrate its robustness and generalizability in predicting material properties under various conditions. It is important to note the very different nature of the materials and properties analysed among the three datasets. The two public datasets contain mainly simple bulk materials with properties extracted from electronic structure calculations, and the one developed in this work contains molecular materials using structural information and ADPs from experimental data.

5.2.1 JARVIS dataset results. The JARVIS 3D DFT Dataset (2021.8.18)[12,42] is a comprehensive dataset consisting of approximately 55k materials, basically bulk 3D materials, where various DFT properties were computed. The geometries of the crystal structures were optimized using the OptB88vdW (OPT) functional,[43] which gives accurate lattice parameters. The same functional was employed for the calculation of the different properties. Although in the case of the band gap, to get a better estimation of the value, additionally a small subset was also calculated with the Tran–Blaha modified Becke Johnson (MBJ) potential.[44] The dataset provides fundamental material properties: (i) formation energy: the energy change when forming a compound from its elements indicating thermodynamic stability, (ii) band gap (OPT): the energy difference between valence and conduction bands from standard DFT calculations, (iii) total energy: the ground-state energy of the crystal structure, (iv) band gap (MBJ): band gap computed using the Tran–Blaha modified Becke Johnson potential for improved accuracy, and (v) ehull: the energy above the convex hull, measuring stability against decomposition into other phases. Notably, the dataset contains only 18k samples for the band gap (MBJ) property, making it a low-data scenario.

In this study, we compared the results of our method against those of other previously reported methods, including Matformer,[21] PotNet,[22] eComFormer,[23] and iComFormer.[23] To ensure a consistent and fair evaluation, we followed the methodology proposed by Matformer[21] and used their proposed data splits. We use the mean absolute error (MAE) as our evaluation metric and report its mean and standard deviation across four random initialization seeds to confirm that our model's performance is robust rather than dependent on initialization. Section S4 in the ESI† provides the detailed CartNet modifications and training configurations used for predicting each property.

As illustrated in Table 1, our model consistently performs best across all evaluated properties. The improvements are most notable in the total energy and band gap (OPT) predictions, where our model demonstrates an improvement of 7.71% and 5.48%, respectively, over the next-best models. The low-data band gap (MBJ) scenario outperforms the next-best model by approximately 2.68%.

Our model's ability to excel across various properties, including those with fewer data samples, such as the band gap (MBJ), highlights its adaptability and robustness. These results suggest that our approach performs well in traditional tasks and thrives in challenging low-data environments, providing a comprehensive solution for crystal structure property prediction.

5.2.2 The Materials Project dataset results. The Materials Project Dataset-2018.6.1 (ref. 41) contains approximately 69k structures collected from the Materials Project.[13] The dataset consists of inorganic crystalline materials, primarily bulk materials, where various DFT properties have been computed. The structures were optimized using the PBE[45]-D3(BJ)[46] level of theory. The dataset provides several essential material properties: (i) formation energy, (ii) band gap, (iii) bulk moduli: measures the resistance of a material to deformation under shear stress, reflecting how it deforms when forces are applied parallel to a surface and (iv) shear moduli: quantifies the resistance of a material to uniform compression, indicating

Table 1 MAE results for the different tested architectures in the test split from the JARVIS dataset. The best result is in bold and second best underlined. Arrows indicate the direction of improvement for each metric

Method	Form energy (meV per atom) ↓	Band gap (OPT) (meV) ↓	Total energy (meV per atom) ↓	Band gap (MBJ) (meV) ↓	Ehull (meV) ↓
Matformer[21]	32.5	137	35	300	64
PotNet[22]	29.4	127	32	270	55
eComFormer[23]	28.4	124	32	280	<u>44</u>
iComFormer[23]	<u>27.2</u>	<u>122</u>	<u>28.8</u>	<u>260</u>	47
CartNet	**27.05 ± 0.07**	**115.31 ± 3.36**	**26.58 ± 0.28**	**253.03 ± 5.20**	**43.90 ± 0.36**

how much it compresses under external pressure. Notably, the dataset contains only around 5.5k shear and bulk modulus samples, making these properties particularly challenging. We directly compared our results with those of previous studies[21–23] without retraining these models, maintaining consistency across evaluations and using the same data splits. Similarly to the previous experiment, we employed the MAE as the evaluation metric with the splits defined in Matformer[21] and ran experiments using four different random seeds, reporting both the mean and standard deviation of the MAE. The CartNet training configurations for each property are detailed in Section S4 in the ESI.†

As shown in Table 2, our method achieves the best performance across all evaluated properties. The improvements are particularly notable for form. energy, yielding an approximately 4.33% improvement. Similarly, for bulk moduli's low-data scenario, our model improves the MAE by approximately 13.16% over the next-best model. In the shear moduli task, another low-data scenario, our method achieves the same metric as the best-known reported.

These results underscore the robustness of our model, especially in low-data environments such as bulk and shear moduli tasks, where limited training data pose significant challenges. The consistent improvements across all properties confirm that our approach is well suited for predicting a broad range of crystal structure properties and offers substantial advantages over existing methods.

