Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

CrystalX: High-accuracy Crystal Structure Analysis Using Deep Learning

Yuqiang Li

liyuqiang@pjlab.org.cn

Shanghai Artificial Intelligence Laboratory https://orcid.org/0000-0001-6756-6154

Kaipeng Zheng

Shanghai Jiao Tong University

Weiran Huang

Shanghai Jiao Tong University https://orcid.org/0000-0003-1193-6157

Wanli Ouyang

Shanghai AI lab

Han-Sen Zhong

Shanghai Artificial Intelligence Laboratory

Physical Sciences - Article

Keywords:

Posted Date: October 7th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-5159628/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

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;

Abstract

Atomic structure analysis of crystalline materials is a paramount endeavor in both chemical and material sciences. This sophisticated technique necessitates not only a solid foundation in crystallography but also a profound comprehension of the intricacies of the accompanying software, posing a significant challenge in meeting the rigorous daily demands. For the first time, we confront this challenge head-on by harnessing the power of deep learning for ultra-precise structural analysis at the full-atom level. To validate the performance of the model, named CrystalX, we employed a vast dataset comprising over 50,000 X-ray diffraction measurements derived from authentic experiments, demonstrating performance that is commensurate with human experts and adept at deciphering intricate geometric patterns. Remarkably, CrystalX revealed that even peer-reviewed publications can harbor errors that are stealthy to human scrutiny, yet CrystalX adeptly rectifies them. This deep learning model revolutionizes the time frame for crystal structure analysis, slashing it down to seconds. It has already been successfully applied in the structure analysis of newly discovered compounds in the latest research without human intervention. Overall, CrystalX marks the beginning of a new era in automating routine structural analysis within self-driving laboratories.

Introduction

Crystal structure analysis, as a pivotal technical tool to gain valuable insights into the microscopic realm of matter, has significantly transformed humanity's perception of the world, helping scientists ascertain the structures of vital substances, such as nucleic acids [1], penicillin [2], vitamin B12 [3], and insulin [4]. These breakthroughs have fueled major scientific advancements. As technology progresses with automated synthesis [5–9] and high-throughput crystallization [10–12], the demand for routine structure analysis has surged. Yet, this task remains both labor-intensive and time-consuming, relying on the specialized skills of crystallographers.

Establishing a routine method to automate daily structure analysis with both precision and speed has long been a key objective to accelerate chemical discoveries. Thanks to continuous advancements in modern crystallographic software [13–17], phasing diffraction data into coarse electron density has become highly automated. Recent breakthroughs in deep learning-based phasing [18] have further improved this. However, the automation of routine structure analysis continues to be impeded by the next critical step: the accurate interpretation of coarse electron density. This step remains heavily dependent on human experts who painstakingly work to populate the electron density map with the appropriate elements and hydrogen atoms. These experts require a deep understanding of crystallography and extensive experience with specialized software [16, 19–23]. The process demands a meticulous comparison of the proposed solution with experimental diffraction data, iteratively refining the fit until the analysis is complete. The complexity of this task escalates dramatically with the increasing number of atoms.

We show that artificial intelligence (AI) holds the potential to overcome this challenge. Through deep learning, AI has become a formidable tool for automating routine tasks and advancing scientific discovery [24–28]. Particularly, pioneering studies in areas such as computer-aided diagnosis [24], material innovation [25], and molecular property prediction [26] have underscored AI's capability in extracting hidden information from complex patterns that are challenging to discern manually. In crystallography, deep learning has been applied to analyze crystal symmetry in powder diffraction [29–31] and electron diffraction [32]. Recent research also highlights its promise in phasing low-quality X-ray diffraction data [18]. Nonetheless, the application of AI to fully automate routine structure analysis remains unexplored.

This work introduces the first deep learning model, CrystalX, to fully automate routine structure analysis with ultra-precision, bringing the time required for the entire process down to seconds. CrystalX's effectiveness is validated using a dataset of over 50,000 real-world experimental X-ray diffraction entries from the Crystallography Open Database (COD) [33], encompassing a diverse range of organic, organometallic, and inorganic compounds across 83 elements and 86 space groups. CrystalX accurately determines non-hydrogen atoms and hydrogen atoms with an accuracy of 99.80% and 99.63%, respectively, rivaling the performance of human experts in crystallographic comparison. Additionally, we investigate CrystalX's intrinsic behavior, confirming the high interpretability of the learned geometric relationships. CrystalX has already demonstrated its practical application. For example, 9 errors have been identified and corrected from 1,559 instances in Journal Citation Reports (JCR) [34] Q1 journals,

