Supplementary Information for: CrystalX: High-accuracy Crystal Structure Analysis Using Deep Learning

Kaipeng Zheng[1,2], Weiran Huang[1,2*], Wanli Ouyang[2], Han-Sen Zhong[2*], Yuqiang Li[2*]

[1]MIFA Lab, Qing Yuan Research Institute, SEIEE, Shanghai Jiao Tong University, Shanghai, China.
[2]Department, Shanghai Artificial Intelligence Laboratory, Shanghai, China.

*Corresponding author(s). E-mail(s): weiran.huang@sjtu.edu.cn; zhonghansen@pjlab.org.cn; liyuqiang@pjlab.org.cn;

Contents

List of Figures

List of Tables

1 Supplementary Methods

1.1 Experimental X-ray Diffraction Dataset from Crystallography Open Database

1.1.1 Data preparation

The dataset is derived from the Crystallography Open Database (COD), a widely used open-access repository of crystal structures. It aggregates comprehensive data from crystal structures published in top-tier journals across chemistry, crystallography, and materials science. Our focus is on entries that include authentic experimental observation records. After excluding samples lacking these records, the COD database contains a total of 77,615 entries.

Each entry includes both an HKL file, containing the experimental diffraction data, and a Crystallographic Information File (CIF). The diffraction data typically consist of a set of Miller indices (h, k, l) along with their corresponding observed intensities (F^2_{obs}) and associated uncertainties $(\sigma(F^2_{obs})$. Some samples provide individual HKL files recording diffraction data, while others embed the data in CIF files. For these, we use PLATON [1] to extract diffraction data and generate corresponding HKL files. The atomic positions recorded in the CIF align with the structure analysis achieved by human experts from the experimental observations.

1.1.2 Coarse electron density phasing from real-world experimental observations

Modern software programs [2–6] widely support automatic phasing from experimental diffraction measurements to generate coarse electron density. We use SHELXT [2] for this purpose due to its simplicity and widespread use, but it is by no means the only option. We follow standard procedures in real-world applications by constructing the

initial INS file required by SHELXT using only pre-existing crystal information. Once the INS and corresponding HKL files are available, the coarse electron density can be automatically calculated with commands such as `shelxt filename`. The resulting RES file contains a point cloud derived from this coarse electron density.

1.1.3 Target matching between coarse electron density and human experts' analysis

The atom positions documented in the CIF can be interpreted as the ground truth for structure analysis, as determined by expert crystallographers. We leverage this information to annotate the coarse electron density for model training. Specifically, we begin by calculating the distances between hydrogen atoms and each non-hydrogen atom, allowing us to label the number of hydrogen atoms for each non-hydrogen atom. To annotate non-hydrogen atoms, we calculate the distance between each peak and each non-hydrogen atom by applying crystal symmetry operations and periodic transformations to determine the pairing result.

Human experts can freely translate atoms without violating crystal symmetry. For instance, in the $P2_1$ space group, movement along the b-axis can be arbitrary. However, such translations are often unknown, making data matching challenging. Additionally, handling disorder—where multiple atoms occupy several positions with a combined occupancy of 1—can complicate peak matching. Due to these complexities, we excluded samples with disorder or unknown translation and symmetry operations from our dataset. The final experimental dataset contains 51,433 entries, representing 66% of the total. The overview of the data preparation process is shown in Fig. 10.

1.1.4 Crystallographic Statistical Details of the Dataset

We present the statistical details of the dataset in Fig. 11. The unit cell volumes cover a wide range, with the largest exceeding 10000 $\mathring{\text{A}}^3$. A total of 86 space groups are included, with the six most common being P1 21/c 1, P -1, P1 21/n 1, C 1 2/c 1, P 21 21 21, and P b c a. We also provide information on the experimental conditions for the diffraction data, including the maximum diffraction angle and signal-to-noise ratio. For most of the data, the maximum diffraction angles fall between 20 and 40 degrees. A smaller portion of the data involves high-angle diffraction, with maximum diffraction angles around 70 degrees. The signal-to-noise ratio reflects the quality of the diffraction data produced, covering a wide range of data quality.

1.2 Details for non-hydrogen elemental determination

1.2.1 Model configuration

The model for determining non-hydrogen elements leverages TorchMD-NET [7]. We utilize the official implementation, which incorporates a representation learning module called `TorchMD_ET` and an output module named `EquivariantScalar`. We adhere to the default hyperparameter settings provided in the official implementation, as outlined in Table 10.

1.2.2 Training details

The coarse electron density is characterized by peaks, each defined by its coordinates and corresponding charge density. We perform a coordinate transformation to convert the peak coordinates from a fractional coordinate system to the Cartesian coordinate system. Charge density can distinguish elements with large differences in atomic numbers. However, it struggles to distinguish between elements with similar atomic numbers, such as carbon, nitrogen, and oxygen, which are most common in organic compounds. In our implementation, we start by assigning initial elemental guesses to peaks based on charge density alone. For peaks with similar charge density values, we introduce variability by randomly shuffling these initial element assignments, serving as data augmentation to enhance model robustness. This also implies that knowing the exact elemental composition isn't necessary for our method. As a result, the model training process can be viewed as making accurate predictions from a diverse range of random guesses, as shown in Fig. 12.

For model training, the input consists of the 3D Cartesian coordinates of the peaks and the categorical encoding of the initially guessed element types. We use the Adam optimizer [8] with an initial learning rate set to 0.0001, which is adjusted according to a cosine decay schedule based on the number of training epochs. The model was trained from scratch for 100 epochs using a cross-entropy loss function. The training was efficient, requiring 2-3 hours on a single RTX 4090 GPU, making it highly scalable for future data. To avoid bias in the training-validation split, we performed further 10-fold cross-validation on the training-validation set, achieving a consistent elemental accuracy of $99.80 \pm 0.01\%$. The model with the best performance on the validation set was used to evaluate the performance on the reserved test set.

1.2.3 Comparison of model architecture

We utilize TorchMD-NET as our model architecture. Although it is not the only option, we show its high efficiency compared to other geometric deep learning models. Specifically, we evaluate various representative models for non-hydrogen atom determination on both accuracy and inference speed (i.e., iterations per second (it/s)), as detailed in Table 11.

TorchMD-NET demonstrates superior performance and high inference speed, surpassing the recent CoMeNet architecture, making it highly efficient for practical applications. Furthermore, the results suggest that better performance is typically yielded by a more comprehensive grasp of geometric patterns. While SchNet relies solely on distances, models like DimeNet, SphereNet, CoMeNet, and TorchMD-NET extract angular patterns or equivariant features of coordinates, which contribute to their enhanced performance.

1.2.4 Reliability of model's confidence

We further validated the reliability of the model's predicted probabilities using the Receiver Operating Characteristic (ROC) curve. Ideally, the predicted probability should accurately represent the likelihood of a correct prediction, with higher probabilities indicating a greater certainty of correctness. This reliability is quantified by

the area under the ROC curve (AUC). Fig. 13 displays the ROC curves for the nine most common elements, all of which have AUC values exceeding 0.99, approaching the theoretical maximum of 1. This demonstrates the high reliability of the model's predicted probabilities.

1.2.5 Model robustness under severe noise

To further evaluate the model's robustness, we introduced substantial noise to simulate the challenge of interpreting electron density maps with very low resolution. Specifically, we degraded the coarse electron density by adding independent Gaussian noise to the 3D Cartesian coordinates of the peaks along each of the three dimensions. The model's performance was then tested on this degraded data, with the results displayed in Fig. 14.