5.2.3 ADP dataset. Our method has been compared against two other previously reported methods for the ADP dataset for material property prediction. eComformer and iComformer[23] have been selected for comparison since the other methods are based only on distance encoding and do not encode the geometry referenced to a 3D basis needed for the correct prediction of the ADP direction. To evaluate the ADPs we computed the MAE between the U matrices. Also, the similarity index (S_{12}) was also calculated since it is widely used in previous studies[47] to compare ADPs. S_{12} is based on the Bhattacharyya distance and represents the percentage error of the overlap between two multivariate Gaussian distributions.[47]

As shown in Fig. 4, small ADPs differ by several orders of magnitude compared to larger ones. This disparity could lead to biased conclusions when analysing the results using MAE, although the S_{12} does not have this problem. Nevertheless, the issue with the S_{12} is that it is not highly discriminative. Most of the errors are less than 1% using this metric, which might give the impression that the predictions are accurate. To address these issues, we also employed an Intersection over Union (IoU) metric over the ADP graphical ellipsoid representations. The IoU metric provides a measurement independent of the ADP's size, enabling us to assess whether smaller or larger ellipsoids are well-predicted. Furthermore, if an ADP is not well-predicted, it will have a higher impact on the IoU metric, making it more restrictive. The IoU is computed by voxelizing the 3D space. More details about the IoU implementation and the training configurations can be found in Sections S3 and S4 in the ESI,† respectively. For all models, we ran experiments using four initialization seeds to ensure robustness against initialization effects, reporting the mean and standard deviation.

Table 3 shows that our method achieves the best performance in all evaluation metrics, MAE, S_{12} and IoU, compared with other methods. In our experiments, CartNet outperforms the second-best model by 10.87% in MAE, 17.58% in S_{12}, and 2% in IoU. Furthermore, CartNet needs to train 49% fewer parameters than the second-best model. This result suggests that our approach can achieve state-of-the-art results without needing specific layers to enforce the rotational equivariance. The error introduced by using our data augmentation instead of using equivariant layers is discussed and evaluated in Section 6.5.

Fig. 11 shows a visual comparison between the compared methods. Even though all tested models can encode the 3D geometry, only our approach and iComformer can encode the

Table 2 MAE results for the different tested architectures in test split from the Material Project dataset. The best result is in bold and second best underlined. Arrows indicate the direction of improvement for each metric

Method	Form energy (meV per atom) ↓	Band gap (meV) ↓	Bulk moduli (log(GPa)) ↓	Shear moduli (log(GPa)) ↓
Matformer[21]	21	211	0.043	0.073
PotNet[22]	18.8	204	0.04	<u>0.065</u>
eComFormer[23]	<u>18.16</u>	202	0.0417	0.0729
iComFormer[23]	18.26	<u>193</u>	<u>0.038</u>	**0.0637**
CartNet	**17.47 ± 0.38**	**190.79 ± 3.14**	**0.033 ± 0.94 × 10^{-3}**	**0.0637 ± 0.0008**

Table 3 Results for the different tested architectures in the test split from the ADP dataset. The best result is in bold and second best underlined. Arrows indicate the direction of improvement for each metric

Method	MAE (Å^2) ↓	S_{12} (%) ↓	IoU (%) ↑	#Params ↓
eComformer[23]	$6.22 \times 10^{-3} \pm 0.01 \times 10^{-3}$	2.46 ± 0.01	74.22 ± 0.06	5.55 M
iComformer[23]	<u>$3.22 \times 10^{-3} \pm 0.02 \times 10^{-3}$</u>	<u>0.91 ± 0.01</u>	<u>81.92 ± 0.18</u>	<u>4.9 M</u>
CartNet	**$2.87 \times 10^{-3} \pm 0.01 \times 10^{-3}$**	**0.75 ±0.01**	**83.56 ±0.01**	**2.5 M**

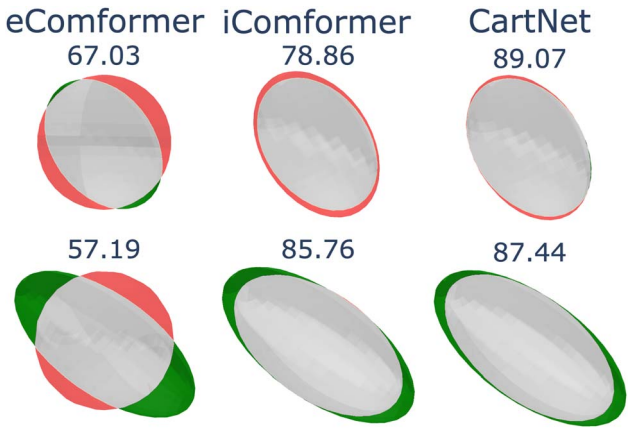

Fig. 11 Visual comparison of eComformer, iComformer, and CartNet on the ADP test split. The top row shows an ellipsoid with average anisotropy, and the bottom row shows the one with high anisotropy. Green indicates experimental values, red shows predicted values, and grey is their intersection. The IoU for each ellipsoid is shown above it.

geometry so that the ellipsoids can be oriented. On the other hand, eComformer only creates spherical ellipsoids. If we take a closer look at the eComformer architecture, it creates an invariant descriptor based on distance. This descriptor is then updated by an equivariant layer based on spherical harmonics and then by a few invariant layers based on distance. Our intuition suggests that even though this equivariant layer enriches the invariant information and can improve the invariant predictions, it is not able to successfully encode the needed information for a correct ADP orientation prediction.

