

Received 24 July 2025

Accepted 19 September 2025

Edited by Th. Proffen, Oak Ridge National Laboratory, USA

‡ These authors contributed equally to this work.

Keywords: crystal structure solution; phase seeding; artificial intelligence; AI phasing.

Supporting information: this article has supporting information at journals.iucr.org/j

The AI-based phase-seeding (AI-PhaSeed) method: early applications and statistical analysis

Benedetta Carrozzini,‡ Francesca Fedele,‡ Anna Moliterni,* Liberato De Caro, Corrado Cuocci, Cinzia Giannini, Rocco Caliandro and Angela Altomare

Institute of Crystallography, National Research Council, via Amendola 122/o, Bari, 70126, Italy. *Correspondence e-mail: anna.moliterni@cnr.it

The crystallographic challenge of structure determination is nowadays effectively supported by advanced computational methods, such as direct methods and Patterson techniques, implemented in sophisticated software. With the rapid expansion of artificial intelligence (AI) across diverse scientific domains, exploring its potential contribution to structure solution and its ability to overcome the limitations of traditional approaches has become increasingly compelling. This work builds upon and extends the findings of two recent studies on AI-driven phasing. The first, by Larsen *et al.* [*Science* (2024), **385**, 522–528], focused on designing and applying a neural network architecture to solve small structures (with unit-cell volumes up to 1000 Å³), primarily within the most common centrosymmetric space group $P2_1/c$. The second, by Carrozzini *et al.* [*Acta Cryst.* (2025), **A81**, 188–201], introduced a novel phase-seeding method applicable to both centrosymmetric and non-centrosymmetric crystal structures of varying complexity, from small molecules to proteins. Although designed with AI integration in mind, this latter method had not yet been tested within an AI framework. In this paper, we apply the method proposed by Carrozzini *et al.* to cases where seed phases are generated by the AI network developed by Larsen *et al.* We demonstrate that this combined approach, termed AI-PhaSeed, successfully extends the applicability of Larsen's neural network to structures with unit-cell volumes exceeding 1000 Å³, even under conditions of limited experimental resolution. The proposed procedure has been extensively tested on a set of structures taken from the Crystallography Open Database, proving it to be a powerful and reliable tool for structure solution. We also provide insights into the use of AI for crystallographic phasing and introduce statistical tools to evaluate the robustness of the solution process based on AI-calculated phases.

1. Abbreviations

AI: artificial intelligence.

DM: direct methods.

MPE: mean phase error.

 R_f : crystallographic agreement factor.

 N_{asym} : number of non-hydrogen atoms in the asymmetric unit.

 N_{ref} : number of measured symmetry-independent reflections.

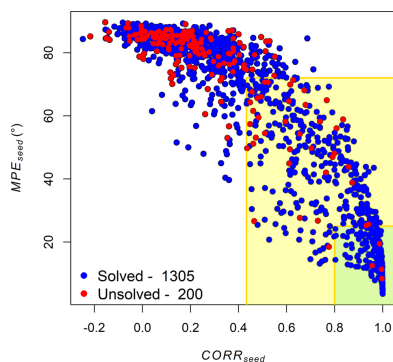
EDM: electron-density modification.

 E_h : normalized structure factor for reflection h .

CORR: the correlation coefficient between phased and true electron-density maps.

2. Introduction

The phase problem in crystallography has long represented a central challenge, complicating the determination of crystal structures for organic, inorganic or metal–organic structures regardless of their asymmetric unit size, whether small (<80 non-hydrogen atoms), medium (<300) or large (≥ 300). Today,



OPEN ACCESS

Published under a CC BY 4.0 licence

the routine solution of small- and medium-sized single-crystal structures using X-ray diffraction data collected in standard laboratories is largely feasible, provided the experimental data are of sufficient quality and resolution. This progress is the result of substantial theoretical and methodological advancements, including the development of direct methods (Giacovazzo, 2013), Patterson techniques (Rius, 2014) and dual-space approaches such as charge flipping (Oszlányi & Sütő, 2011). These approaches have been further empowered by significant improvements in diffractometer instrumentation and, crucially, in software packages that enhance modern computational capabilities (Burla *et al.*, 2015; Rius, 2011; Palatinus & Chapuis, 2007). Nevertheless, structure solution remains a challenge in some cases, especially for large molecules. Difficulties also emerge when the experimental resolution is far from atomic, ~ 1.2 Å according to the ‘Sheldrick rule’ (Sheldrick, 1990), which is the threshold at which individual atoms can be clearly resolved in an electron-density map.

The application of artificial intelligence (AI) is currently being explored across various scientific disciplines. In the field of structural crystallography, one of the most significant breakthroughs has been the development of *AlphaFold* by Google DeepMind (Jumper *et al.*, 2021), which has demonstrated unprecedented accuracy in predicting protein 3D structures directly from amino acid sequences. This milestone has underscored the transformative potential of AI in structural biology and opened new avenues for accelerating structure determination beyond traditional experimental approaches.

In a pioneering study, Larsen *et al.* (2024) demonstrated that AI can be effectively harnessed to address the phase problem in crystallography, successfully solving *ab initio* crystal structures with unit-cell volumes up to 1000 Å^3 using only experimental amplitude data. Their deep learning architecture was specifically designed for small structures, primarily within the space group $P2_1/c$, and was trained on a dataset comprising millions of artificial crystal structures. While the work of Larsen *et al.* (2024) was focused on a relatively narrow class of structures, some of which may be solvable through traditional direct methods or Patterson techniques, its findings represent a major step forward. Notably, the study demonstrated that AI can determine missing phases, a task that has traditionally relied on decades of theoretical and algorithmic development in crystallography. Beyond its immediate achievements, this work opened new avenues for applying AI to more complex scenarios that remain challenging for conventional methods, particularly when the experimental resolution is limited. These opportunities can be further explored using the recently developed phase-seeding method (Carrozzini *et al.*, 2025), which was explicitly designed to be AI compatible and applicable to crystal structures of any size and space group. The method is based on the accurate identification of a small subset of seed phases; when these are combined with experimentally derived amplitudes and randomly initialized phases, they can drive the reconstruction of the full electron-density map. This is accomplished by iterative procedures in both

direct and reciprocal space, enhanced by electron-density modification (EDM) cycles to extend and refine the phase set. For non-centrosymmetric structures, the method introduces a discretization of both seed and random phases into a limited number of distinct values, thereby enabling classification and refinement through an AI-driven algorithm. Although this method has been robustly defined and tested *in silico*, it had not yet been implemented in a fully AI-based framework prior to the present work.

This paper presents the first application of the phase-seeding method, in combination with the neural network developed by Larsen *et al.* (2024), extending its use to structures still within the space group $P2_1/c$ but with unit-cell volumes ranging from 1000 Å^3 up to 3500 Å^3 , well beyond the previously studied limit below 1000 Å^3 . These structures include organic, inorganic and metal–organic compounds. Despite the original design constraints of the neural network, the combined AI-based phase-seeding approach has demonstrated promising performance, benefiting from the advances introduced by the phase-seeding method, which requires reliable phase estimates for only a limited subset of reflections.