Remarkably, even as the level of perturbation increased, the model's performance showed only a slight decline. When the noise standard deviation reached 0.1 Å—a level sufficient to disrupt most atomic distances and angles beyond their normal ranges—the model still maintained a unit-cell-level accuracy of 84.15%, underscoring its robustness and its potential to effectively analyze electron density maps with extremely low resolution.

1.3 Details for hydrogen determination

1.3.1 Joint modeling of intramolecular and intermolecular interaction patterns

To describe the chemical environment of hydrogen atoms, we identify equivalent atoms within a 3.2 Å radius of each non-hydrogen atom considering the crystal's symmetry and periodicity. These equivalent atoms act as auxiliary atoms, capturing intermolecular interactions crucial for determining hydrogen atoms, as shown in Fig. 15.

1.3.2 Model configuration and training details

The hydrogen determination model also leverages Torch-MD, utilizing the official implementation. The model configuration is the same as the one used for non-hydrogen elemental determination (Table 10). For model training, the input data includes the 3D Cartesian coordinates and types of non-hydrogen elements, along with any auxiliary atoms. The model is trained to predict the number of hydrogen atoms associated with each non-hydrogen atom. We employed the Adam optimizer [8] with an initial learning rate of 0.0001, which was progressively reduced using a cosine decay scheduler. The model was trained from scratch for 100 epochs, utilizing a cross-entropy loss function.

1.3.3 Reliability of model's confidence

Fig. 16 presents the ROC curves for hydrogen atom determination, all of which exhibit AUC values exceeding 0.99, demonstrating the high effectiveness of the model's predicted probabilities.

1.3.4 Hydrogen atoms in other chemical environments

The main types of hydrogen atoms examined in our study include: a single hydrogen atom bonded to a boron atom, one to three hydrogen atoms bonded to a carbon atom, one to three hydrogen atoms bonded to a nitrogen atom, and a single hydrogen atom bonded to an oxygen atom, encompassing various hybridization modes. These scenarios account for over 98.89% of all cases. The remaining instances typically occur in rare chemical environments with limited available data. For these rare cases, we also report the model's performance in predicting the number of hydrogen atoms, as illustrated in Table 12.

Compared to typical hydrogen environments, predicting hydrogen atoms in rare chemical environments—such as those involving phosphorus, sulfur, and chlorine—shows reduced accuracy. This decline is primarily due to the substantially smaller dataset available for these uncommon environments, which hinders model performance.

1.4 Details for crystallography computation and error correction

1.4.1 Standardized crystallography comparison between the model and human experts

We employed SHELXL to evaluate the crystallographic metrics, R_1 and S, which serve as indicators of the alignment between structure analysis and the experimental data. To ensure a fair comparison between the analysis results of human experts and the model, we maintained the default settings of the SHELX suite for merging reflections and avoided any omissions, ensuring that the comparison was based on an identical set of reflections. Additionally, we standardized the refinement process across all evaluations, focusing on a concise, automated procedure that included essential steps: anisotropic refinement, placement of hydrogen atoms, and weight adjustments.

1.4.2 Heavy (non-hydrogen) element comparison

We began by separately writing the analysis of non-hydrogen atoms from the model and human experts into an INS file, followed by 10 cycles of least squares anisotropic refinement. All other lines in the INS file were kept completely consistent, with only the atom lines different. The parameters follow the default settings of SHELXL.

1.4.3 The placement of hydrogen atoms

We employed the riding model in SHELXL to position hydrogen atoms (i.e., HFIX and AFIX commands). The placement process generally follows a consistent procedure across different environments: based on the expected number of hydrogen atoms and the structure of non-hydrogen atoms, along with the residual electron density identified in the difference Fourier map, the initial positions of hydrogen atoms are determined. These positions are then refined through subsequent least-squares adjustments. We conduct hydrogen atom placement in typical environments. In other environments, the placement method remains similar to that used in common scenarios. However,

due to the absence of straightforward commands to facilitate this, we skipped placing hydrogen atoms in these cases for simplicity.

Finally, we conducted an additional 10 cycles, adjusting the weights as recommended by SHELXL. This process yielded the final crystallographic metrics.

1.4.4 Additional details for error correction

Error correction is performed by comparing the R_1 values obtained from expert human analyses with those from model analyses. The R_1 value indicates the consistency between the calculated structure factors and the observed diffraction intensities from the experiment. A lower R_1 value signifies a better fit, with a smaller residual. Since the reflection set and refinement conditions are kept consistent, a correct structure is expected to yield a lower R_1 value. If the R_1 value from human analysis is significantly higher than that from the model, it suggests a high probability of error.

Additionally, the errors corrected by the model were published in top-tier (JCR Q1) journals, which had already undergone rigorous review by human experts. These errors are subtle and typically cannot be identified through AB-level structure alerts provided by CheckCIF [1], which is widely recognized as a very rigorous publication standard for crystal structure analysis. For instance, despite the absence of AB-level alerts in the CheckCIF reports for [9–13], structural errors were present and detected by the model. On the other hand, interpreting CheckCIF alerts requires specialized knowledge, and identifying errors is not always straightforward. It is also common to encounter AB-level structure alerts without real structural errors. In contrast, the corrections provided by our model are straightforward.

1.5 Details of the corrected errors in JCR Q1 journals

Detailed information about the corrected errors in JCR Q1 journals by the model is provided. In these cases, the structure analysis provided by the model aligns more closely with the real experimental observations and is intuitively more rational. A thorough review of the original literature confirmed the presence of these errors. We visualized both the original and corrected structures using Olex2 [14] and reported a detailed comparison of the crystallographic metrics. The details are as follows:

(a) The data was sourced from [9], with the entry numbered `7244616` in the COD database. The error was due to the misassignment of carbon and nitrogen atoms (Fig. 1, Table 1).

(b) The data was sourced from [15], with the entry numbered `7246090` in the COD database. The error was due to the misassignment of carbon and nitrogen atoms (Fig. 2, Table 2).

(c) The data was sourced from [16], with the entry numbered `4345059` in the COD database. The error was due to the incorrect hydrogen atoms on the oxygen and carbon atoms (Fig. 3, Table 3).

(d) The data was sourced from [10], with the entry numbered `4125905` in the COD database. The reason for the error is that one hydrogen atom was mistakenly omitted from the nitrogen atom (Fig. 4, Table 4).

(e) The data was sourced from [17], with the entry numbered `4123283` in the COD database. The reason for the error is that one hydrogen atom was mistakenly omitted from the oxygen atom (Fig. 5, Table 5).

(f) The data was sourced from [11], with the entry numbered `1540524` in the COD database. The reason for the error is that an extra hydrogen atom was incorrectly added to the carbon atom (Fig. 6, Table 6).

(g) The data was sourced from [12], with the entry numbered `1555760` in the COD database. The reason for the error is that an extra hydrogen atom was incorrectly added to the carbon atom (Fig. 7, Table 7).

(h) The data was sourced from [13], with the entry numbered `1555474` in the COD database. The reason for the error is that one hydrogen atom was mistakenly omitted from the nitrogen atom (Fig. 8, Table 8).

(i) The data was sourced from [18], with the entry numbered `1547704` in the COD database. The reason for the error is that hydrogen atoms were mistakenly omitted from the carbon atoms (Fig. 9, Table 9).

1.6 Practical applications

1.6.1 A web application

We have developed a preliminary web application to facilitate the practical use of the model. An example of practical usage for structure analysis is as follows: the user uploads the initial INS file and diffraction data file (HKL), as shown in Fig. 17.

The model can return the structure analysis results within seconds, as shown in Fig. 18. We render the structure analysis results, displaying the crystallographic calculations, and perform structure checks and diffraction data quality checks based on CheckCIF's A and B alerts. The analysis results are available for download in a zip file. This package includes a CIF, a CheckCIF report, XYZ and GJF files for easy viewing, SHELXL format files for seamless software integration, and PROB files that display model-predicted probabilities.