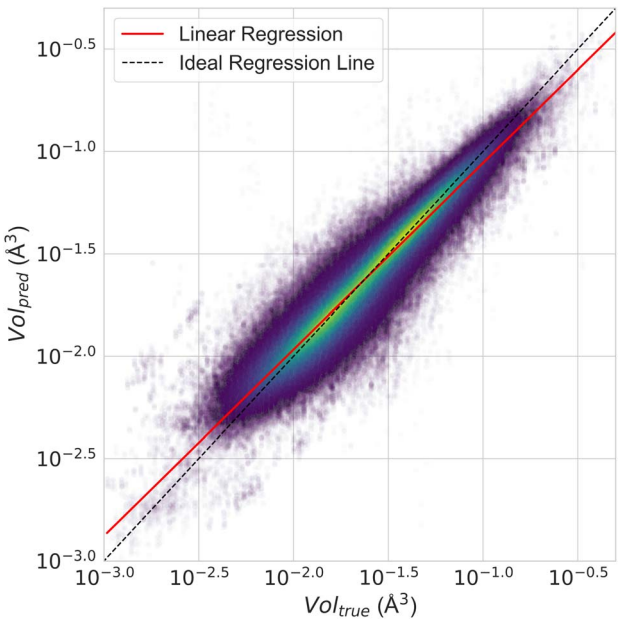

Fig. 12 Scatter plot of true *versus* predicted ADP volumes for the test split of the ADP dataset using CartNet. A linear regression fitted to the data yields $y = 1.0013x + 0.0021$ with an R^2 score of 0.91, indicating a correlation. The scale is logarithmic for both axes.

Entire crystal structure comparison between methods can be found in Section S8 in the ESI.†

Fig. 12 presents a scatter plot of the predicted ADP volumes *versus* the experimental values. The model achieves a coefficient of determination (R^2) of 0.91, demonstrating a strong linear correlation between the predictions and the actual volumes. This near-perfect linear regression indicates the model's high accuracy in predicting ADP volumes.

In Section 3, we discussed the data imbalance concerning temperature, volume, and elemental composition in the ADP dataset. We analysed the error distribution for temperature, volume, and element to determine if the model's errors align with these imbalances. Fig. 13 shows CartNet's predicted IoU as a function of temperature for the test split of the ADP dataset. The IoU decreases noticeably when the temperature drops below 80 K and when it increases above 400 K. This pattern corresponds with the data distribution depicted in Fig. 3, indicating higher errors in temperature regions that are underrepresented in the dataset. Nevertheless, the IoU is stable for the rest of the temperatures. Fig. 14 illustrates the IoU metric for CartNet's predictions across different volume ranges in the test split. Comparing this with Fig. 4, we observe that higher errors occur in volume ranges with fewer data points, especially at the extremes of the chart. Fig. 15 presents the IoU per element for CartNet's predictions on the ADP dataset. Lower metrics correspond to lower representation in the dataset for some elements, such as beryllium, technetium, caesium, and

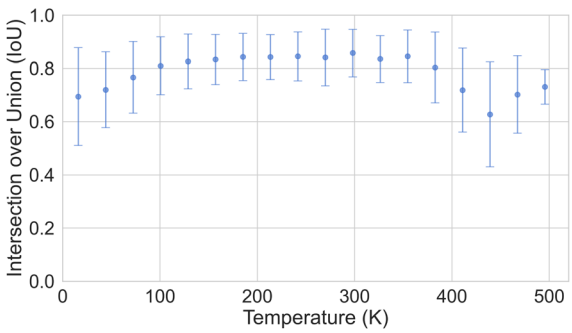

Fig. 13 Plot of the IoU error with standard deviation as a function of temperature for CartNet's predictions on the ADP test dataset.

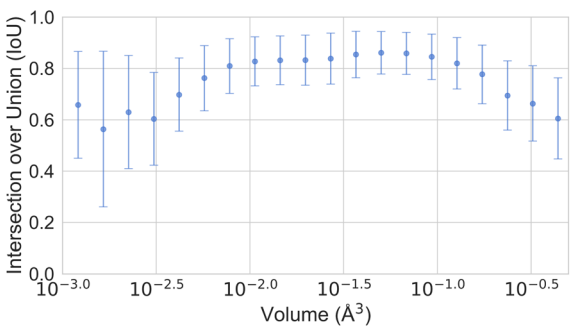

Fig. 14 Plot of the IoU error with standard deviation as a function of the ADP volume for CartNet's predictions on the ADP test dataset.

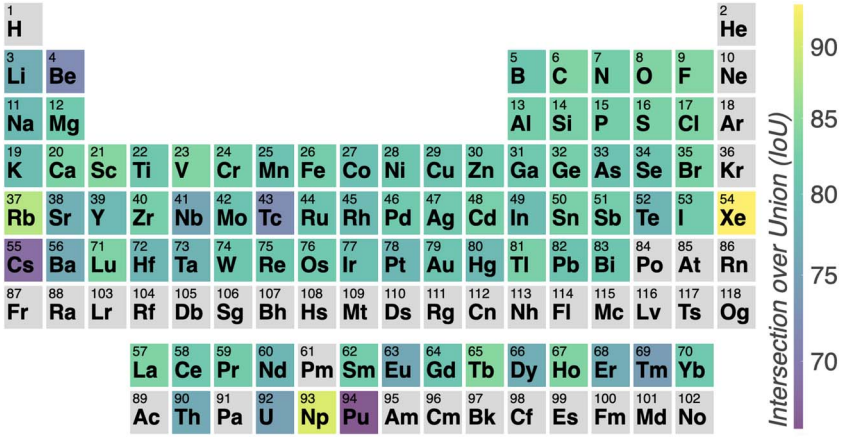

Fig. 15 Heatmap illustrating the IoU metric per element for the CartNet's predictions in the test split from the ADP dataset. Lighter colours represent higher IoU, while darker colours indicate lower IoU.