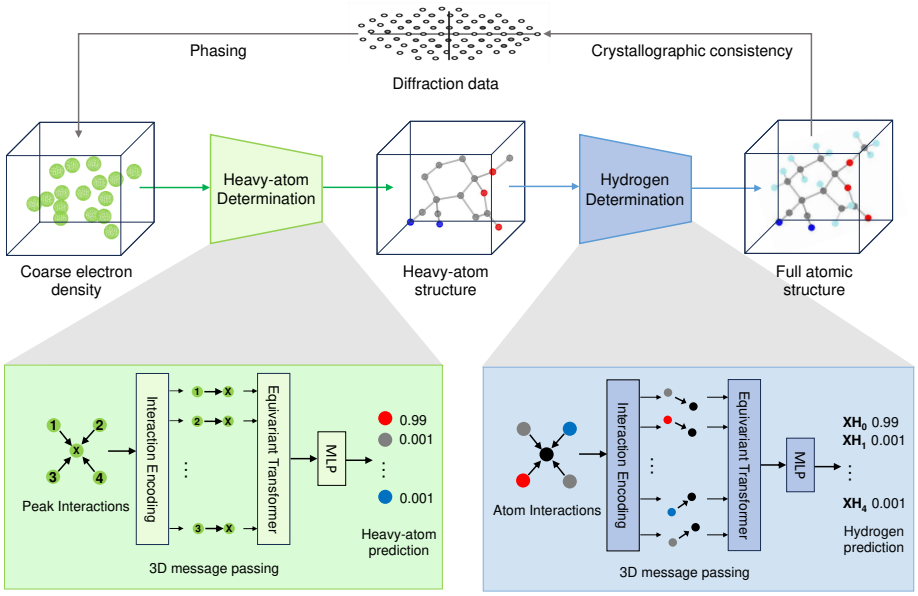


Fig. 1 The CrystalX neural network approach to crystal structure analysis. The inputs consist of coarse electron density peaks, which are processed using an Equivariant Transformer to extract geometric interaction patterns for heavy-atom (non-hydrogen atoms) determination. Following this, a separate Equivariant Transformer is developed to capture interactions for hydrogen determination. The dataset is derived from automatically phased coarse electron density in experimental diffraction measurements, with expert human interpretations serving as annotations.

revealing that even peer-reviewed journals contain errors that escape human detection. It has also been successfully utilized in the structure analysis of newly discovered compounds in cutting-edge research, entirely without human intervention.

Results

The neural network approach

Crystal structure analysis typically begins by phasing experimental X-ray diffraction data into an initial, coarse electron density map. Our approach is designed to predict the complete atomic structure from the coarse electron density that is automatically produced by routine software [13–16]. To achieve this objective, we split it into a pipeline that separately determines the heavy atoms (i.e., non-hydrogen atoms) and the hydrogen atoms, as depicted in Fig. 1.

The coarse electron density can be represented as a point cloud, where each peak corresponds to a specific location with a charge density value. Distinguishing elements with similar atomic numbers, such as the ubiquitous boron, carbon, nitrogen, oxygen, and fluorine, based on charge density can be challenging. This difficulty is particularly pronounced in real-world experimental measurements where data quality may

not be optimal. The foundation for elemental determination hinges on precisely capturing atomic geometric interaction patterns, such as distances, angles, and dihedral angles, derived from electron density peaks. With the advancements in geometric deep learning [35–39], these complex patterns can now be learned by deep-learning models in an end-to-end manner, removing the need for rule-based design or intricate handcrafted features. We harness the power of an advanced Equivariant Transformer model, TorchMD-NET [39], to decode geometric interaction patterns from electron density peaks, enabling accurate identification of heavy elements. We also show that various geometric deep learning methods can effectively address this problem, with TorchMD-NET emerging as a particularly competitive option due to its superior balance of accuracy and efficiency, making it a highly practical solution.

Hydrogen atoms are often indistinct in electron density maps obtained from traditional X-ray diffraction [40, 41]. The critical aspect of determining hydrogen atoms involves accurately identifying the exact number of hydrogen atoms to be added to each non-hydrogen atom. Once this number is established, the hydrogen atoms can be systematically positioned using theoretical models, taking into account any residual electron density. The key to determining hydrogen atoms lies in precisely identifying the number of hydrogen atoms that should be added to each non-hydrogen atom. Once this number is determined, the placement of hydrogen atoms can be automatically carried out based on theoretical models and any possible residual electron density. Typically, hydrogen atoms are located on the outer edges of a molecule, and their potential to form hydrogen bonds with neighboring molecules serves as an important criterion in determining hydrogen atoms. As a result, understanding the interaction patterns between molecules becomes essential for fully describing the chemical environment. We build another TorchMD-Net model for hydrogen atom determination, incorporating joint modeling of both intra-molecular and inter-molecular interactions. Specifically, we identify equivalent atoms within a 3.2 Å radius of each non-hydrogen atom, a distance that aligns with the typical maximum bond length of a hydrogen bond, taking into account the crystal's symmetry and periodicity. These equivalent atoms serve as auxiliary atoms in the model, introducing necessary intermolecular interaction patterns to predict the number of hydrogen atoms added to each non-hydrogen atom. We show that incorporating both intra- and intermolecular interactions yields a significant performance improvement of over 10% compared to using only intramolecular interactions.