1.6.2 Analyzing complex structures

We show that the model can fully automate intricate analysis of complex structures with precise results in seconds. Fig. 19 and Fig. 19 illustrate the model's results on complex structures containing up to 370 heavy atoms within the test set. While the model completes this analysis almost instantaneously, the same task would be significantly more difficult and time-consuming for human experts.

1.6.3 Applying to newly discovered compounds and high-throughput crystallization methods

We further demonstrate the model's effectiveness by applying it to the structural analysis of newly discovered compounds. As illustrated in Fig. 21 and Table 13, Fig. 22 and Table 14, the results show that the model successfully automates precise analysis, with all crystallographic metrics falling within expected ranges and CheckCIF validation revealing no AB-level alerts. The structures have been submitted to the Cambridge

Crystallographic Data Centre (CCDC). Additionally, we highlight the model's compatibility with Encapsulated Nanodroplet Crystallization in Fig. 23 and Table 15, an advanced high-throughput crystallization technique.

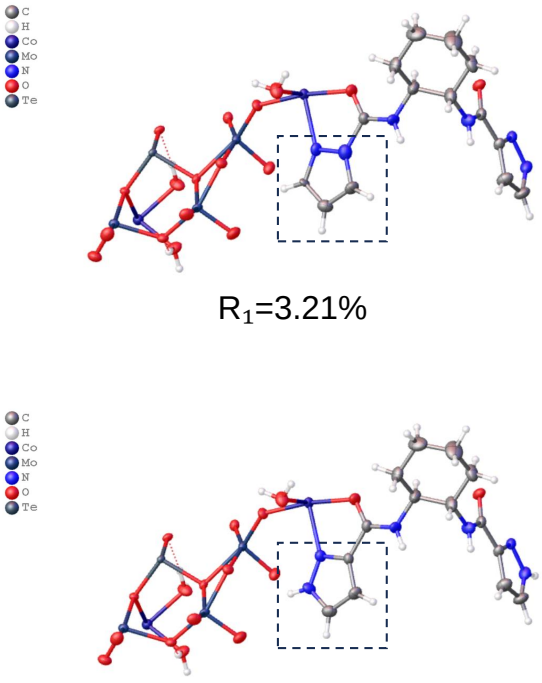

Fig. 1 The corrected and original structure of COD_7244616.

Table 1 Crystallographic results of the corrected and original structure of COD_7244616

	original	**corrected**
$\mathbf{R_1(obs)(\%)}$	3.21	3.02
$\mathbf{wR_2(obs)(\%)}$	6.57	5.89
$\mathbf{R_1(all)(\%)}$	4.63	4.45
$\mathbf{wR_2(all)(\%)}$	7.17	6.44
S	1.022	1.021
Peak, hole ($\mathbf{e^-\AA^{-3}}$)	1.12, -0.71	1.10, -0.69

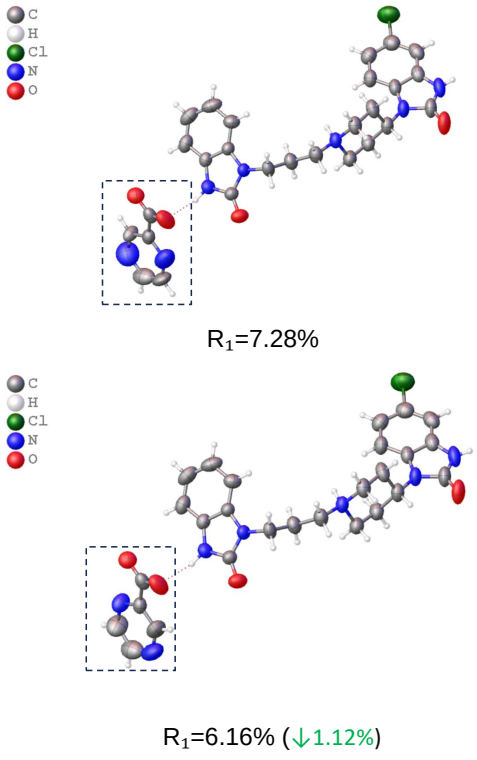

Fig. 2 The corrected and original structure of COD_7246096.

Table 2 Crystallographic results of the corrected and original structure of COD_7246096

	original	**corrected**
$\mathbf{R_1(obs)(\%)}$	7.28	6.16
$\mathbf{wR_2(obs)(\%)}$	19.11	15.42
$\mathbf{R_1(all)(\%)}$	15.23	13.99
$\mathbf{wR_2(all)(\%)}$	24.66	21.90
S	1.274	1.042
Peak, hole ($\mathbf{e^- \AA^{-3}}$)	0.39, -0.41	0.40, -0.33

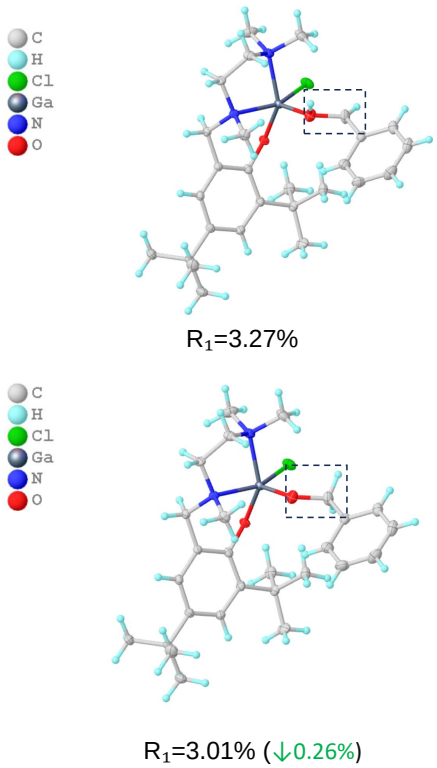

Fig. 3 The corrected and original structure of COD_4345059.

Table 3 Crystallographic results of the corrected and original structure of COD_4345059

	original	**corrected**
$\mathbf{R_1(obs)(\%)}$	3.27	3.01
$\mathbf{wR_2(obs)(\%)}$	11.40	7.94
$\mathbf{R_1(all)(\%)}$	4.12	3.87
$\mathbf{wR_2(all)(\%)}$	12.24	8.84
S	0.936	1.123
Peak, hole ($\mathbf{e^- \AA^{-3}}$)	0.69, -0.89	0.51, -0.37

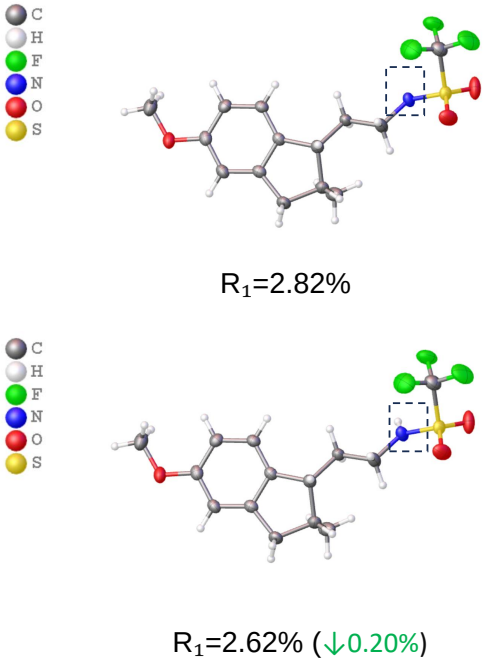

Fig. 4 The corrected and original structure of COD_4125905.