plutonium. Conversely, elements such as xenon and neptunium exhibit near-perfect IoU scores despite being underrepresented. Additionally, most of the other elements show similar IoU values, suggesting that the network is able to learn from atom numbers that appear more frequently in the dataset and generalize that information to other atomic numbers. Further discussion on the performance of different atom types in specific chemical interactions such as hydrogen bonding, π–π stacking, and *tert*-butyl groups, as well as additional results on different polymorphs, can be found in ESI Section S5.†

Moreover, we aimed to investigate whether our model could predict ADPs at various temperatures while using the same crystal geometry. To this end, we selected a crystal not included in our dataset, which had been synthesised at different temperatures. The ESI† from previous studies[48] indicated that the series of crystal structures of guanidinium pyridinium naphthalene-1,5-disulfonate, with the CSD refcode DOWVOC,[49] met these criteria. The CSD contains 14 entries for this crystal, covering a temperature range between 155 K and 283 K, all with well-defined ellipsoids. The list of CSD refcodes and the respective temperature can be found in Section S7 in the ESI.† We conducted two experiments to assess the ability of our model to predict ADPs at any temperature fixing the geometry. In the first experiment, we examined whether our model could accurately predict the ADPs for all data points of this crystal using the experimental geometries and temperatures. In the second experiment, we evaluated whether our system could predict ADPs by employing a fixed geometry while varying the input temperature. We used the fixed geometry at 213 K and systematically adjusted the input temperature values provided to CartNet. We computed the mean of the ellipsoid volumes from experiments, predicted, and predicted from the geometry for each temperature point.

Fig. 16 presents the results of these experiments. The results demonstrate that our model can predict ADPs across the entire temperature range. However, the predicted volume diverges when using a fixed geometry and varying the input temperature. These results suggest that cell expansion due to temperature significantly affects ADPs, as the error increases with the difference between the temperature at which we fixed the geometry and the temperature at which ADPs should be predicted. Nonetheless, when predicting ADPs around the temperature from the fixed geometry, they are estimated with high accuracy. Further discussion about using the fixed geometry at 150 K and 283 K can be found in Section S7 in the ESI.†

Finally, the last experiment was to compare with traditional theoretical methods. We compared the ADPs of our method with DFT calculations. Due to the high computational cost of computing ADPs with DFT, a single crystal structure (5,5′-dimethyl-2,2′-bipyrazine, CSD refcode: ETIDEQ) has been computed. The electronic structure calculation was performed using the Vienna *Ab initio* Simulation Package (VASP c6.4.3)[36–38] package with the optimized structure at the PBE[45]-D3(BJ)[46] level of theory. More details about the configuration used in electronic structure calculations can be seen in Section S6 in the ESI.†

In the case of DFT calculations, the choice of the central geometry for the atomic displacements can have a strong

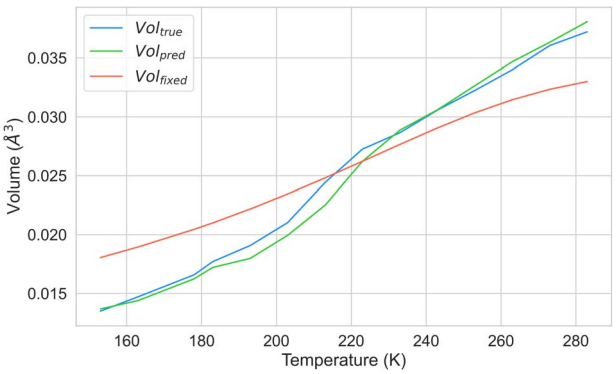

Fig. 16 Comparison of the mean ellipsoid volume as a function of temperature for the guanidinium pyridinium naphthalene-1,5-disulfonate (CSD refcode: DOWVOC) crystal structure. The blue line shows experimental data, the green line shows predicted volumes using experimental geometry and temperature, and the orange line shows predictions with a fixed 215 K geometry and varying input temperature in CartNet.

Table 4 Comparative ADP results between CartNet and DFT for the 5,5′-dimethyl-2,2′-bipyrazine crystal structure (CSD refcode: ETIDEQ). For the DFT calculations, three configurations have been tested. First, using atomic relaxation with a fixed volume, obtained by solving the Vinet equation. Second, atomic relaxation with a fixed lattice. Third, full optimization of the geometry. DFT calculations were done using 56 CPU cores from the MareNostrum 5 (ref. 50) HPC, while CartNet calculations were done using 1 GPU and 1 CPU core from our setup described in Section 5.1. The best result is in bold. Arrows indicate the direction of improvement for each metric

Method	MAE (Å^2) ↓	S_{12} (%) ↓	IoU (%) ↑	Time (s) ↓
DFT (Vinet)	1.32×10^{-2}	3.09	57.33	$\sim 2.88 \times 10^{6}$
DFT (fix latt.)	1.43×10^{-2}	4.12	70.75	$\sim 1.44 \times 10^{6}$
DFT (full opt.)	3.25×10^{-3}	0.49	86.27	$\sim 2.88 \times 10^{6}$
CartNet	$\mathbf{2.12 \times 10^{-3}}$	**0.17**	**92.31**	$\sim \mathbf{10^{-2}}$