The novelty of this study lies in its clear demonstration that the AI model developed by Larsen *et al.* (2024) can be extended beyond its initial volumetric limitations through the integration of the phase-seeding strategy. We provide a comprehensive assessment of the capabilities and limitations of the AI-based phase-seeding (AI-PhaSeed) method. Validated on a large dataset of experimentally determined structures, AI-PhaSeed proves to be a robust and competitive alternative to traditional phasing approaches such as direct methods (DM).

Particularly noteworthy is the section addressing phasing from limited-resolution data, where we demonstrate that AI-PhaSeed can successfully solve structures that remain intractable using traditional methods.

3. Methods

3.1. The AI-based phase-seeding (AI-PhaSeed) method

The AI-based phase-seeding method integrates the concept of phase seeding with AI to solve crystal structures from X-ray diffraction data, once the unit cell and space group have been determined. Its effectiveness has been tested on real small-sized structures with unit-cell volumes ranging from 1000 to 3500 Å^3 and space group $P2_1/c$, deposited in the Crystallography Open Database (COD) (Gražulis *et al.*, 2009; Downs & Hall-Wallace, 2003).

AI-PhaSeed involves the following steps outlined in Fig. 1:

(i) For each real structure in the centrosymmetric space group $P2_1/c$, a subset of experimental reflections is selected to meet the input size requirements of the PhAI neural network (Larsen *et al.*, 2024). The selected reflections follow the constraint on Miller indices (hkl) such that $(h^2 + k^2 + l^2)^{1/2} \leq 10$ and, due to the Laue symmetry $2/m$, $-10 < h < 10$, $0 < k < 10$ and $0 < l < 10$. This selection is performed automatically by a built-in tool within *SIR2024*, the latest updated version of the

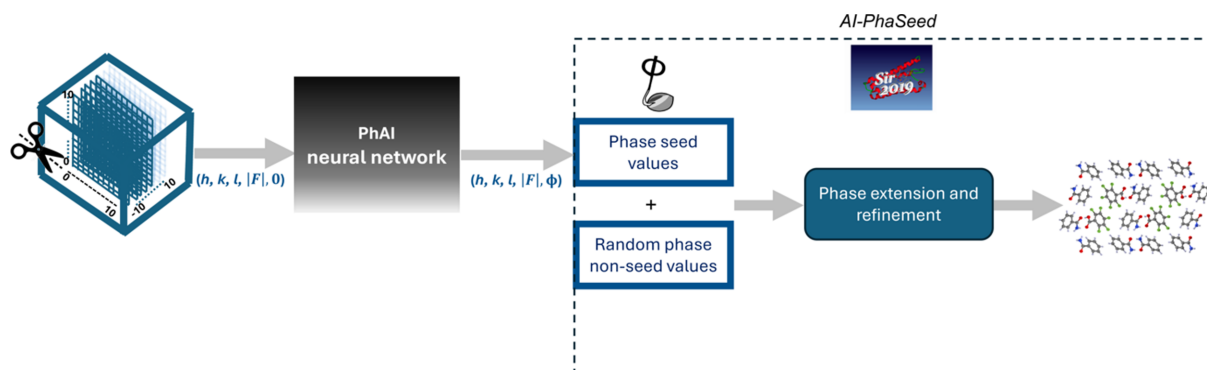


Figure 1

Scheme of the AI-PhaSeed method. The image of the phase seeding is taken from Madsen (2025) and reproduced with permission.

SIR2014 package (Burla *et al.*, 2015). For each hkl reflection, the amplitude and an initial phase set to zero are provided as input to PhAI, and the run was set with five phasing cycles.

This subset constitutes only a small fraction of the total number of reflections for each structure.

(ii) The phases predicted by the PhAI network (restricted to 0 or π) provide the phase-seed values required by the phase-seeding procedure (Carrozzini *et al.*, 2025). The efficiency of the AI-generated phases is evaluated using the mean phase error (MPE_{seed}), calculated as the average difference between the AI-derived phases and those derived from the known structural model, as well as the correlation coefficient ($\text{CORR}_{\text{seed}}$) as defined by Larsen *et al.* (2024).

(iii) For non-seed reflections, phases are randomly assigned as either 0 or π . The subsequent phase extension and refinement procedure employs iterative EDM cycles, using the experimental amplitudes as constraints (Carrozzini *et al.*, 2025). The final electron-density map is computed during this process, which is executed using the *SIR2024* software (Burla *et al.*, 2015). Three validation parameters are used to evaluate the reliability and accuracy of both the phasing process and the resulting structure solution: (1) the final mean phase error ($\text{MPE}_{\text{final}}$), calculated in reciprocal space as the deviation between the refined and model-derived phases; (2) the final correlation coefficient ($\text{CORR}_{\text{final}}$); and (3) the crystallographic agreement factor (R_f) obtained by comparing calculated and observed amplitudes.

3.2. AI-PhaSeed combined with direct methods

In the AI-PhaSeed method, the phases generated by AI are extended and refined using direct-space methods applying cycles of electron-density map modifications (Burla *et al.*, 2010) (path A in Fig. 2). The conventional phasing protocol, typically performed using DM for small- and medium-sized compounds, involves initial phase determination followed by phase extension and refinement in direct space, the same approach as adopted in the AI-PhaSeed method (path B in Fig. 2).

In this study we present the first applications of the AI-PhaSeed method and compare its performance with the classical DM phasing procedure (path A versus path B in Fig. 2).

In addition, we propose a strategy, called DM+AI-PhaSeed (see Section 4.2), that combines the two procedures (path C in Fig. 2), based on the theory outlined in Section 3.3. In this approach, the AI-generated phase seed is actively integrated into DM to drive the multi-solution process more effectively towards a reliable phase set.

3.3. AI-based tangent formula

The direct methods approach exploits structure invariants, phase combinations unaffected by origin shifts, typically derived from a set of strong reflections ($|E_h| > 1.2$), referred to as N_{large} . Phases are estimated from known amplitudes via the tangent formula (Karle & Hauptman, 1956), which provides probabilistic phase values based on statistical relationships. The phasing process starts with random phase assignment to N_{large} reflections followed by iterative refinement. Since convergence is not always achieved, the procedure is repeated with different random inputs, a multi-solution approach that increases the chance of identifying correct phase estimates.

According to the formalism described by Giacovazzo (1998, pp. 123–127), the total probability $P(\theta_h)$ that the phase of E_h involved in r structure-invariant relationships (specifically triplets, in this case) equals θ_h can be expressed as a suitably normalized product of r specific probabilities,

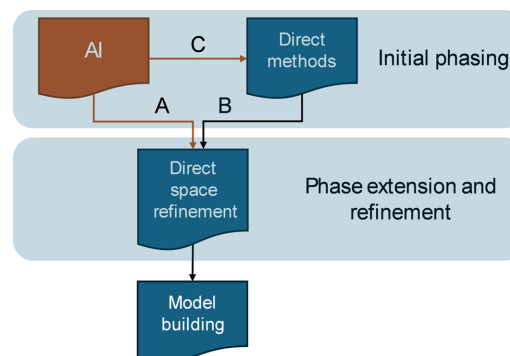


Figure 2

Workflow of the AI-PhaSeed method (path A), the classical phasing procedure (path B) and their combination DM+AI-PhaSeed (path C).