Table 4 Crystallographic results of the corrected and original structure of COD_4125905

	original	**corrected**
$\mathbf{R_1(obs)(\%)}$	2.82	2.62
$\mathbf{wR_2(obs)(\%)}$	7.87	6.90
$\mathbf{R_1(all)(\%)}$	2.90	2.69
$\mathbf{wR_2(all)(\%)}$	7.93	6.94
S	1.013	1.106
Peak, hole ($\mathbf{e^- \AA^{-3}}$)	0.46, -0.31	0.19, -0.32

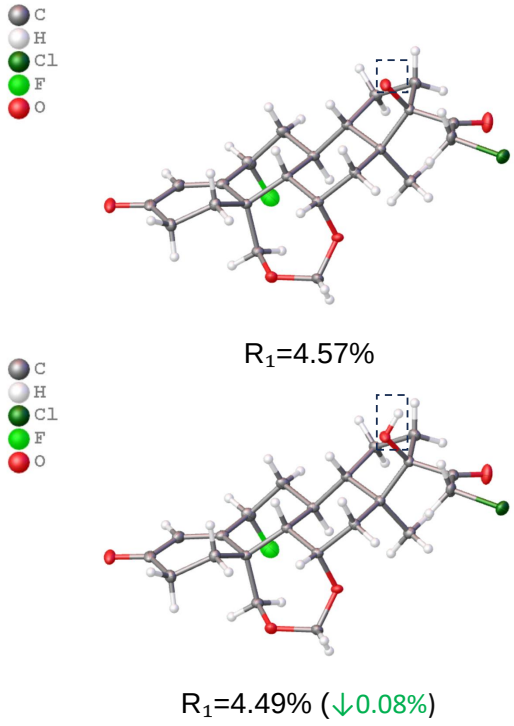

Fig. 5 The corrected and original structure of COD_4123283.

Table 5 Crystallographic results of the corrected and original structure of COD_4123283

	original	**corrected**
R_1(obs)(%)	4.57	4.49
wR_2(obs)(%)	12.58	12.21
R_1(all)(%)	4.74	4.66
wR_2(all)(%)	12.76	12.39
S	1.084	1.090
Peak, hole (e^-Å^{-3})	0.58, -0.49	0.58, -0.49

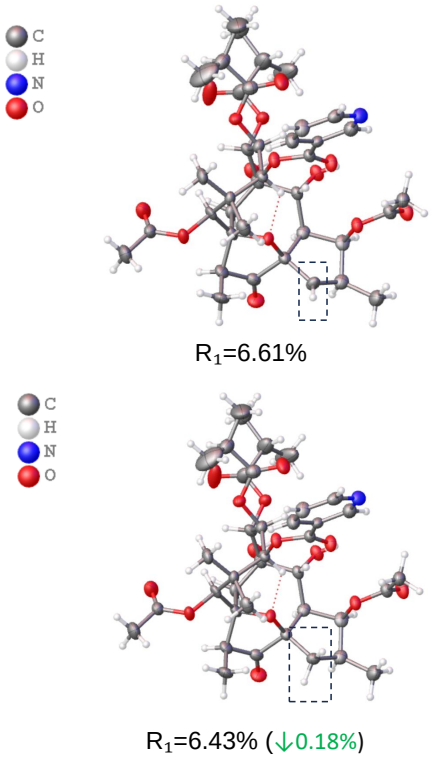

Fig. 6 The corrected and original structure of COD_1540524.

Table 6 Crystallographic results of the corrected and original structure of COD_1540524

	original	**corrected**
R_1(obs)(%)	6.61	6.43
wR_2(obs)(%)	12.44	11.49
R_1(all)(%)	15.24	15.09
wR_2(all)(%)	15.21	14.06
S	0.998	0.995
Peak, hole (e^- Å^{-3})	0.39, -0.27	0.34, -0.26

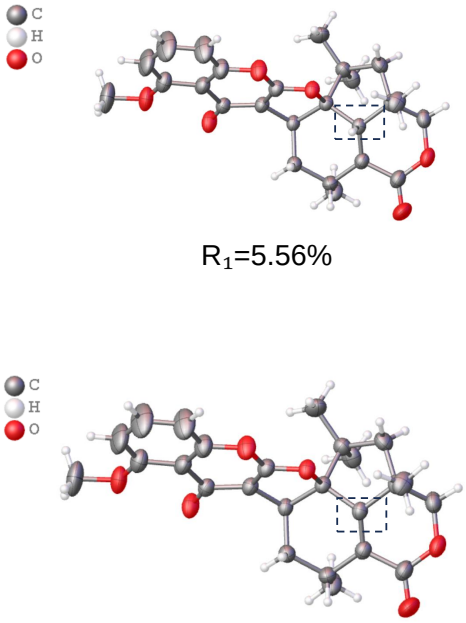

Fig. 7 The corrected and original structure of COD_1555760.

Table 7 Crystallographic results of the corrected and original structure of COD_1555760

	original	**corrected**
R_1(obs)(%)	5.56	5.15
wR_2(obs)(%)	14.45	13.07
R_1(all)(%)	5.88	5.47
wR_2(all)(%)	14.73	13.32
S	1.063	1.080
Peak, hole (e^- Å^{-3})	0.24, -0.67	0.23, -0.24

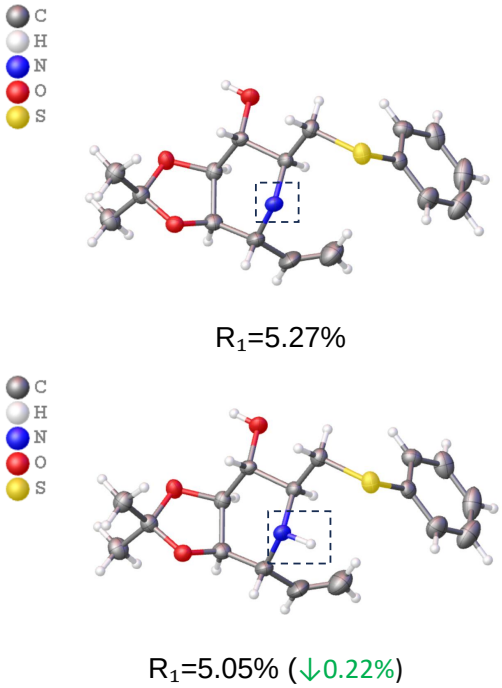

Fig. 8 The corrected and original structure of COD_1555474.

Table 8 Crystallographic results of the corrected and original structure of COD_1555474

	original	**corrected**
R_1(obs)(%)	5.27	5.05
wR_2(obs)(%)	12.69	11.15
R_1(all)(%)	7.03	6.82
wR_2(all)(%)	14.09	12.30
S	1.025	1.034
Peak, hole (e^- Å^{-3})	0.65, -0.26	0.61, -0.26

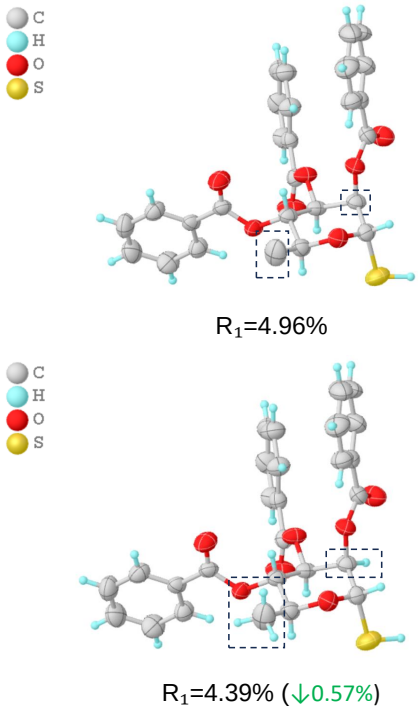

Fig. 9 The corrected and original structure of COD_1547704.