Table 5 Ablation results in the test split of the ADP dataset. Exp. no. 1 involves full CartNet using all contributions, exp. no. 2 creates the graph without the hydrogens, exp. no. 3 was trained without using the envelope to equalize the neighbours, exp. no. 4 was trained without using the direction unit vector between the neighbours, exp. no. 5 was trained without using the temperature of the crystal structure as input, and exp. no. 6 was trained without the SO(3) data augmentation proposed method. The best result is in bold and second best underlined. Arrows indicate the direction of improvement for each metric

Exp. no.	Method	MAE (Å^2) ↓	S_{12} (%) ↓	IoU (%) ↑
1	CartNet	$\mathbf{2.88 \times 10^{-3}}$	**0.75**	**83.53**
2	w/o hydrogens	3.28×10^{-3}	0.94	81.74
3	w/o envelope	3.04×10^{-3}	<u>0.77</u>	<u>83.21</u>
4	w/o $\hat{\mathbf{v}}_{ij}$	6.23×10^{-3}	2.46	74.17
5	w/o temperature	3.04×10^{-3}	0.85	82.22
6	w/o SO(3) Aug	<u>3.02×10^{-3}</u>	0.81	82.78

impact on the accuracy of the calculated ADPs. In this case, three geometries were tested: (i) a full optimization, considering atomic positions and lattice parameters, (ii) geometry relaxation including atoms but fixing the lattice to its crystallographic dimensions and (iii) an atomic relaxation with a fixed volume, obtained by solving the Vinet equation of state. The latter calculation is the most sophisticated since it involves the calculation of the change in free energy with respect to the compression and expansion of the cell to derive a cell volume at a given temperature. However, the best results were obtained for the full optimization, which is a simpler method in comparison. Regarding the computational cost of this calculation, it is interesting to observe how the fixed lattice calculation required half the geometrical displacements compared to Vinet and full optimization, as this geometry retained inversion symmetry after structural optimization. Thus, the displacements over symmetry equivalent atoms were redundant and, hence, omitted.

Table 4 shows the numerical results for this comparison. As can be seen, our model still improves the MAE by 34.77% and the IoU by 6.04%. Additionally, CartNet shows real improvement in the computation time, which was reduced by several orders of magnitude. Fig. 17 compares the ADPs from CartNet and the best DFT results (full optimization). In both cases, ellipsoids closely match the experimental reference, in line with their low MAE values.

6 Ablation studies

We conduct comprehensive ablation studies to assess the contribution of each component in our proposed model. By systematically removing or altering specific elements of the architecture, we aim to understand the impact of each part on the overall performance. This analysis allows us to identify which components are crucial for achieving high accuracy and provides insights into the model's inner workings.

Table 5 presents the results of removing each component of our proposed method. The following subsections explain the detailed experiments done and discuss their implications in the design of our model.

6.1 Impact of the hydrogens

The ADP dataset used in this study does not contain experimental ADP data for the hydrogen atoms, which makes determining how to handle these atoms during the graph

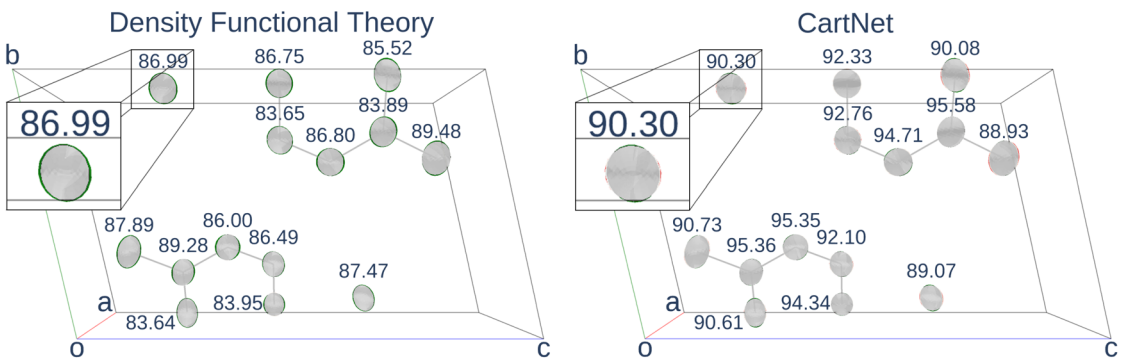

Fig. 17 Thermal ellipsoids representations from experimental ADPs for the 5,5′-dimethyl-2,2′-bipyrazine crystal structure (CSD refcode: ETIDEQ) predicted using DFT and CartNet, respectively. The green regions represent the experimental values, the red ones represent the prediction values, and the grey represents the intersection between them. The numbers in each atom represent the IoU between the experimental and the calculated ADP. Highlighted can be seen a sample ellipsoid predicted using the DFT and the same ellipsoid using CartNet. The parallelepiped represents the unit cell, and the red, green, and blue lines correspond to the *a*, *b*, and *c* unit cell axes.

construction process challenging. Our proposed methodology takes advantage of the available 3D coordinates of the hydrogen atoms by including them as additional atoms in the graph but inferring ellipsoids solely for the non-hydrogen atoms. However, we also tested an approach that ignored all hydrogen atoms during the creation of the graph and computed the ellipsoids using only the remaining atoms.

As seen in the experiment number 2 from Table 5, the results indicate that excluding the hydrogen atoms decreases the IoU from 83.53% to 81.74%, resulting in a 1.32% decrease in performance. For the other metrics, an improvement of 14.89% can be seen in MAE and 0.19% in S_{12}.