Table 1

Summary of key features of the test structures, presented as the variable names (first column), their description (second column) and the corresponding variability range (third column).

Variable name	Description	Variability range
N_{asym}	Number of non-H atoms in the asymmetric unit	9.5–55
maxCellSize (Å)	Maximum linear dimension of the unit cell (maximum between a , b , c)	10.7–43.1
minCellSize (Å)	Minimum linear dimension of the unit cell (minimum between a , b , c)	3.68–14.8
Perc	Percentage of phase-seed reflections that meet the neural network's input size criteria, over the total number of symmetry-independent reflections	5.03–53.9%†
Compl (%)	Crystallographic data completeness	29.7–100%
max W	Atomic weight of the heaviest element	~12–238
Pseudo (%)	Pseudosymmetry percentage affecting atoms in the structure	0–96%
N_{ref}	Number of measured symmetry-independent reflections	1508–20169
RES (Å)	Experimental data resolution	0.49–0.98
REFLEC _{seed}	Number of reflections phased by AI	894–1042
RES _{seed} (Å)	Resolution for reflections phased by AI	0.59–1.37
Vol (Å ³)	Volume of the unit cell	1005–3495

† On average, Perc exceeds 10%, ensuring sufficient data for effective phase prediction.

$$P(\theta_h) = \prod_{j=1}^r P_j(\theta_h) = L^{-1} \exp \left[\sum_k w_{h,k} \cos(\theta_h - \phi_k - \phi_{h-k}) \right], \quad (1)$$

where L is a normalizing factor, and ϕ_k and ϕ_{h-k} are phases of reflections contributing to the invariants. Usually, for the tangent formula, we have

$$w_{h,k} = G_{h,k} = 2\sigma_3\sigma_2^{-3/2} |E_h E_k E_{h-k}|, \quad (2)$$

where $\sigma_n = \sum_{j=1}^N Z_j^n$ and Z_j is the atomic number of the j th atom.

When *a priori* phase information is available, such as AI-derived phase estimates for a subset of N_{large} reflections, this information can be directly combined into equation (1), with random phases assigned to the rest. *A priori* phase information could be generated by running a neural network multiple times with different random seeds. However, results from this study indicate that performing multiple random trials to generate phase histograms offers limited benefit. Consequently, a single AI-derived phase seed, as used in this work, is sufficient to drive the phasing process effectively, reducing computational cost without compromising accuracy.

Thus, incorporating *a priori* knowledge from AI-derived phase values with reliability G_h^{AI} , the total probability $P_{\text{tot}}(\theta_h)$ that the phase of E_h is θ_h can be written as (Giacovazzo, 1998, pp. 564–565)

$$P_{\text{tot}}(\theta_h) = \prod_{j=1}^r P_j(\theta_h) P^{\text{AI}}(\theta_h) = L^{-1} \exp \left[\sum_k G_{h,k} \cos(\theta_h - \phi_k - \phi_{h-k}) + G_h^{\text{AI}} \cos(\theta_h - \theta_h^{\text{AI}}) \right]. \quad (3)$$

If we treat the AI-derived phase estimates as triplets, we can always estimate G_h^{AI} through equation (2) and we have

$$\tan \theta_h = \frac{\sum_k [G_{h,k} \sin(\phi_k + \phi_{h-k})] + G_h^{\text{AI}} \sin \theta_h^{\text{AI}}}{\sum_k [G_{h,k} \cos(\phi_k + \phi_{h-k})] + G_h^{\text{AI}} \cos \theta_h^{\text{AI}}} = \frac{T_h}{B_h}, \quad (4)$$

with statistical weight given by

$$Z_{h,k}^{\text{AI}} = \left\{ [G_{h,k} \sin(\theta_h - \phi_k - \phi_{h-k}) + G_h^{\text{AI}} \sin(\theta_h - \theta_h^{\text{AI}})]^2 + [G_{h,k} \cos(\theta_h - \phi_k - \phi_{h-k}) + G_h^{\text{AI}} \cos(\theta_h - \theta_h^{\text{AI}})]^2 \right\}^{1/2}. \quad (5)$$

Equations (3)–(5) generalize the tangent formula to account for *a priori* knowledge from AI-derived phase estimates. When AI estimates of both ϕ_k and ϕ_{h-k} are available, they jointly contribute to equation (5), enhancing the reliability of the corresponding structure invariant. If only one phase is available, it still provides a partial but meaningful enhancement to the triplet's reliability. This approach allows external information (AI estimates) to be incorporated in a natural and consistent manner within the formalism of DM. The *a priori* phase information does not override the estimates derived from triplet relationships, but rather merges with them according to their relative reliability. Mathematically, this is equivalent to the weighted vector sum of phase estimates in the complex plane.

3.4. Test structures and relevant variables

Test structures were obtained from the COD (Grazulis *et al.*, 2009; Downs & Hall-Wallace, 2003) using tools integrated into the software package *EXPO* (Altomare *et al.*, 2013). We selected 1505 crystal structures in the $P2_1/c$ space group, all solved using single-crystal X-ray diffraction data. This choice aligns with the PhAI neural network (Larsen *et al.*, 2024) used in our study to generate AI phases, which was specifically trained on structures in the $P2_1/c$ space group. In the AI-PhaSeed method, the AI-generated phases are subsequently extended and refined. Additionally, in accordance with the capabilities of the PhAI neural network designed to perform phasing on datasets with unit-cell volumes below 1000 Å³ and constrained Miller indices (hkl), we restricted our selection to structures with unit-cell volumes between 1000 and 3500 Å³. This range ensures compatibility with the AI-PhaSeed method, which, according to Carrozzini *et al.* (2025), requires phase estimates for a subset (typically 10%) of the total

reflections. The main features of the selected structures are summarized in Table 1, and they are also illustrated as box plots in Fig. S1 in the supporting information.

4. Results

As a first check, we verified the performance of the standalone PhAI method (without phase seeding or EDM cycles) on the dataset used in this study. The results, shown in Fig. S2, indicate that, even for structures for which the phase values assigned by PhAI are reliable ($\text{MPE}_{\text{seed}} < 20^\circ$), the low-resolution electron-density map that can be calculated using them cannot be properly interpreted in terms of an atomic model. As a result, the percentage of correctly positioned atoms is less than 20% for most of the test structures. Thus, adopting the phase-seeding strategy to generate the PhAI phases is essential for the structures considered in this study.