Table 9 Crystallographic results of the corrected and original structure of COD_1547704

	original	**corrected**
$\mathbf{R_1(obs)(\%)}$	4.96	4.39
$\mathbf{wR_2(obs)(\%)}$	12.35	8.86
$\mathbf{R_1(all)(\%)}$	7.76	7.23
$\mathbf{wR_2(all)(\%)}$	14.05	9.90
S	1.015	1.014
Peak, hole ($\mathbf{e^-\,\AA^{-3}}$)	0.34, -0.19	0.15, -0.20

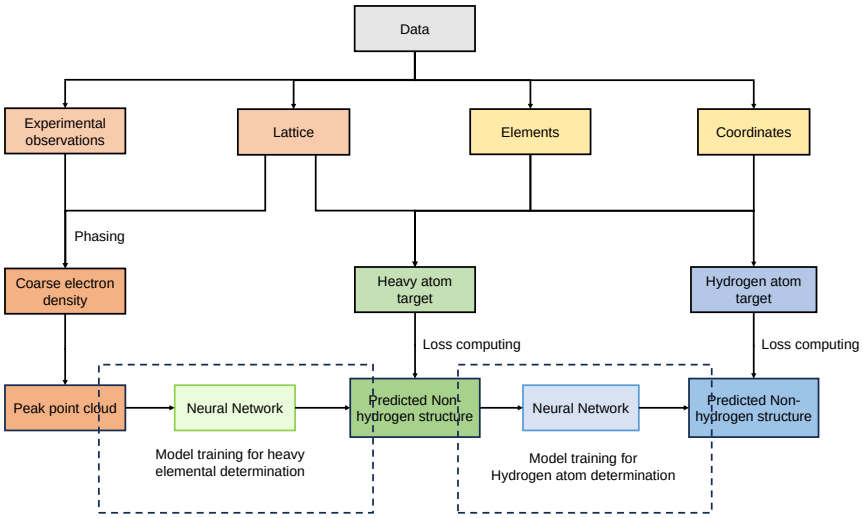

Fig. 10 The pipeline of data processing.

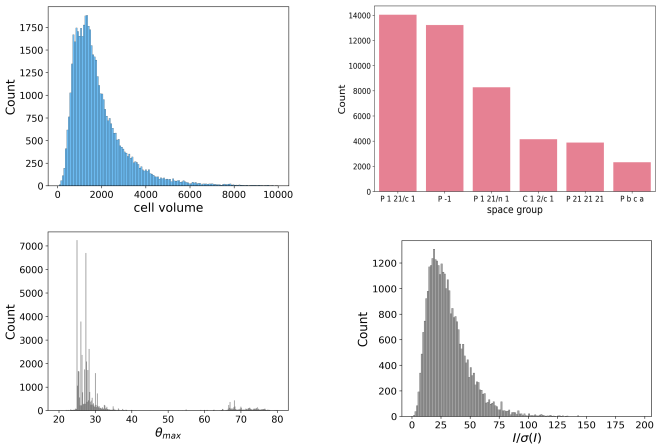

Fig. 11 The statistical information of the dataset. We provide a comprehensive overview of the dataset, detailing crystal information (including unit cell volume and space groups) and experimental conditions (such as maximum diffraction angle and signal-to-noise ratio).

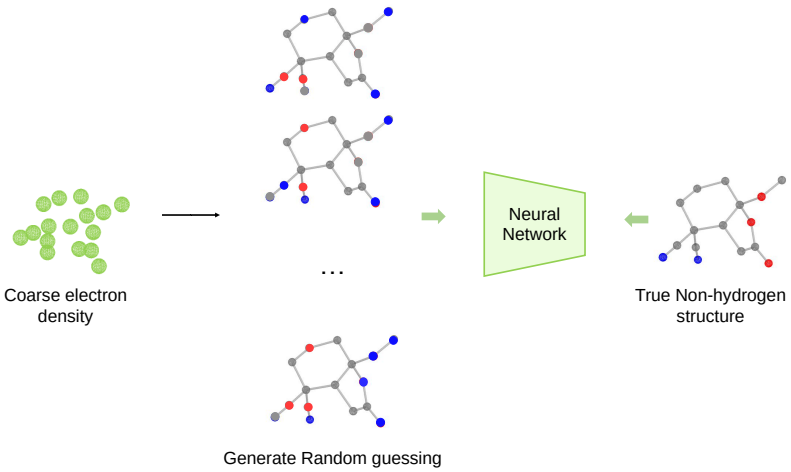

Fig. 12 The training scheme for non-hydrogen atom determination. We introduce diversity by randomly shuffling element assignments for peaks with similar charge densities. The model is trained to make correct predictions from a broad spectrum of these randomized inputs.

Table 10 Hyper-parameter settings of the neural network. We follow the default settings of the TorchMD-NET official implementation.

hidden_channels	256
num_layers	8
num_rbf	64
rbf_type	“expnorm”
trainable_rbf	False
activation	“silu”
attn_activation	“silu”
neighbor_embedding	True
num_heads	8
distance_influence	“both”
cutoff_lower	0.0
cutoff_upper	5.0
max_num_neighbors	32
layernorm_on_vec	None

Table 11 Comparison of model architecture

Model	**Accuracy (%)**	**Inference speed (it/s)**
SchNet [19]	89.86	132.22
DimeNet [20]	95.77	51.55
SphereNet [21]	95.91	28.96
CoMeNet [22]	93.57	96.19
TorchMD-NET [7]	**96.23**	95.13

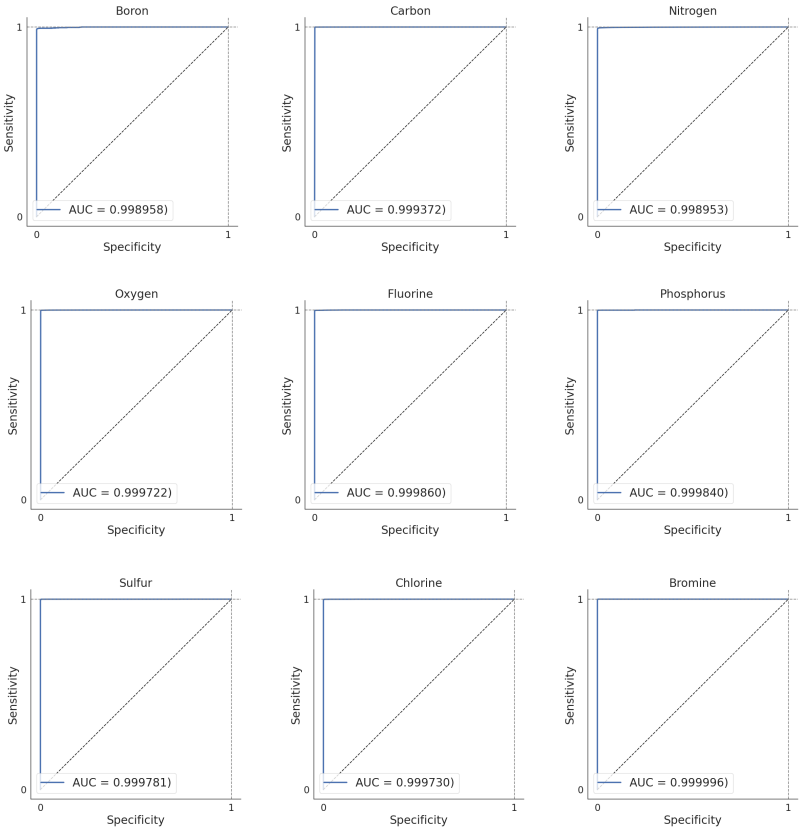

Fig. 13 Reliability of model's confidence in non-hydrogen elemental determination. We present the Receiver Operating Characteristic (ROC) curves for the nine most prevalent non-hydrogen elements, all of which exhibit Area Under the Curve (AUC) values surpassing 0.99.