The results suggest that hydrogen atoms, while not directly involved in inferring ellipsoids, contribute valuable contextual information to the graph. This additional data benefit the model, allowing it to more accurately infer the ADPs of non-hydrogen atoms by incorporating the effects of the covalently bonded hydrogens and include the relevant hydrogen bond interactions and other intermolecular forces.

6.2 Neighbour equalization

The neighbour equalization technique is a novel method for equalizing the number of neighbours. It uses an envelope function to weight the contributions of atoms with respect to their distance. It addresses the challenge of distant atoms disproportionately influencing the aggregation process and helps detect the peaks of the different interatomic interactions.

Experiment number 3 in Table 5 shows the results when training CartNet without the envelope function. Compared to full CartNet (experiment no. 1), the ADP prediction experiments show a drop of 0.32% for the IoU, 0.02% for the S_{12}, and 5.26% for the MAE. This suggests that using neighbour equalization with the envelope function is highly effective in equalizing the neighbours.

6.3 Cartesian axis

This ablation study investigates the effect of incorporating the Cartesian direction vector ($\hat{\mathbf{v}}_{ij}$) between atoms in the edge encoder on the CartNet model. The direction vector captures important geometric information about the relative orientation of atoms, which is expected to influence the accuracy of ADP predictions.

The results presented in experiment number 4 in Table 5 clearly show the significant impact of incorporating the Cartesian direction unit vector in the edge encoder. The 9.36% decrease in IoU, the 1.71% drop in the S_{12}, and the 53.77% reduction in MAE highlight the ability of the model to capture the spatial relationships between atoms when using the direction vector. This suggests that the direction unit vector is a crucial input feature for effectively encoding the geometric information necessary for accurate ADP prediction.

6.4 Temperature

This ablation study explores the impact of including temperature as an input feature on the performance of the model in predicting ADPs. Temperature plays a crucial role in atomic displacements, and incorporating it as an input feature can enhance the ability of the model to capture temperature-dependent behaviours in ADPs.

The experiment 5 in Table 5 shows the results when not including the temperature information in the input of the CartNet model. The decreases 1.31% in IoU, 0.08% in the S_{12}, and 4.63% in MAE demonstrate that the ability of the model to capture the spatial extent of atomic displacements is enhanced when the temperature is included. These results suggest that temperature is essential for improving the performance of models in predicting ADPs, particularly when considering the thermal motion of atoms.

6.5 SO(3) data augmentation

The augmentation was explicitly designed to enhance the model's ability to learn to generalize unseen rotations, which is critical when dealing with 3D molecular and crystal structures where the orientation of the input can vary.

For these experiments, we trained our model on the ADP dataset with and without rotation SO(3) augmentation. The configuration without augmentation exposed the model to the data in its original orientation. In contrast, the configuration utilizing rotation SO(3) augmentation involved randomly applying three-dimensional rotations to the direction vectors of the atoms during training, thereby encouraging the model to learn features equivariant to spatial orientation.

Experiment number 6, compared with experiment number 1 from Table 5, demonstrates the positive impact of the rotation SO(3) augmentation on model performance. The decrease of 0.75% in IoU, 0.06% in S_{12}, and 4.63% in MAE demonstrates that the model is more adept at generalizing to unseen orientations of the input data, making it better suited for real-world applications where molecules and crystal structures can appear in various spatial configurations.

Since CartNet does not explicitly enforce SO(3) rotation equivariance but learns it through data augmentation, we wanted to evaluate the error this method introduces into the final predictions. To quantify this, we defined two variables: $\mathbf{U}_{\text{orig}}$ represents the ADP predictions from the original, unrotated crystal structures and $\mathbf{U}_{\text{rot}}$ represents the ADP predictions from the crystal structures after they have been rotated by a rotation matrix $\mathbf{R}$. We then compared $\mathbf{U}_{\text{rot}}$ with the rotated versions of $\mathbf{U}_{\text{orig}}$, calculated as $\mathbf{R}\mathbf{U}_{\text{orig}}\mathbf{R}^{\mathbf{T}}$. This methodology allows us to isolate rotation-induced errors, ensuring that any observed discrepancies are attributable solely to rotation effects rather than a combination of prediction and rotation errors. To perform this comparison, we conducted a Monte Carlo experiment, applying 100 different random rotation matrices to the test set.

The results of this experiment yielded a Mean Absolute Error (MAE) of 1.01×10^{-3} Å^2 $\pm$ 1.68×10^{-3} Å^2, an S_{12} score of 0.65% $\pm$ 0.17%, and an Intersection over Union (IoU) of 94.96% $\pm$ 2.46%. Fig. 18 illustrates the results of this experiment.

In all cases, the ADP predictions for the rotated crystal structures ($\mathbf{U}_{\text{rot}}$) closely matched the rotated predictions of the original structures ($\mathbf{R}\mathbf{U}_{\text{orig}}\mathbf{R}^{\mathbf{T}}$). These results confirm that

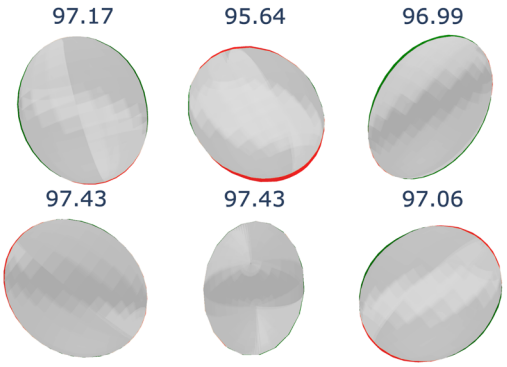

Fig. 18 Visualization of rotational errors by comparing the rotated ADP predictions from the original crystal structures (green ellipsoids, $RU_{orig}R^T$) with the ADP predictions from the rotated crystal structures (red ellipsoids, U_{rot}). The overlapping regions are shaded in grey, representing the intersection between the two predictions. The IoU values are displayed above each pair of ellipsoids.