CrystalX enables precise and rapid full atomic structure analysis on large-scale real experimental data.

We construct a dataset with 51,334 X-ray diffraction data gathered from real-world experimental measurements through the Crystallography Open Database (COD). This includes the diffraction data of organic compounds, organometallics, and inorganic substances, encompassing a total of 83 elements that nearly span the entire periodic table, and covering 86 space groups. The dataset is split into an approximately 8:2 ratio, with 41,334 instances allocated for training and validation, and 10,000 instances reserved for testing. We perform 10-fold cross-validation on the training and validation sets. The model demonstrating the best performance during validation is used

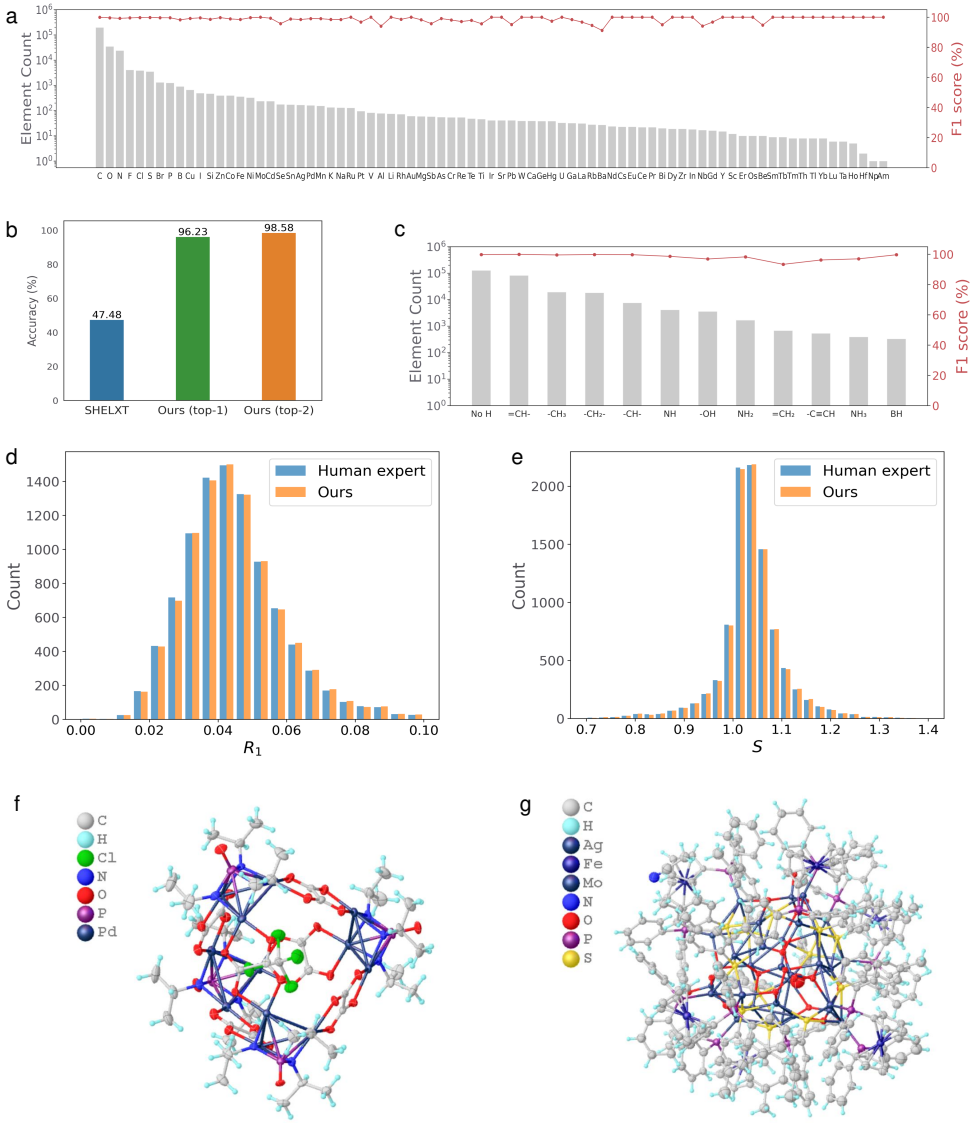


Fig. 2 Performance overview on large-scale real experimental data. **(a)** Element-level F1 score for non-hydrogen elements. **(b)** Unit-cell-level accuracy for non-hydrogen elements. **(c)** F1 score for prevalent hydrogen atoms. **(d, e)** Comparison of crystallographic metrics (i.e., R_1 and S) between the model and human experts. **(f, g)** Complex structure analysis (with up to 370 heavy atoms) automated by the model.