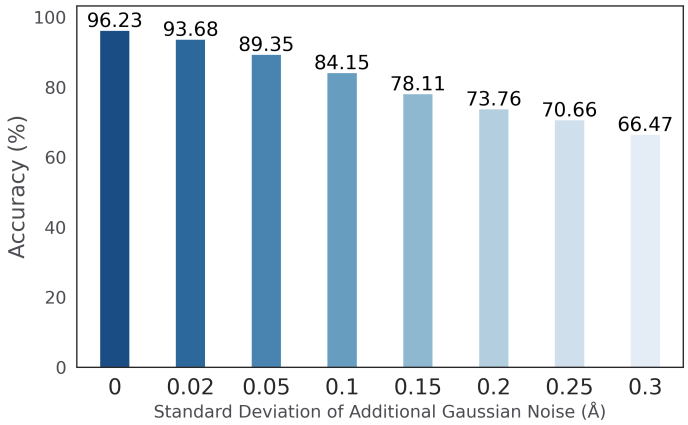

Fig. 14 The model's performance (unit-cell-level accuracy) under severe noise perturbations. We tested the robustness of the model in non-hydrogen elemental determination. We introduced additional Gaussian noise, independently applied across each coordinate dimension, with a standard deviation varying from 0 to 0.3 Å, to the experimental coarse electron density to simulate extreme cases of lower resolution. The model demonstrates only a slight decline in performance as noise levels increase. It maintains a unit-cell-level accuracy above 80% when the standard deviation reaches 0.1 Å—a magnitude significant enough to disrupt typical atomic geometric relationships.

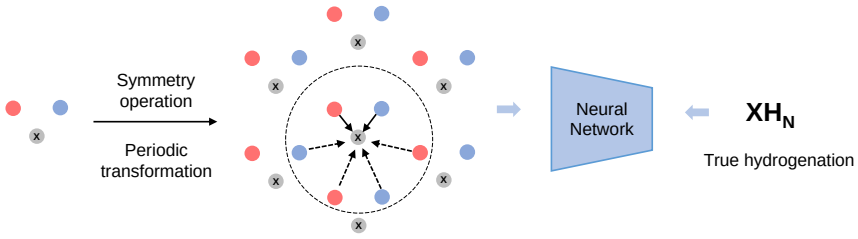

Fig. 15 The training scheme for hydrogen atom determination. We describe the chemical environment of hydrogen atoms by identifying equivalent atoms within a radius around each non-hydrogen atom, taking into account the crystal's symmetry and periodicity, ensuring that intermolecular interactions are properly incorporated.

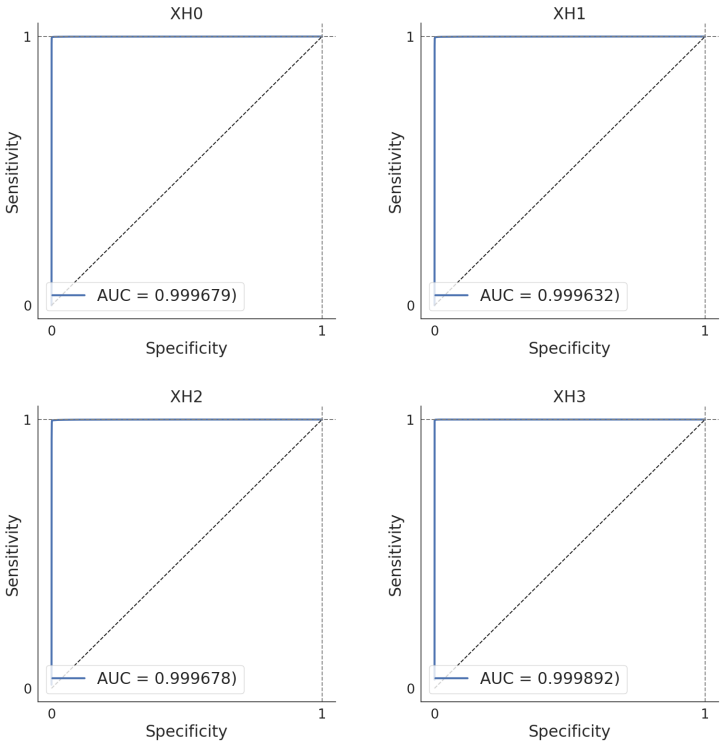

Fig. 16 Reliability of model's confidence in hydrogen atom determination. We present the Receiver Operating Characteristic (ROC) curves for hydrogen atom determination, with all Area Under the Curve (AUC) values exceeding 0.99.

Table 12 Additional results for hydrogen atoms in other environments. We also report the model's performance in predicting the number of hydrogen atoms in other chemical environments. The observed performance decline can be attributed to the significantly smaller dataset associated with these uncommon hydrogen atom environments.

	precision	**recall**	**F1-score**	**count**
H_2O	0.92	0.99	0.95	2781
NH_4^+	0.90	1.00	0.95	26
H_3O^+	0.75	0.24	0.36	25
BH_2	0.70	0.67	0.68	21
PH_1	0.91	0.53	0.67	19
PH_2	1.00	1.00	1.00	18
SH_1	0.93	0.72	0.81	18
BH_3	0.93	0.87	0.90	15
FH_1	1.00	0.33	0.50	12
SiH_1	1.00	0.90	0.95	10
RuH_1	0.60	0.30	0.40	10
FeH_1	0.33	0.43	0.38	7
IrH_1	0.80	0.80	0.80	5
RuH_2	0.40	1.00	0.57	4
BH_4	1.00	1.00	1.00	3
SiH_2	0.67	0.67	0.67	3
OsH_1	1.00	1.00	1.00	2

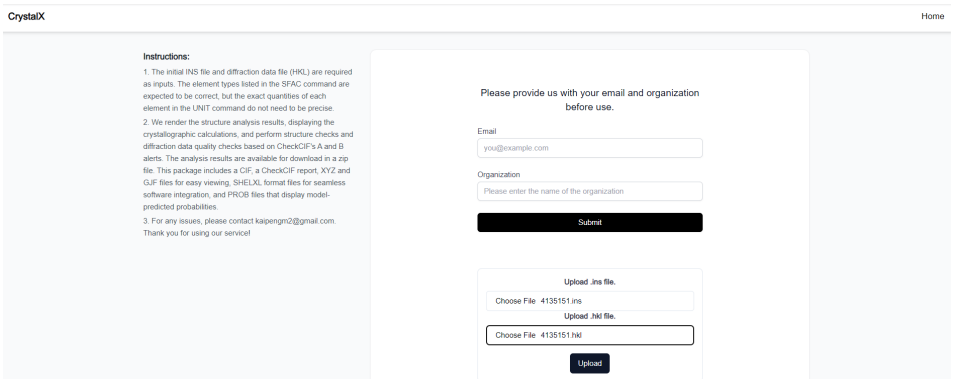

Fig. 17 The data upload page of the web application.