CartNet effectively generalizes to unseen rotations despite not explicitly enforcing rotation equivariance.

7 Conclusions

In this work, we introduced CartNet, a novel GNN architecture designed to predict properties of crystalline molecular-based structures. In the specific test case examined, our model significantly reduces computational costs while demonstrating improved performance relative to the low-level GGA functional DFT calculations and current state-of-the-art learning-based architectures. Nonetheless, more advanced DFT formulations (*e.g.*, hybrid functionals with many-body dispersion), which are considerably more computationally expensive, may offer higher accuracy and warrant further investigation to fully assess the broader performance benefits of our approach. The development of CartNet was driven by the need to address the challenges posed by ADPs, which are crucial for understanding thermal vibrations in crystallography.

CartNet utilizes a novel Cartesian encoding approach that avoids reliance on the unit cell, thereby overcoming limitations faced by previous models. The incorporation of neighbour equalization helps the model to differentiate between various types of bonds and interaction forces between atoms. The Cholesky-based output layer ensures that the model generates valid ADP predictions that align with physical requirements. Additionally, by introducing a rotational generalization through data augmentation, CartNet effectively learns the directional nature of atomic vibrations without relying on specific equivariant layers. The evaluation of CartNet demonstrated its robustness and accuracy. It outperformed previously reported methods in other benchmarks, JARVIS and The Materials Project, that focused on bulk materials, instead of molecular systems, and contained structure and properties that have been computed with DFT calculations, instead of experimental structures and ADPs.

In addition to the model, we curated and presented a comprehensive ADP dataset containing over 200k crystal structures of molecular systems from the Cambridge Structural Database (CSD). This dataset spans a wide range of temperatures and atomic environments, providing a valuable resource for further research on predicting anisotropic displacements and thermal behaviours in crystalline structures.

This work provides a more efficient and accurate method for predicting properties in crystal structures, opening new possibilities for studying different material properties and designing new materials. Therefore, when CartNet is specifically used to predict ADPs, it can be used to evaluate the experimental results of new systems in cases where their experimental determination using diffraction techniques presents difficulties.

Future work could focus on predicting cell expansion to estimate ellipsoids at other temperatures based on a fixed geometry at a specific temperature. Regarding the specific case of the ADP, future work could explore the creation of equivariant methods to generate valid ADP matrices. Moreover, due to the large number of molecular crystal structures in the ADP dataset, future work can study using this dataset as pre-training for other crystal structure tasks. Finally, the efficiency and accuracy of CartNet also highlight its potential for future crystal structure prediction challenges, such as the 7th Blind Test[51,52] organized by the Cambridge Crystallographic Data Centre (CCDC),[53] where handling highly flexible or disordered systems remains a critical obstacle.

Data availability

The code to generate the ADP dataset and recreate the results of the paper can be found at: **https://github.com/imatge-upc/CartNet**. The project website with online demo available at: **https://www.ee.ub.edu/cartnet**.

Author contributions

Àlex Solé: data curation, conceptualization, methodology, software, investigation, validation, writing – review & editing. Albert Mosella-Montoro: conceptualization, supervision, resources, writing – review & editing. Joan Cardona: conceptualization, writing – review & editing. Silvia Gómez-Coca: conceptualization, supervision, resources, writing – review & editing. Daniel Aravena: conceptualization, supervision, resources, validation, writing – review & editing. Eliseo Ruiz: conceptualization, supervision, resources, writing – review & editing, funding acquisition. Javier Ruiz-Hidalgo: conceptualization, supervision, resources, writing – review & editing, funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Financial support from Ministerio de Ciencia e Innovación (project PID2020-117142GB-I00, PID2021-122464NB-I00, TED2021-129593B-I00, CNS2023-144561 and Maria de Maeztu CEX2021-001202-M) is acknowledged. We also acknowledge the

Generalitat de Catalunya for the 2021-SGR-00286 grant. E. R. also acknowledges the Generalitat de Catalunya for an ICREA Academia grant. We thank BSC for the computational resources.