for evaluation on the reserved test set. Across the entire test set, CrystalX achieves a 99.80% accuracy in determining non-hydrogen atoms (Fig. 2a) and a 99.63% accuracy

for hydrogen atoms (Fig. 2c). Specifically, CrystalX generally performs well across a total of 83 elements, precisely distinguishing key elements with similar charges in organic compounds and organometallics, such as carbon (precision: 99.91%, recall: 99.95%), nitrogen (precision: 99.45%, recall: 99.03%), oxygen (precision: 99.50%, recall: 99.73%), and fluorine (precision: 99.88%, recall: 99.29%). At the unit cell level, where all atoms within the unit cell are accurately identified, the model achieves an accuracy of 96.23% for non-hydrogen atoms and 94.26% for hydrogen atoms. In contrast, the widely-used crystallography tool SHELXT achieves only 47.48% accuracy for non-hydrogen atoms. CrystalX therefore represents a 48.75% improvement (Fig. 2b). Furthermore, we show that CrystalX's predicted probabilities are highly reliable in the Supplementary Information, with AUC values exceeding 0.99 for each of the nine most common non-hydrogen elements. Therefore, additional corrections can be effortlessly made based on the predicted probabilities. For structures with a single recognition error, adjusting the prediction to the model's second-highest probability can achieve an accuracy of 98.58% for non-hydrogen atom determination and 97.48% for hydrogen atoms. Additionally, we evaluate the structure analysis produced by CrystalX against crystallographic metrics: R_1, and S (Fig. 2de), which reflects the consistency between the structure analysis and the experimental observations. CrystalX consistently produces results that are highly comparable to those achieved by human experts. Notably, CrystalX can process the entire test set in just a few minutes, far surpassing human experts. We have also demonstrated CrystalX's capabilities in analyzing complex structures, which poses exponential challenges for human experts. Yet, CrystalX can fully automate the complex structure analysis (with up to 370 heavy atoms) precisely in seconds (Fig. 2fg). Furthermore, beyond evaluating CrystalX on the current dataset, we also demonstrate its ability to generalize by successfully analyzing the structures of newly discovered compounds in the Supplementary Information, where CrystalX's structure analysis can pass standard validation [42] directly without any human intervention.

Visualization and interpretability of the model's underlying behavior

How does the model capture latent patterns for analysis? To gain a deeper understanding of the learned analysis mechanisms, we employ Attention Rollout [43] to visualize the attention maps of the Equivariant Transformer during elemental determination from coarse electron density. Fig. 3a-d illustrates the geometric interactions the model has learned from the coarse electron density of various organic and organometallic compounds. The colors and shading in the visualizations highlight class activations, reflecting both their relevance and importance. The model clearly captures key geometric features critical for recognizing diverse elements. Additionally, as shown in the t-SNE projection of the learned embeddings in Fig. 3ef, different elements and hydrogen types exhibit clear and distinct clustering.