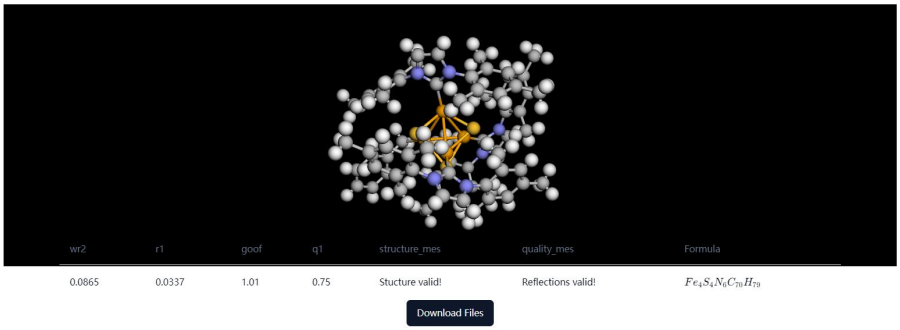

Fig. 18 The results display page of the web application.

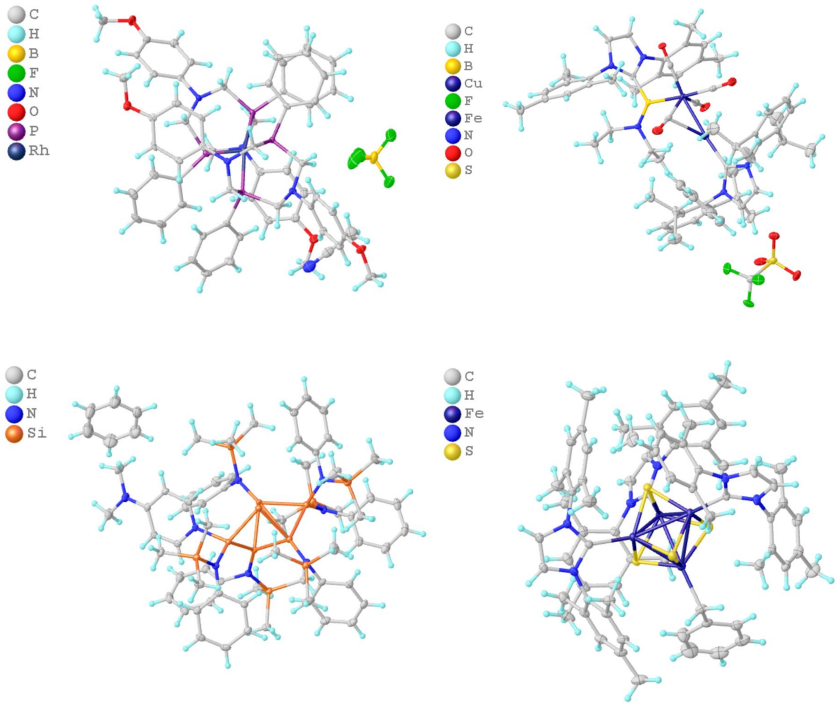

Fig. 19 Complex structure analysis (a): The model automated the precise structural analysis of complex structures in the test set (with up to 370 heavy atoms) within just a few seconds. The entries illustrated are `COD_4123903`, `COD_4132661`, `COD_7709435`, and `COD_4135151`.

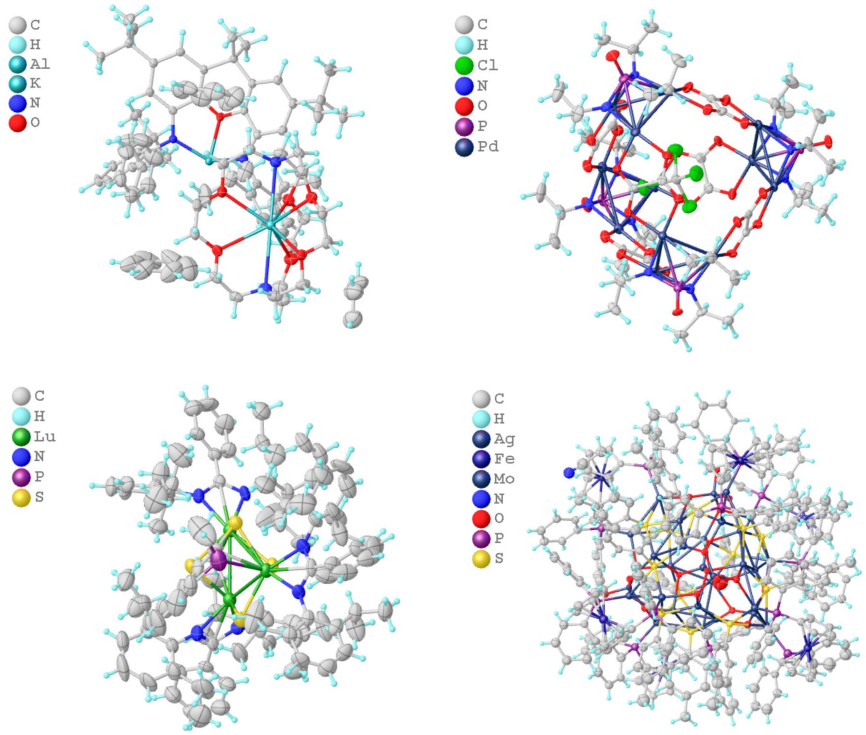

Fig. 20 Complex structure analysis (b): The model automated the precise structural analysis of complex structures in the test set (with up to 370 heavy atoms) within seconds. The entries illustrated are `COD_4127406`, `COD_4342696`, `COD_4126931`, and `COD_1545292`.

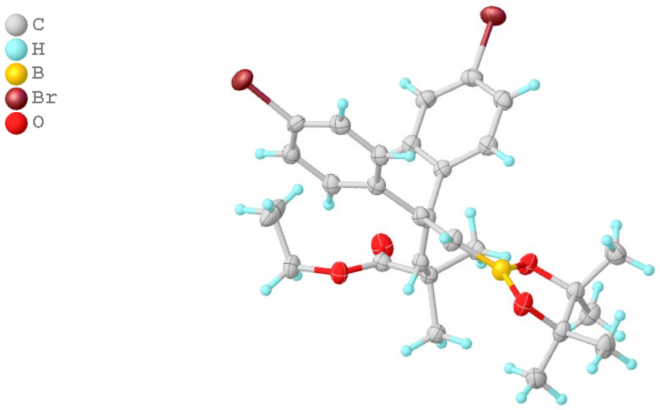

Fig. 21 The structure analysis of a newly discovered compound achieved by the model.

Table 13 Crystallographic results

$\mathbf{R_1(obs)(\%)}$	3.73
$\mathbf{wR_2(obs)(\%)}$	10.15
$\mathbf{R_1(all)(\%)}$	4.50
$\mathbf{wR_2(all)(\%)}$	10.59
S	1.054
Peak, hole ($\mathbf{e^- \AA^{-3}}$)	0.62, -0.69

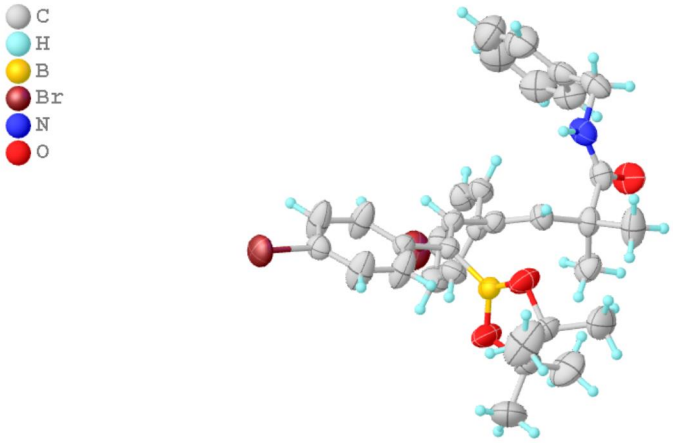

Fig. 22 The structure analysis of a newly discovered compound achieved by the model.