Given the size of the test dataset (1505 structures), the performance of the AI-PhaSeed method must be evaluated by a robust metric capable of automatically and unambiguously distinguishing solved structures from unsolved ones. To this end, we considered the space defined by three validation parameters ($\text{MPE}_{\text{final}}$, $\text{CORR}_{\text{final}}$ and R_f), each calculated with respect to the published structure, to evaluate the outcome of the structure solution process. Representative points of each test structure are projected into this space [Fig. 3(a)] and clustered by a k -means clustering procedure with $k = 2$ (Hartigan & Wong, 1979). This results in two identified clus-

ters: one representing the solved structures [blue points in Fig. 3(a)] and the other representing the unsolved structures [red points in Fig. 3(a)]. The two clusters are clearly separated, with the points corresponding to the solved structures accumulating around the optimal values of the validation parameters. As a result, the silhouette plot shown in Fig. 3(b) is dominated by the cluster of solved structures, most of which have a silhouette width larger than 0.8.

The unambiguous definition of the solved structures allows us to calculate the efficiency of the AI-PhaSeed method as the ratio of solved to total structures, which comes to 86.7%. This value should be compared with that obtained using a random phase seed (random phases replacing AI phases), which is 84.3%, and that from a true phase seed (phases calculated from published structures replacing AI phases), which is 97.1%. As expected, the AI-PhaSeed method performs slightly better than the random-phase case but is quite far from what can be obtained using known phases. This gap is mainly due to a subset of structures for which AI fails to assign reliable phase values to seed reflections. However, if we focus on structures whose AI-generated seed phases exhibit a mean phase error (MPE_{seed}) below 25° and a reflection correlation ($\text{CORR}_{\text{seed}}$) greater than 0.8, *i.e.* the 300 structures for which the AI predictions can be considered reliable, the efficiency of the AI-PhaSeed method reaches 98%, closely aligning with the 97% achieved by the true-phase method on the same subset of structures. This is better shown in Fig. 4, where the k -means clusters of solved and unsolved structures are projected

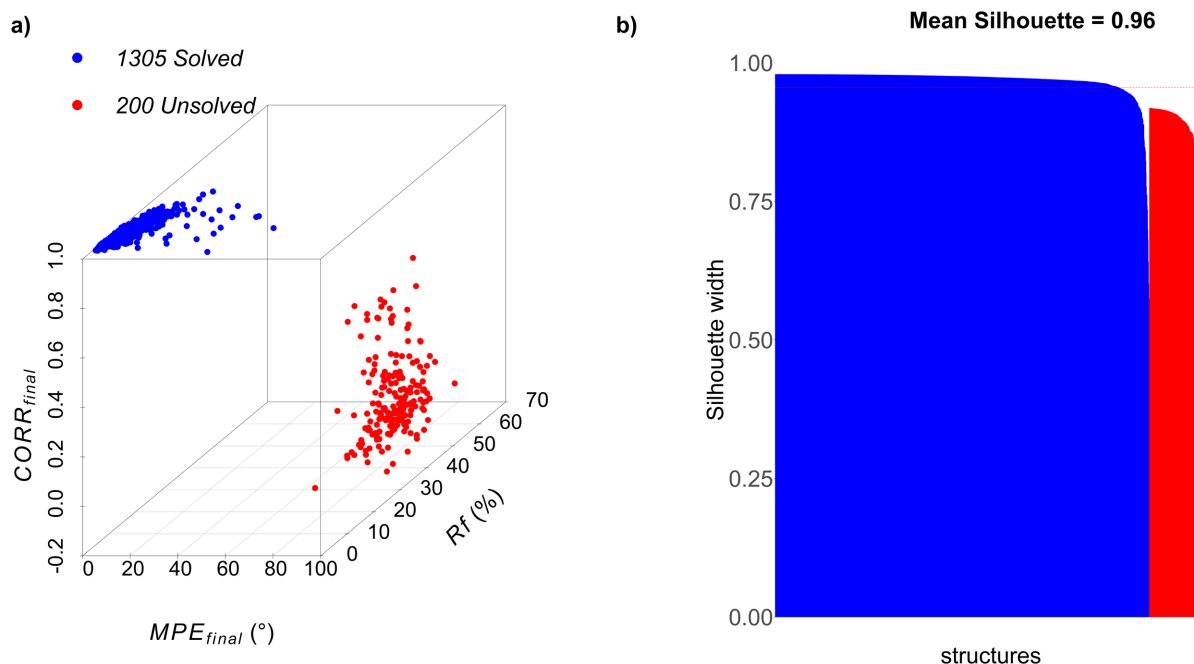


Figure 3

(a) A 3D scatter plot illustrating the k -means clustering results, with $k = 2$ (Hartigan–Wong algorithm), based on $\text{MPE}_{\text{final}}$ (x axis), R_f (y axis) and $\text{CORR}_{\text{final}}$ (z axis). Data points are colour-coded to distinguish between ‘Solved structures’ (blue) and ‘Unsolved structures’ (red), with the legend indicating the number of observations in each category. (b) A silhouette plot for k -means clustering with $k = 2$. Each block contains a number of vertical bars corresponding to the elements assigned to that cluster (blue for solved and red for unsolved structures). The height of each bar represents the silhouette width, indicating how well each element fits within its assigned cluster compared with the other one. Silhouette values closer to 1 correspond to better-defined clustering. The value of the mean silhouette width, shown above the plot, quantifies the overall clustering quality; values above 0.7 are generally considered indicative of strong and well separated clusters (Kaufman & Rousseeuw, 1990).

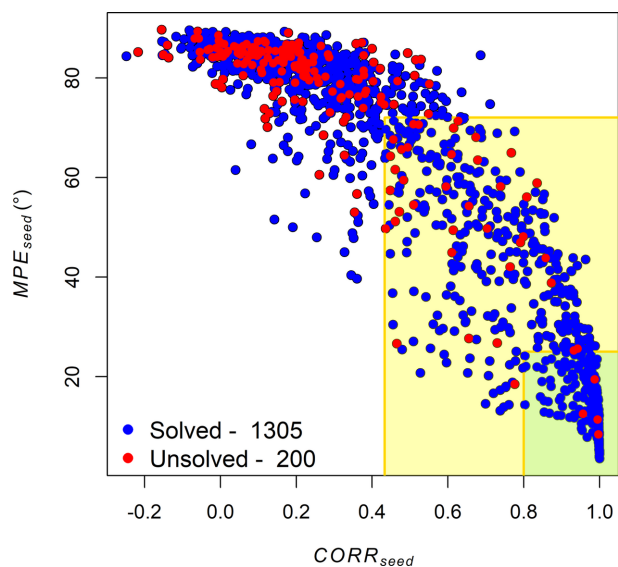


Figure 4

Scatter plot of MPE_{seed} versus $CORR_{seed}$, showing clusters of solved (blue dots) and unsolved (red dots) structures. The green region corresponds to the area defined by $MPE_{seed} \leq 25$ and $CORR_{seed} \geq 0.8$ where the AI-generated phases can be considered reliable, *i.e.* sufficiently accurate to approximate the true phases. The yellow region corresponds to the area defined by optimizing the values $MPE_{seed} \leq Q1$ and $CORR_{seed} \geq Q3$ (Class 1). This region predominantly contains solved structures, which tend to accumulate at $CORR_{seed}$ values close to 1.

in the space defined by the two variables MPE_{seed} and $CORR_{seed}$ (see Section 3.1): the green box contains the 300 structures with MPE_{seed} below 25° and $CORR_{seed}$ greater than 0.8, and only four of them (red dots in the green box) belong

to the unsolved k -means cluster. This outcome suggests that the AI-PhaSeed method performs as proposed by Carrozzini *et al.* (2025), assuming the AI-predicted seed phases are reliable.