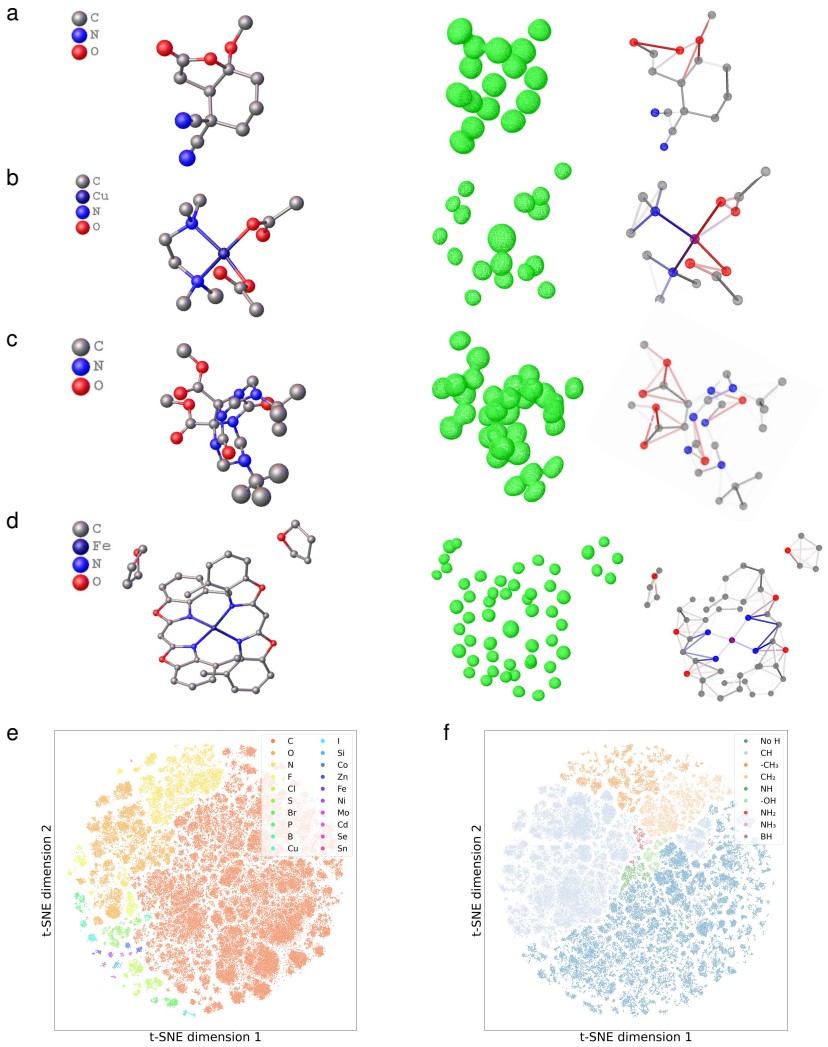


Fig. 3 Visualization and interpretability of the model's underlying behavior. (**a-d**) Visualization of attention maps from Equivalent Transformers using Attention Rollout [43] with class activation demonstrates precise identification of elements like carbon, nitrogen, oxygen, and metals (b,d). This identification is achieved solely by analyzing the geometric patterns of coarse electron density, without the need for any prior information. The interaction intensity is shaded according to class activation levels. (**e,f**) Two-dimensional t-SNE [44] projections of the model's learned representations for non-hydrogen atom identification (e) and hydrogen atom identification (f).

CrystalX corrects errors in JCR Q1 journals

The interpretations recorded by human experts in Crystallographic Information Files (CIF) are generally regarded as the standard for analyzing the corresponding diffraction data. However, could these manual interpretations potentially contain errors?

When evaluating the model on the test set, comparing the crystallographic metrics achieved by both human experts and our models under identical conditions (Fig. 4a), we observed that even for publications in JCR Q1 journals, the structure analysis by the model achieved significantly better performance than those of human experts. Upon verification, it became evident that these discrepancies were due to errors made by the human experts, while the model's analysis was correct. Three typical errors are illustrated in Fig. 4: Incorrect identification of non-hydrogen atoms (Fig. 4b, `COD_7244616`), Incorrect placement of hydrogen atoms (Fig. 4c, `COD_4345059`), Missing hydrogen atoms (Fig. 4d, `COD_4125905`). These subtle errors are difficult to detect, even passing the rigorous scrutiny of JCR Q1 journals.

Out of the 1,559 entries from JCR Q1 journals in the entire test dataset, 10 data points were automatically filtered. The models outperformed human experts in these cases, indicating a high likelihood of interpretation errors by the latter. Upon manual verification, the 9 interpretation errors by human experts were confirmed. The details are shown in the Supplementary Information. This demonstrates that our method has reached the level of professional crystallographers and has the capability to verify the analysis results of human experts.

Discussions

The definitive objective of crystal structure analysis is to precisely determine the full atomic structure. Recent advances in modern software [13–17] have significantly automated the phasing step in routine structure analysis. The current bottleneck impeding automation lies in interpreting the obtained coarse electron density. Human experts, relying on extensive knowledge and software skills, must manually iterate the process of filling electron density maps. As the number of atoms increases, the time and labor required grow exponentially, making the task increasingly resource-intensive. This work represents the first data-driven approach using deep learning to achieve routine structure analysis at the crystallographer level, enabling thorough structure analysis in seconds. CrystalX is entirely trained and evaluated on data obtained from real experimental measurements, encompassing a wide range of elements and substances. Its performance not only rivals that of human experts but also corrects errors identified in high-impact JCR Q1 publications, demonstrating its effectiveness in resolving the structures of newly discovered compounds. These strengths make CrystalX ready for practical applications, poised to become an integral component in self-driving laboratories [45].

X-ray diffraction remains the most widely used technique for routine structure analysis. However, recent advancements in novel characterization methods, such as three-dimensional electron diffraction [46–48] and serial femtosecond X-ray crystallography [49, 50], have attracted significant attention. These techniques are particularly effective for resolving crystal structures with challenging characteristics, including low symmetry, nano- and micro-sized crystals, and radiation sensitivity. Despite the differences in these characterization techniques, the structure analysis still adheres closely