Notes and references

1 D. W. J. Cruickshank, *Acta Crystallogr.*, 1956, **9**, 754–756.
2 D. W. J. Cruickshank, *Acta Crystallogr.*, 1956, **9**, 1005–1009.
3 S. C. Capelli, A. Albinati, S. A. Mason and B. T. M. Willis, *J. Phys. Chem. A*, 2006, **110**, 11695–11703.
4 R. Stoffel, C. Wessel, M.-W. Lumey and R. Dronskowski, *Angew. Chem., Int. Ed.*, 2010, **49**, 5242–5266.
5 K. Jarzembska, A. Hoser, R. Kamiński, A. Madsen, K. Durka and K. Woźniak, *Cryst. Growth Des.*, 2014, **14**, 3453–3465.
6 D. W. J. Cruickshank, *Acta Crystallogr.*, 1956, **9**, 1010–1011.
7 A. L. Goodwin, M. Calleja, M. J. Conterio, M. T. Dove, J. S. O. Evans, D. A. Keen, L. Peters and M. G. Tucker, *Science*, 2008, **319**, 794–797.
8 P. Pavone, K. Karch, O. Schütt, D. Strauch, W. Windl, P. Giannozzi and S. Baroni, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **48**, 3156–3163.
9 V. L. Deringer, J. George, R. Dronskowski and U. Englert, *Acc. Chem. Res.*, 2017, **50**, 1231–1239.
10 J. George, A. Wang, V. L. Deringer, R. Wang, R. Dronskowski and U. Englert, *CrystEngComm*, 2015, **17**, 7414–7422.
11 About the Cambridge Structural Database|CCDC — ccdc.cam.ac.uk,**https://www.ccdc.cam.ac.uk/solutions/about-the-csd/**, accessed 04-02-2024.
12 K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, *et al.*, *npj Comput. Mater.*, 2020, **6**, 173.
13 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.
14 M. N. Burnett and C. K. Johnson, *ORTEP-III: Oak Ridge Thermal Ellipsoid Plot Program for Crystal Structure Illustrations*, Oak Ridge National Laboratory, Technical Report ORNL-6895, 1996.
15 N. J. Lane, S. C. Vogel, G. Hug, A. Togo, L. Chaput, L. Hultman and M. W. Barsoum, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2012, **86**, 214301.
16 A. Togo, *J. Phys. Soc. Jpn.*, 2023, **92**, 012001.
17 F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini, *IEEE Trans. Neural Networks*, 2009, **20**, 61–80.
18 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *International Conference on Machine Learning*, 2017, pp. 1263–1272.
19 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
20 G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang and G. Ke, *The Eleventh International Conference on Learning Representations*, 2023.
21 K. Yan, Y. Liu, Y. Lin and S. Ji, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 15066–15080.
22 Y. Lin, K. Yan, Y. Luo, Y. Liu, X. Qian and S. Ji, *International Conference on Machine Learning*, 2023, pp. 21260–21287.
23 K. Yan, C. Fu, X. Qian, X. Qian and S. Ji, *International Conference on Learning Representations*, 2024.
24 J. Gasteiger, J. Groß and S. Günnemann, *International Conference on Learning Representations (ICLR)*, 2020.
25 J. Gasteiger, S. Giri, J. T. Margraf and S. Günnemann, *Machine Learning for Molecules Workshop*, NeurIPS, 2020.
26 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 11423–11436.
27 G. Simeon and G. De Fabritiis, Advances in Neural Information Processing Systems, *TensorNet: Cartesian Tensor Representations for Efficient Learning of Molecular Potentials*, ed. A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, Curran Associates, Inc., 2023, vol. 36, pp. 37334–37353, **https://proceedings.neurips.cc/paper_files/paper/2023/file/75c2ec5f98d7b2f50ad68033d2c07086-Paper-Conference.pdf**.
28 M. Geiger and T. Smidt, *arXiv*, 2022, preprint, arXiv:2207.09453, DOI: **10.48550/arXiv.2207.09453**.
29 B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán and S. Alvarez, *Dalton Trans.*, 2008, 2832–2838.
30 S. Alvarez, *Dalton Trans.*, 2013, **42**, 8617–8636.
31 Y. Bengio, R. Ducharme and P. Vincent, Neural Probabilistic Language Models, *Innovations in Machine Learning: Theory and Application*, Springer, 2006, pp. 137–186.
32 S. Elfwing, E. Uchibe and K. Doya, *Neural Network*, 2018, **107**, 3–11.
33 V. Nair and G. E. Hinton, *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
34 PyTorch—pytorch.org, **https://pytorch.org**, Accessed 22-09-2024.
35 PyG Documentation—pytorch-geometric.readthedocs.io, **https://pytorch-geometric.readthedocs.io**, Accessed 22-09-2024.
36 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
37 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15–50.
38 G. Kresse and J. Hafner, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1993, **47**, 558–561.
39 A. Togo, *J. Phys. Soc. Jpn.*, 2023, **92**, 012001.
40 A. Togo, L. Chaput, T. Tadano and I. Tanaka, *J. Phys.: Condens. Matter*, 2023, **35**, 353001.
41 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
42 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 185.
43 J. c. v. Klimeš, D. R. Bowler and A. Michaelides, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **83**, 195131.
44 D. Rai, M. Ghimire and R. Thapa, *Semiconductors*, 2014, **48**, 1411–1422.
45 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

46 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
47 A. E. Whitten and M. A. Spackman, *Acta Crystallogr., Sect. B*, 2006, **62**, 875–888.
48 A. van der Lee and D. G. Dumitrescu, *Chem. Sci.*, 2021, **12**, 8537–8547.
49 C. Shi, B. Wei and W. Zhang, *Cryst. Growth Des.*, 2014, **14**, 6570–6580.
50 MareNostrum 5—bsc.es, **https://www.bsc.es/marenostrum/marenostrum-5**, accessed 27-09-2024.
51 L. M. Hunnisett, J. Nyman, N. Francia, N. S. Abraham, C. S. Adjiman, S. Aitipamula, T. Alkhidir, M. Almehairbi, A. Anelli, D. M. Anstine, *et al.*, *Struct. Sci.*, 2024, **80**, 517–547.
52 L. M. Hunnisett, N. Francia, J. Nyman, N. S. Abraham, S. Aitipamula, T. Alkhidir, M. Almehairbi, A. Anelli, D. M. Anstine, J. E. Anthony, *et al.*, *Struct. Sci.*, 2024, **80**, 548–574.
53 Advancing Structural Science|CCDC — ccdc.cam.ac.uk, **https://www.ccdc.cam.ac.uk**.