However, this high efficiency is achieved for only approximately 20% of the total structures, specifically those that closely match the features of the neural network's training set. To investigate the nature of the decline in AI efficiency and identify the structures for which this occurs, we developed a classification model to explain the AI-PhaSeed results.

To this end, we focused on analysing the intermediate result of the method, the one obtained after application of AI and related to the seed phases, by considering the two variables MPE_{seed} and $CORR_{seed}$.

Well separated distributions, with a minimal overlap, are obtained when comparing the MPE_{seed} and $CORR_{seed}$ values for structures belonging to the solved k -means cluster with those belonging to the unsolved k -means cluster. This result is illustrated in Fig. S3. The observed separation, while confirming the discriminative power of MPE_{seed} and $CORR_{seed}$ in distinguishing between solved and unsolved structures and thus confirming the impact of AI-generated phases on the performance of the AI-PhaSeed method, also allows the identification of a threshold value for both variables, *i.e.* the first quartile (Q1) for MPE_{seed} and the third quartile (Q3) for $CORR_{seed}$, both of the unsolved cluster. This ensures the identification of the maximum number of solved structures, while simultaneously minimizing the number of unsolved ones. These thresholds identify the yellow box in Fig. 4, which mostly contains solved structures accumulating at $CORR_{seed}$ values close to 1.

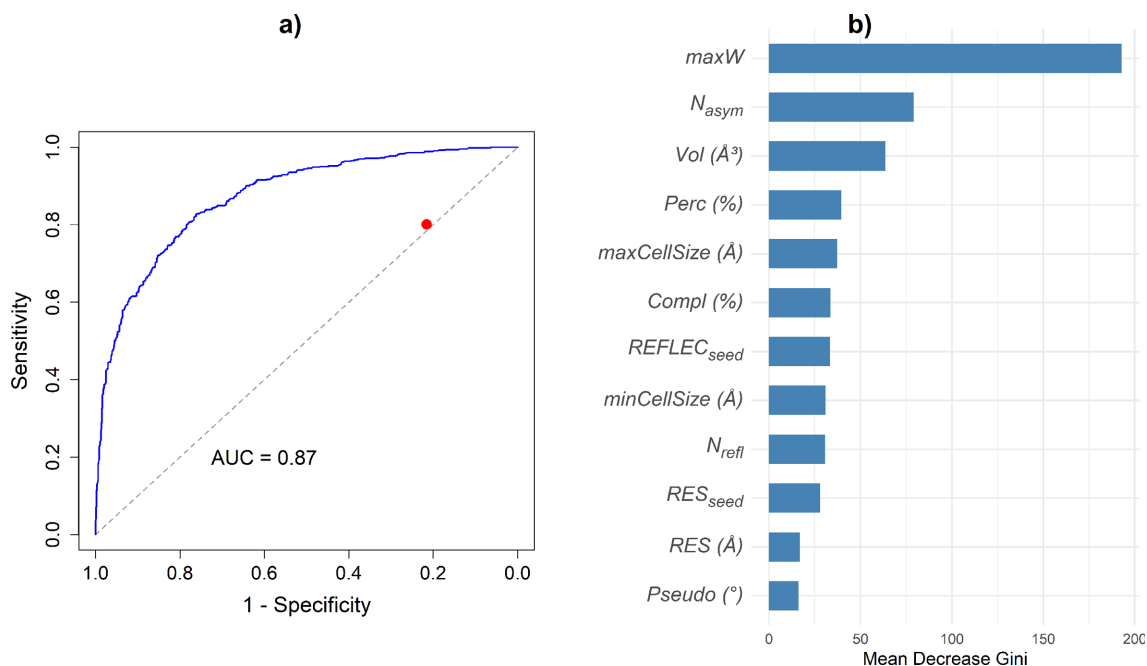


Figure 5

(a) The ROC curve of the RF model, illustrating sensitivity and specificity values across varying thresholds. The red dot marks the optimal cut-off point, identified as the highest specificity value corresponding to a specificity of at least 80%, for the correct classification of a significant portion of Class 1 structures. The AUC is reported as a measure of the overall performance of the model. (b) Feature importance, ordered by the descending mean decrease Gini (Kaufman & Rousseeuw, 1990) resulting from the RF model.

Table 2

Confusion matrix for the tenfold cross-validation RF model, showing predicted and actual class distributions.

Class labels were assigned using an optimal probability threshold, as determined by identifying the highest sensitivity value corresponding to a specificity of at least 80% (for correct classification of a significant portion of Class 1 structures) based on the global ROC curve [red dot in Fig. 5(a)]. Class 1 corresponds to $MPE_{seed} \leq Q1$ and $CORR_{seed} \geq Q3$, and Class 0 to $MPE_{seed} > Q1$ and $CORR_{seed} < Q3$.

	Actual: 0	Actual: 1
Predicted: 0	657	133
Predicted: 1	183	532

We employed a random forest (RF) classification model (Breiman, 2001) as a supervised learning algorithm, labelling structures within the yellow region in Fig. 4 as Class 1 and all others as Class 0. The model was thus trained to classify as successful (Class 1) those structures with $MPE_{seed} \leq Q1$ and $CORR_{seed} \geq Q3$, *i.e.* those falling within the yellow box in Fig. 4, and as unsuccessful (Class 0) those lying outside this region. The main goal in building the classification model is to study how the quality of the AI phases influences the outcome of the structure solution. Consequently, a small number of structures (approximately 3% of the overall dataset) that remain unsolved even with the true phase seed are excluded from this stage of the analysis to avoid biases unrelated to phase quality.

The RF results are shown in Fig. 5. The optimal probability threshold on the receiver operating characteristic (ROC) curve (Fawcett, 2006) [red dot in Fig. 5(a)] was determined to classify correctly a significant portion of Class 1 structures by selecting the highest sensitivity value corresponding to a specificity of at least 80%. The area under the curve (AUC) value is 0.87, indicating an excellent overall model performance. The resulting confusion matrix for the tenfold cross-validated RF model is reported in Table 2.

The RF model can be effectively used to optimize the decision-making process and resource allocation in crystal

structure determination by applying the AI-PhaSeed method only to structures classified as Class 1. For these structures, the efficiency of AI-PhaSeed increases to 90%, compared with 85% with a random phase seed. This indicates that when AI provides reliable seed phases the AI-PhaSeed method performs close to the optimal scenario where true phases are used as seeds (96% efficiency).