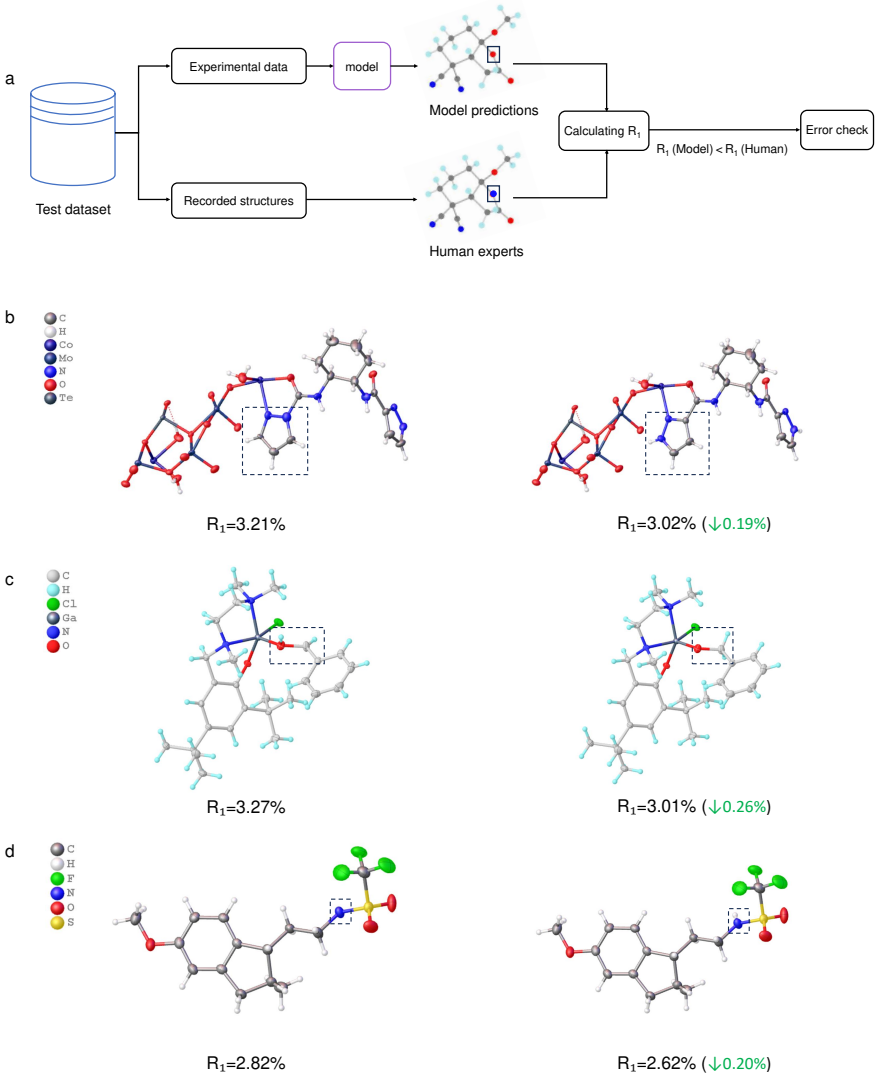


Fig. 4 CrystalX corrects errors in JCR Q1 journals. a, Error corrections through crystallographic comparison of the structure analysis results between the model and human experts. **b-d**, The model corrects three types of errors found in the published literature: (b) misidentification of atoms with similar charges, (c) incorrect placement of hydrogen atoms, and (d) missing hydrogen atoms. These errors are often difficult to detect, passing the rigorous reviews of JCR Q1 journals. The model's analysis offers both improved structural rationality and superior crystallographic metrics.

to the standard procedures used in X-ray diffraction [46–50]. This implies that CrystalX holds the potential to be extended to these cutting-edge techniques, enabling crystal structure analysis that surpasses traditional X-ray diffraction.

Our approach sets itself apart from previous methods that rely heavily on simulated data for evaluation by being entirely based on real experimental measurements. This foundation ensures its seamless applicability to real-world scenarios. CrystalX exhibits ultra-precision and reliability in automating structure analysis of 10,000 real-world experimental data, including complex structures with up to 370 heavy atoms—challenges that are exponentially difficult for human experts. Its effectiveness is further highlighted by its ability to correct errors even in JCR Q1 journals, which have undergone rigorous expert review. Moreover, CrystalX has showcased successful in automating the structure analysis of newly discovered compounds and is fully compatible with high-throughput crystallization technology. An additional advantage of CrystalX is its ability to bypass the need for precise prior knowledge of each element's content, which human experts typically require and often obtain through additional experimental measurements.

Conclusions

In conclusion, this work introduces the first deep learning model, CrystalX, to fully automate routine structure analysis. CrystalX is trained and evaluated entirely on real experimental data, covering a wide range of elements and substances. It not only demonstrates crystallographer-level performance but also offers interpretable learned representations. Remarkably, it even corrects errors that have previously passed scrutiny from human reviewers in leading JCR Q1 journals, effectively facilitating the automated structural analysis of newly discovered compounds. These features make CrystalX well-suited to meet the rigorous demands of daily crystal structure analysis and position it for integration into self-driving laboratories, driving forward chemical discovery.