Table 14 Crystallographic results

$\mathbf{R_1(obs)(\%)}$	3.27
$\mathbf{wR_2(obs)(\%)}$	9.16
$\mathbf{R_1(all)(\%)}$	3.43
$\mathbf{wR_2(all)(\%)}$	9.29
S	1.090
Peak, hole ($\mathbf{e^- \AA^{-3}}$)	0.36, -0.46

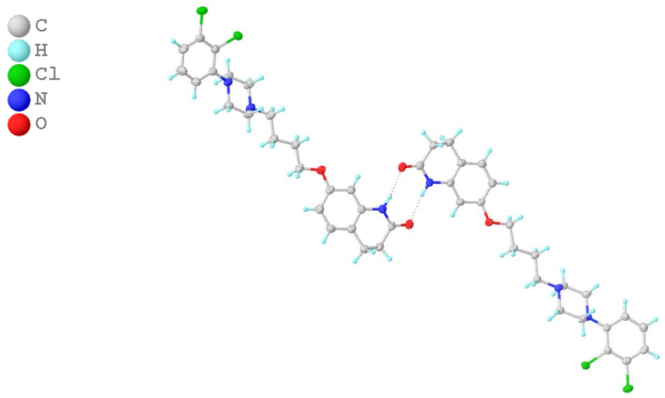

Fig. 23 The structure analysis of an Encapsulated Nanodroplet crystallized compound [23] achieved by the model.

Table 15 Crystallographic results

$\mathbf{R_1(obs)(\%)}$	3.87
$\mathbf{wR_2(obs)(\%)}$	8.92
$\mathbf{R_1(all)(\%)}$	6.04
$\mathbf{wR_2(all)(\%)}$	9.99
S	0.993
Peak, hole ($\mathbf{e^{-}\AA^{-3}}$)	0.26, -0.26

References

[1] Spek, A.L.: Structure validation in chemical crystallography. Acta Crystallographica Section D: Biological Crystallography **65**(2), 148–155 (2009)

[2] Sheldrick, G.M.: Shelxt–integrated space-group and crystal-structure determination. Acta Crystallographica Section A: Foundations and Advances **71**(1), 3–8 (2015)

[3] Sheldrick, G.M.: A short history of shelx. Acta Crystallographica Section A: Foundations of Crystallography **64**(1), 112–122 (2008)

[4] Oszlányi, G., Sütő, A.: Ab initio structure solution by charge flipping. Acta Crystallographica Section A: Foundations of Crystallography **60**(2), 134–141 (2004)

[5] Altomare, A., Burla, M.C., Camalli, M., Cascarano, G.L., Giacovazzo, C., Guagliardi, A., Moliterni, A.G., Polidori, G., Spagna, R.: Sir97: a new tool for crystal structure determination and refinement. Journal of applied crystallography **32**(1), 115–119 (1999)

[6] Burla, M.C., Caliandro, R., Carrozzini, B., Cascarano, G.L., Cuocci, C., Giacovazzo, C., Mallamo, M., Mazzone, A., Polidori, G.: Crystal structure determination and refinement *via SIR2014*. Journal of Applied Crystallography **48**(1), 306–309 (2015)

[7] Thölke, P., Fabritiis, G.D.: Equivariant transformers for neural network based molecular potentials. In: International Conference on Learning Representations (2022)

[8] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)

[9] Zhang, J.-Y., Zhang, Y.-C., Wang, X.-L., Chang, Z.-H., Zhang, Z., Lin, H.-Y., Cui, Z.-W.: Polyoxometalate-based cuii/coii complexes tuned using various metal–pyrazole loops: design, diverse architectures and catalytic activity toward the oxidation of thioether derivatives. CrystEngComm **24**, 3172–3178 (2022)

[10] Zhang, Z., Wang, J., Li, J., Yang, F., Liu, G., Tang, W., He, W., Fu, J.-J., Shen, Y.-H., Li, A., Zhang, W.-D.: Total synthesis and stereochemical assignment of delavatine a: Rh-catalyzed asymmetric hydrogenation of indene-type tetrasubstituted olefins and kinetic resolution through pd-catalyzed triflamide-directed c–h olefination. Journal of the American Chemical Society **139**(15), 5558–5567 (2017)

[11] Reis, M.A., André, V., Duarte, M.T., Lage, H., Ferreira, M.-J.U.: 12,17-cyclojatrophane and jatrophane constituents of euphorbia welwitschii. Journal of Natural Products **78**(11), 2684–2690 (2015)

[12] Qiao, C., Zhang, W., Han, J.-C., Li, C.-C.: Catalytic enantioselective total synthesis of hypocrolide a. Organic Letters **18**(19), 4932–4935 (2016)

[13] Moynihan, L., Chadda, R., McArdle, P., Murphy, P.V.: Allylic azide rearrangement in tandem with huisgen cycloaddition for stereoselective annulation: Synthesis of c-glycosyl iminosugars. Organic Letters **17**(24), 6226–6229 (2015)

[14] Dolomanov, O.V., Bourhis, L.J., Gildea, R.J., Howard, J.A., Puschmann, H.: Olex2: a complete structure solution, refinement and analysis program. Journal of applied crystallography **42**(2), 339–341 (2009)

[15] Rekha Rout, S., Kenguva, G., Giri, L., Dandela, R.: Novel salts of the antiemetic drug domperidone: synthesis, characterization and physicochemical property investigation. CrystEngComm **25**, 513–524 (2023)

[16] Kremer, A.B., Andrews, R.J., Milner, M.J., Zhang, X.R., Ebrahimi, T., Patrick, B.O., Diaconescu, P.L., Mehrkhodavandi, P.: A comparison of gallium and indium alkoxide complexes as catalysts for ring-opening polymerization of lactide. Inorganic Chemistry **56**(3), 1375–1385 (2017)

[17] Renata, H., Zhou, Q., Dünstl, G., Felding, J., Merchant, R.R., Yeh, C.-H., Baran, P.S.: Development of a concise synthesis of ouabagenin and hydroxylated corticosteroid analogues. Journal of the American Chemical Society **137**(3), 1330–1340 (2015)

[18] Doyle, L.M., O'Sullivan, S., Di Salvo, C., McKinney, M., McArdle, P., Murphy, P.V.: Stereoselective epimerizations of glycosyl thiols. Organic Letters **19**(21), 5802–5805 (2017)

[19] Schütt, K., Kindermans, P.-J., Sauceda Felix, H.E., Chmiela, S., Tkatchenko, A., Müller, K.-R.: Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. Advances in neural information processing systems **30** (2017)

[20] Gasteiger, J., Groß, J., Günnemann, S.: Directional message passing for molecular graphs. In: International Conference on Learning Representations (2020)

[21] Liu, Y., Wang, L., Liu, M., Lin, Y., Zhang, X., Oztekin, B., Ji, S.: Spherical message passing for 3d molecular graphs. In: International Conference on Learning Representations (2022)

[22] Wang, L., Liu, Y., Lin, Y., Liu, H., Ji, S.: ComENet: Towards complete and efficient message passing for 3d molecular graphs. In: Oh, A.H., Agarwal, A.,

Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)

[23] Tyler, A.R., Ragbirsingh, R., McMonagle, C.J., Waddell, P.G., Heaps, S.E., Steed, J.W., Thaw, P., Hall, M.J., Probert, M.R.: Encapsulated nanodroplet crystallization of organic-soluble small molecules. Chem **6**(7), 1755–1765 (2020)