As noted above, a key advantage of developing a classification model to assess the performance of AI-PhaSeed is its ability to evaluate the importance of different structural variables in determining the success of the phasing process. The feature importance, ordered by the descending mean decrease Gini coefficient (Kaufman & Rousseeuw, 1990), is shown in Fig. 5(b) and reveals that $maxW$, *i.e.* the largest atomic weight among the elements in the unit cell, is a key variable. This reflects a well established principle in crystallography: the presence of heavy atoms can facilitate the phasing process. What is particularly noteworthy is that the importance of this feature was identified by the RF classification model, even though no explicit information about atomic weight was provided during the neural network's training phase. This represents a significant insight into the application of AI in structure determination, as it highlights the deep learning model's capacity to extract physically meaningful information from data autonomously.

Subsequent relevant variables, ordered by the mean decrease Gini (Kaufman & Rousseeuw, 1990) in Fig. 5(b), are N_{asym} , Vol and $maxCellSize$, related to the size of the unit cell in direct space, and Perc and $REFLEC_{seed}$, related to the size of the unit cell in reciprocal space. These findings indicate that another key parameter influencing the efficiency of the AI-based phase assignment is the complexity of the crystal structure (*e.g.* in terms of number of non-H atoms in the asymmetric unit). This is closely linked to the limited size of the phase seed: for a larger crystal structure more effort is required to extend the phase information to all measured reflections. To verify the effect of the variables highlighted by the feature importance analysis, we plot in Fig. 6 the distribution

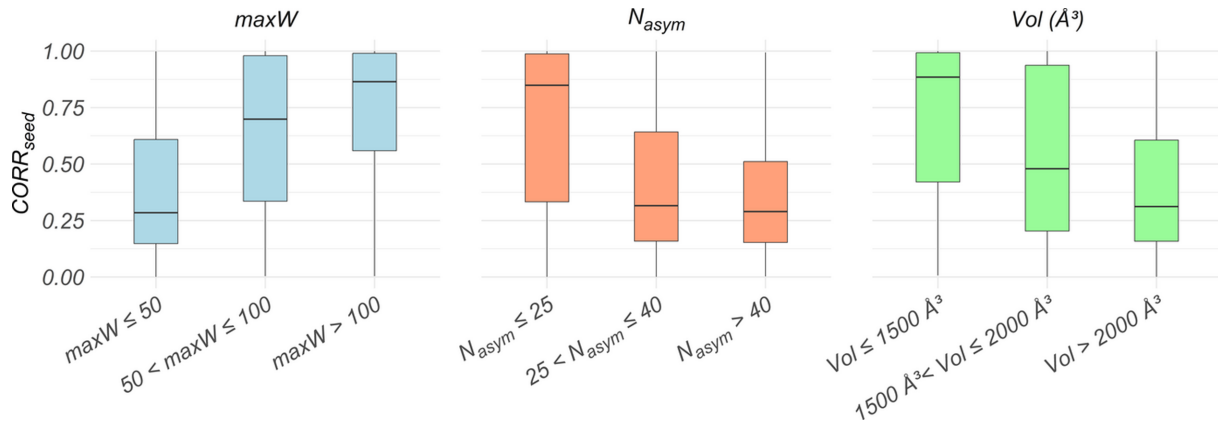


Figure 6

Distribution of $CORR_{seed}$ across different intervals of the three most important features according to the RF mean decrease Gini [Fig. 5(b)]: maximum atomic weight among the structure elements ($maxW \leq 50$, $50 < maxW \leq 100$ and $maxW > 100$), number of non-H atoms in the asymmetric unit ($N_{asym} \leq 25$, $25 < N_{asym} \leq 40$ and $N_{asym} > 40$) and unit-cell volume ($Vol \leq 1500 \text{ \AA}^3$, $1500 < Vol \leq 2000 \text{ \AA}^3$ and $Vol > 2000 \text{ \AA}^3$). Horizontal black bars indicate median values.

of $\text{CORR}_{\text{seed}}$ across different intervals of the three most important features according to the RF mean decrease Gini [Fig. 5(b)]: $\text{max}W$ ($\text{max}W \leq 50$, $50 < \text{max}W \leq 100$ and $\text{max}W > 100$), N_{asym} ($N_{\text{asym}} \leq 25$, $25 < N_{\text{asym}} \leq 40$ and $N_{\text{asym}} > 40$) and Vol ($\text{Vol} \leq 1500$, $1500 < \text{Vol} \leq 2000$ and $\text{Vol} > 2000$). As expected, a clear increase in $\text{CORR}_{\text{seed}}$ values is observed in intervals closer to the optimal values of the variable, *i.e.* for higher $\text{max}W$ values and lower N_{asym} and Vol values.

As a result of the classification method, we can now estimate the performance on Class 1 structures. The AI-PhaSeed method shows improved efficiency, increasing from 86.7% on the entire dataset to 91.1% when applied specifically to the Class 1 subset.

4.1. AI phasing at limited data resolution

Given the challenging task of solving structures at resolutions far from the atomic one, we explored the performance of AI-PhaSeed as the data resolution was progressively decreased, being aware that the size of the seed relative to the total number of observed reflections plays a significant role. The analysis revealed that both MPE_{seed} and $\text{CORR}_{\text{seed}}$ values are distributed differently across the datasets obtained after applying the resolution cut-offs (*i.e.* 1.0, 1.2, 1.4 and 1.6 Å). Additionally, the criteria for distinguishing solved from unsolved structures and for separating Class 0 and Class 1 structures after application of AI become less distinct compared with the corresponding results without a resolution cut-off. However, the mean silhouette width value remains above 0.7 (Kaufman & Rousseeuw, 1990), still indicating an appreciable degree of separation. Figs. S4, S5 and S6 present

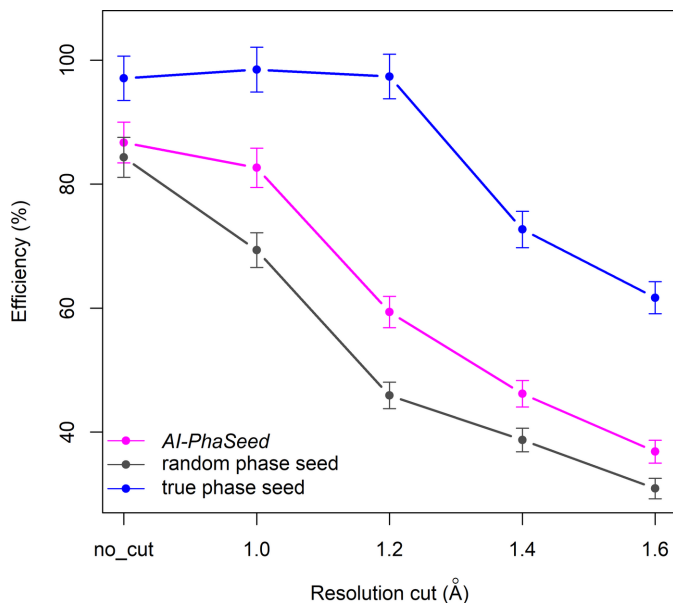


Figure 7

Efficiency of the AI-PhaSeed method as a function of data resolution cut-offs, compared with the efficiency obtained using random phase and true phase seeds. Five datasets were analysed: no resolution cut (no_cut), and cut-offs at 1.0, 1.2, 1.4 and 1.6 Å resolution. Error bars represent the propagated uncertainty, calculated as $\delta E = E(1/N + 1/D)^{1/2}$, where E is the efficiency (%), D is the number of correctly classified structures and N is the total number of structures.

the statistical analysis of the results obtained using a 1.6 Å resolution cut-off, following the same procedure as the tests on native uncut structures. At this resolution, 23 structures exhibited statistical parameters that were unsuitable for applying DM (*e.g.* an insufficient number of structure-invariant relationships) and were therefore excluded from the dataset. The results of the RF model (Fig. S7 and Table S1) and the validation of the feature importance analysis (Fig. S8) are also provided.