References

[1] Watson, J.D., Crick, F.H.: Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature **171**(4356), 737–738 (1953)

[2] Crowfoot, D., Bunn, C.W., Rogers-Low, B.W., Turner-Jones, A.: X-ray crystallographic investigation of the structure of penicillin. Chemistry of penicillin, 310–367 (1949)

[3] DG, H., Pickworth, J., JH, R., KN, T., RJ, P., JG, W.: The crystal structure of the hexacarboxylic acid derived from b12 and the molecular structure of the vitamin. Nature **176**(4477), 325–328 (1955)

[4] Adams, M.J., Blundell, T.L., Dodson, E.J., Dodson, G.G., Vijayan, M., Baker, E.N., Harding, M.M., Hodgkin, D.C., Rimmer, B., Sheat, S.: Structure of rhombohedral 2 zinc insulin crystals. Nature **224**(5218), 491–495 (1969)

[5] Li, J., Ballmer, S.G., Gillis, E.P., Fujii, S., Schmidt, M.J., Palazzolo, A.M., Lehmann, J.W., Morehouse, G.F., Burke, M.D.: Synthesis of many different types

of organic small molecules using one automated process. Science **347**(6227), 1221–1226 (2015)

[6] Trobe, M., Burke, M.D.: The molecular industrial revolution: automated synthesis of small molecules. Angewandte Chemie International Edition **57**(16), 4192–4214 (2018)

[7] Greenaway, R., Santolini, V., Bennison, M., Alston, B., Pugh, C., Little, M., Miklitz, M., Eden-Rump, E., Clowes, R., Shakil, A., *et al.*: High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. Nature communications **9**(1), 2849 (2018)

[8] Blair, D.J., Chitti, S., Trobe, M., Kostyra, D.M., Haley, H.M., Hansen, R.L., Ballmer, S.G., Woods, T.J., Wang, W., Mubayi, V., *et al.*: Automated iterative c sp 3–c bond formation. Nature **604**(7904), 92–97 (2022)

[9] Wang, W., Angello, N.H., Blair, D.J., Tyrikos-Ergas, T., Krueger, W.H., Medine, K.N., LaPorte, A.J., Berger, J.M., Burke, M.D.: Rapid automated iterative small-molecule synthesis. Nature Synthesis, 1–8 (2024)

[10] Metherall, J.P., Carroll, R.C., Coles, S.J., Hall, M.J., Probert, M.R.: Advanced crystallisation methods for small organic molecules. Chem. Soc. Rev. **52**, 1995–2010 (2023)

[11] Bolla, G., Sarma, B., Nangia, A.K.: Crystal engineering of pharmaceutical cocrystals in the discovery and development of improved drugs. Chemical reviews **122**(13), 11514–11603 (2022)

[12] 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)

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

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

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

[16] 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)

[17] 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)

[18] Larsen, A.S., Rekis, T., Madsen, A.Ø.: Phai: A deep-learning approach to solve the crystallographic phase problem. Science **385**(6708), 522–528 (2024)

[19] Sheldrick, G.M.: Crystal structure refinement with shelxl. Acta Crystallographica Section C: Structural Chemistry **71**(1), 3–8 (2015)

[20] Betteridge, P.W., Carruthers, J.R., Cooper, R.I., Prout, K., Watkin, D.J.: Crystals version 12: software for guided crystal structure analysis. Journal of Applied Crystallography **36**(6), 1487–1487 (2003)

[21] 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)

[22] Hübschle, C.B., Sheldrick, G.M., Dittrich, B.: Shelxle: a qt graphical user interface for shelxl. Journal of applied crystallography **44**(6), 1281–1284 (2011)

[23] Farrugia, L.J.: Wingx suite for small-molecule single-crystal crystallography. journal of Applied Crystallography **32**(4), 837–838 (1999)

[24] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. nature **542**(7639), 115–118 (2017)

[25] Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., Cubuk, E.D.: Scaling deep learning for materials discovery. Nature **624**(7990), 80–85 (2023)

[26] Lee, B.K., Mayhew, E.J., Sanchez-Lengeling, B., Wei, J.N., Qian, W.W., Little, K.A., Andres, M., Nguyen, B.B., Moloy, T., Yasonik, J., *et al.*: A principal odor map unifies diverse tasks in olfactory perception. Science **381**(6661), 999–1006 (2023)

[27] Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., *et al.*: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)

[28] Tubiana, J., Schneidman-Duhovny, D., Wolfson, H.J.: Scannet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nature Methods **19**(6), 730–739 (2022)

[29] Park, W.B., Chung, J., Jung, J., Sohn, K., Singh, S.P., Pyo, M., Shin, N., Sohn, K.-S.: Classification of crystal structure using a convolutional neural network.