The performance of the RF model is characterized by a still acceptable AUC of 0.75. The variable Perc shows increased importance compared with its role in the model trained on uncut data, emerging as the second most important variable after $\text{max}W$, as shown in Fig. S7(b) [compare with Fig. 5(b)].

Similar evaluations have been performed for all datasets corresponding to the various resolution cut-offs. The trend in AI-PhaSeed efficiency (E) plotted against the resolution cut-off is shown in Fig. 7 and compared with that obtained by feeding the seed with random or true phase values. In Fig. 7, error bars represent the propagated uncertainty, calculated as $\delta E = E(1/N + 1/D)^{1/2}$, where E is the efficiency (%), D is the number of correctly classified structures and N is the total number of structures. A common decreasing trend is observed across the curves, although the rates of decline differ. In particular, it can be noted that the efficiency of the true phase seed remains unaffected by the resolution cut-off up to 1.2 Å, while it drops substantially at 1.4 Å. Conversely, the random phase seed has a steady decrease in efficiency for data resolution up to 1.2 Å, which becomes less steep when going to 1.4 Å data resolution. The efficiency of the AI-PhaSeed method is always intermediate between the random and true

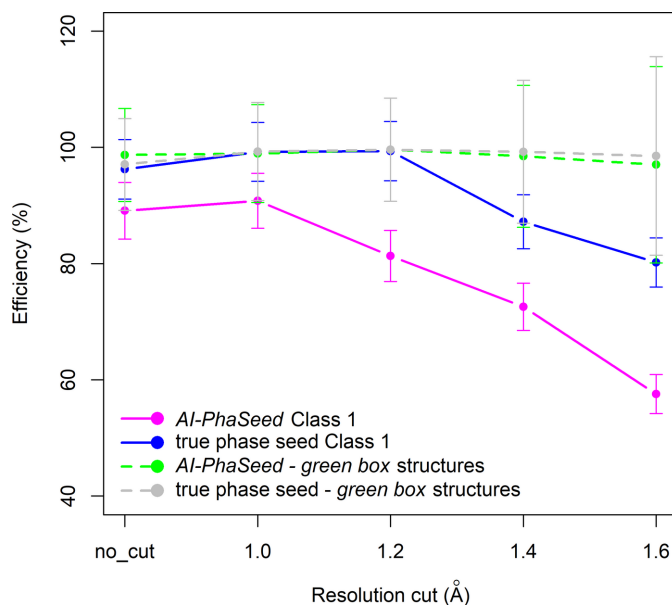


Figure 8

Efficiency of the AI-PhaSeed method applied on subsets of test structures as a function of the data resolution cut-offs, compared with the efficiency obtained using true phases (true phase seed). Full and dashed lines represent the efficiency calculated for structures belonging to their corresponding resolution-cut-off-based yellow (Class 1) and green boxes, as reported in Fig. S6. Error bars are also shown.

phase seed. The decrease in the rate of efficiency follows that of the random-phase-seed curve for data resolution between 1.0 and 1.2 Å, while it is less steep for data resolution <1.0 Å or between 1.2 and 1.4 Å.

An RF classification model was also built for each resolution cut-off, following the same procedure described above. As shown in Fig. 8, an increase in the efficiency of the AI-PhaSeed method for Class 1 structures with respect to the full set of structures (Fig. 7) is observed across all datasets, resulting in a smaller gap with the true phase-seeding efficiency. A satisfactory efficiency, close to 60%, for Class 1 structures is reached even at 1.6 Å resolution. Fig. 8 also shows the AI-PhaSeed efficiency calculated for structures with AI-generated phases very close to the true phases ($MPE_{seed} < 25$ and $CORR_{seed} > 0.8$), located in their corresponding green boxes as shown in Fig. S6 for the case at 1.6 Å; the efficiency is very close to 100%, regardless of the resolution cut-off, and almost coincides with the true phase-seeding values.

4.2. AI-PhaSeed combined with DM

The newly introduced AI-PhaSeed needs to be compared with the gold standard for *ab initio* crystal structure solution of small molecules, *i.e.* DM. With an efficiency of 86.7% on the entire dataset, AI-PhaSeed performs significantly below DM, which achieves 99.1% efficiency. This difference is expected: the DM approach has been under continuous development since the 1970s and is optimized for a wide variety of structures, whereas the potential of AI phasing is only beginning to be explored. With the aim of developing a more efficient phasing procedure than either method alone, we have integrated AI-PhaSeed with DM into a combined approach. The theoretical basis for using AI-derived phases to initiate DM

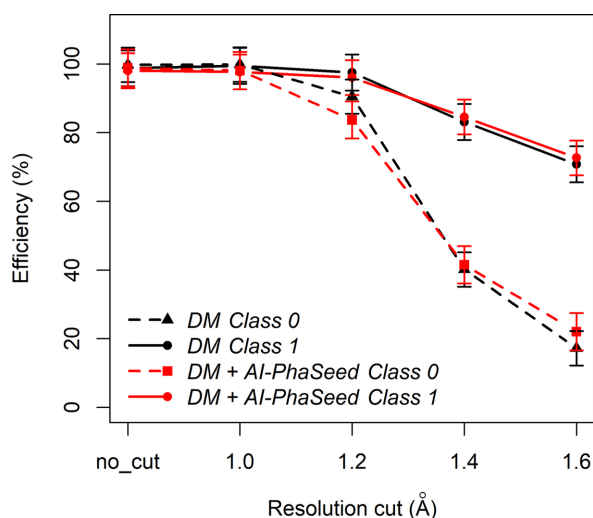


Figure 9

Efficiency of the DM and DM+AI-PhaSeed methods when applied to Class 0 and Class 1 test structures, plotted as a function of applied resolution cut-off level. Five different datasets were compared: no resolution cut (no_cut) and cut-offs at 1.0, 1.2, 1.4 and 1.6 Å resolution (Class 1 corresponds to $MPE_{seed} \leq Q1$ and $CORR_{seed} \geq Q3$, and Class 0 to $MPE_{seed} > Q1$ and $CORR_{seed} < Q3$). Error bars are also shown.