IUCrJ **4**(4), 486–494 (2017)

[30] Ziletti, A., Kumar, D., Scheffler, M., Ghiringhelli, L.M.: Insightful classification of crystal structures using deep learning. Nature communications **9**(1), 2775 (2018)

[31] Oviedo, F., Ren, Z., Sun, S., Settens, C., Liu, Z., Hartono, N.T.P., Ramasamy, S., DeCost, B.L., Tian, S.I., Romano, G., *et al.*: Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. npj Computational Materials **5**(1), 60 (2019)

[32] Kaufmann, K., Zhu, C., Rosengarten, A.S., Maryanovsky, D., Harrington, T.J., Marin, E., Vecchio, K.S.: Crystal symmetry determination in electron diffraction using machine learning. Science **367**(6477), 564–568 (2020)

[33] Gražulis, S., Chateigner, D., Downs, R.T., Yokochi, A., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., Le Bail, A.: Crystallography open database–an open-access collection of crystal structures. Journal of applied crystallography **42**(4), 726–729 (2009)

[34] https://clarivate.com/products/scientific-and-academic-research/research-analytics-evaluation-and-management-solutions/journal-citation-reports/

[35] 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)

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

[37] 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)

[38] 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)

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

[40] Palatinus, L., Brázda, P., Boullay, P., Perez, O., Klementová, M., Petit, S., Eigner, V., Zaarour, M., Mintova, S.: Hydrogen positions in single nanocrystals revealed by electron diffraction. Science **355**(6321), 166–169 (2017)

[41] Clabbers, M.T., Gruene, T., Genderen, E., Abrahams, J.P.: Reducing dynamical electron scattering reveals hydrogen atoms. Acta Crystallographica Section A: Foundations and Advances **75**(1), 82–93 (2019)

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

[43] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190–4197. Association for Computational Linguistics, Online (2020)

[44] Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[45] Abolhasani, M., Kumacheva, E.: The rise of self-driving labs in chemical and materials sciences. Nature Synthesis **2**(6), 483–492 (2023)

[46] Klar, P.B., Krysiak, Y., Xu, H., Steciuk, G., Cho, J., Zou, X., Palatinus, L.: Accurate structure models and absolute configuration determination using dynamical effects in continuous-rotation 3d electron diffraction data. Nature Chemistry **15**(6), 848–855 (2023)

[47] Kapaca, E., Jiang, J., Cho, J., Jordá, J.L., Diaz-Cabanas, M.J., Zou, X., Corma, A., Willhammar, T.: Synthesis and structure of a 22× 12× 12 extra-large pore zeolite itq-56 determined by 3d electron diffraction. Journal of the American Chemical Society **143**(23), 8713–8719 (2021)

[48] Saha, A., Nia, S.S., Rodríguez, J.A.: Electron diffraction of 3d molecular crystals. Chemical Reviews **122**(17), 13883–13914 (2022)

[49] Schriber, E.A., Paley, D.W., Bolotovsky, R., Rosenberg, D.J., Sierra, R.G., Aquila, A., Mendez, D., Poitevin, F., Blaschke, J.P., Bhowmick, A., *et al.*: Chemical crystallography by serial femtosecond x-ray diffraction. Nature **601**(7893), 360–365 (2022)

[50] Takaba, K., Maki-Yonekura, S., Inoue, I., Tono, K., Hamaguchi, T., Kawakami, K., Naitow, H., Ishikawa, T., Yabashi, M., Yonekura, K.: Structural resolution of a small organic molecule by serial x-ray free-electron laser and electron crystallography. Nature chemistry **15**(4), 491–497 (2023)

[51] https://doi.org/10.5281/zenodo.13820303

Data and code availability:

All data required to validate the conclusions of this paper are provided in the manuscript, the supplementary materials, and a data archive containing all relevant

data, code, and model parameters, accessible at [51]. A web application is available at `https://crystalx.intern-ai.org.cn/`, enabling users to upload their own data.

Acknowledgments

This work was done during K.Z.'s internship at Shanghai Artificial Intelligence Laboratory. We thank Professor Guoyin Yin from Wuhan University for providing two newly discovered compounds for testing. This work is supported by the Shanghai Artificial Intelligence Laboratory. W.H. was funded by National Natural Science Foundation of China (62406192), MSR Asia StarTrack Scholars Program, Tencent WeChat Rhino-Bird Focused Research Program, and Doubao LLM Fund.

Contributions:. Y.L. conceived the original idea. W.H., H-S.Z., and Y.L. jointly supervised the entire project. K.Z. implemented the data preparation, model training and evaluation, crystallographic computation, and error corrections. All authors contributed to advancing the research, determining the technical approach, analyzing the results, and writing the manuscript.

Competing interests:. There are no competing interests to declare.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryinformation.pdf