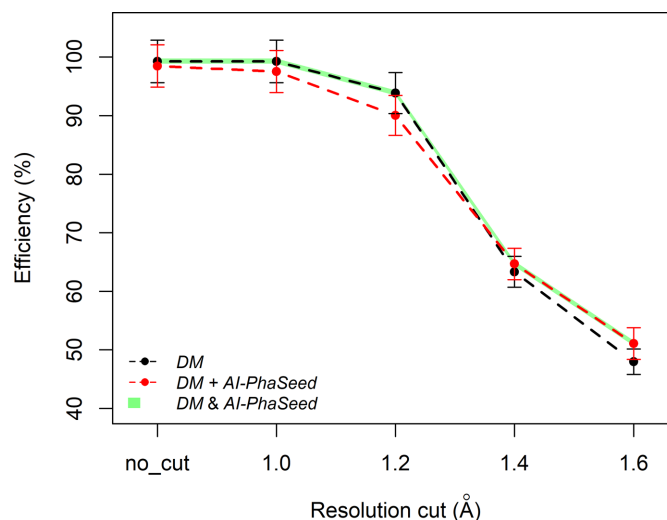


Figure 10

Efficiency of the AI-PhaSeed method combined with DM, plotted as a function of resolution cut-off level, compared with that obtained using DM alone and with the protocol DM&AI-PhaSeed.

phasing is outlined in Section 3.3. This protocol is denoted by DM+AI-PhaSeed. However, we have demonstrated that there are structures for which AI is not able to supply reliable phases, and these cases can be predicted with good accuracy by an RF model. Thus, we envisage that the combination of *AI-PhaseSeed* with DM could be conditioned by the results of the RF model prediction, by applying AI-PhaSeed only if the structure is classified as Class 1 by the RF model. In addition, by evaluating the efficiency of DM+AI-PhaSeed and DM separately for Class 0 and Class 1 structures (Fig. 9), we found that for structures with a data resolution (RES) of about 1.4 Å or worse (*i.e.* higher values such as 1.6 Å) DM+AI-PhaSeed outperforms DM for both classes. Thus, we can define an optimal solution strategy denoted by DM&AI-PhaSeed, according to which DM+AI-PhaSeed is applied to solve structures with lower data resolution (*i.e.* $RES > 1.4$ Å), while AI-PhaSeed alone is applied in cases of structures with higher resolution (*i.e.* $RES \leq 1.4$ Å). From Fig. 10 it can be seen that this protocol has the highest efficiency on the entire structure dataset.

5. Discussion and perspectives

We have demonstrated that the method proposed by Carrozzini *et al.* (2025), when suitably implemented in the AI-PhaSeed procedure, can successfully phase crystal structures that cannot be solved by the currently available stand-alone AI PhAI neural network. When the AI-predicted phase seed is reliable, its performance approaches that achieved using true phase seeds.

The results were assessed in two steps. First, the reliability of the AI-generated phase seed was evaluated, monitored by the variables MPE_{seed} and $CORR_{seed}$. Then, the phases obtained at the end of the AI-based solution process, where the AI-generated phases serve as a seed for traditional crystallographic methods, were assessed, using MPE_{final} and

$\text{CORR}_{\text{final}}$ as indicators. The performance of the first step was optimized using an RF classification model to predict the quality of the AI-generated phase seed. This model enabled us to analyse the dependence of AI performance on key variables related to the test structures, such as the percentage of the phase seed and the presence of heavy atoms.

For structures classified as optimal, the efficiency of AI-PhaSeed does not decline even at resolutions $>1 \text{ \AA}$, indicating that AI-PhaSeed is capable of phasing at resolution lower than the atomic level.

Besides the application of the method proposed by Carrozzini *et al.* (2025), we have made a step forward in integrating AI with *ab initio* phasing techniques, specifically DM, using the values of the phase seed generated by AI to support and reinforce the phase assignment performed by DM. To this end, we developed a DM+AI-PhaSeed integrated approach based on a modified tangent formula that combines AI-generated phases with those assigned by DM starting from random phases for reflections included in the phase seed.

The RF classification model also allowed us to optimize the application of the DM+AI-PhaSeed integrated approach. This was achieved by introducing a decision-making step based on structural features that are readily accessible after synthesis and X-ray measurement. With this fully integrated AI-based approach, we are able to phase real structures more efficiently than DM, even when the data resolution is $>1 \text{ \AA}$. This highlights the strong potential of the AI-PhaSeed method with low-resolution data.

In perspective, we expect that a substantial improvement in the AI-PhaSeed method and its integration with DM will result from the use of a neural network specifically trained on a larger number of investigated reflections. In the present study, we employed the PhAI neural network developed by Larsen *et al.* (2024), trained on reflection grids of fixed size ($21 \times 11 \times 11$). Our analysis identified that a critical factor for AI-PhaSeed performance is the percentage of the seed, defined as the number of seed reflections relative to the total number of experimental reflections.

Two key challenges must be addressed in the future to make this method suitable for routine structural determination of real structures: (i) extending AI-PhaSeed to structures with space groups other than $P2_1/c$, including non-centrosymmetric structures, and (ii) developing alternative criteria for selecting the phase seed. Carrozzini *et al.* (2025) have laid the groundwork for addressing these challenges. In particular, the application to non-centrosymmetric structures could be addressed by applying a phase binning strategy, where continuous phase values in the range $[0, 2\pi]$ are sampled by using two, three, four or six phase values equally distributed in the same range. Implementing these enhancements will require the design of a new neural network, offering opportunities for changes and improvements to its architecture.

6. Conclusions

The AI-based approach proposed by Carrozzini *et al.* (2025) has been implemented by integrating AI with traditional *ab*

initio phasing techniques. Its first application, the AI-PhaSeed method, is presented here using the neural network originally trained by Larsen *et al.* (2024) on real structures of limited size and $P2_1/c$ symmetry.

We have demonstrated that the performance of the AI-PhaSeed method depends on two key factors: (i) the high extension of the phase seed, defined as the number of seed reflections relative to the total number of observed reflections, and (ii) the good quality of the phase values assigned to those seed reflections. When applied to structures matching both conditions, AI-PhaSeed can achieve optimal performance, in some cases surpassing that of classical DM. Notably, this advantage is observed even at data resolution lower than $1 \text{ \AA}$, indicating that the AI approach has the ability to phase at resolution lower than the atomic level.

We have also developed machine learning tools, such as an RF classification model, to introduce advanced decision-making strategies in the phasing process. Guided by this model, we have developed a first-of-its-kind integration of AI-based tools with classical DM approaches, thereby enhancing the overall efficiency and reliability of crystal structure determination. This study highlights the strong potential of developing faster and more robust AI-based *ab initio* phasing methods, particularly for challenging cases involving low-resolution data, incomplete datasets or entirely novel protein folds.

Our results also suggest that the performance of AI-PhaSeed could be further improved by training a dedicated neural network specifically tailored/optimized to medium- and large-sized structures, incorporating the key features emphasized by the AI-PhaSeed approach. This development is planned for future investigations.

Acknowledgements

The authors thank Dr Claudia Favia and Dr Mauro de Feudis for their help with the data analysis. Open access publishing facilitated by Consiglio Nazionale delle Ricerche, as part of the Wiley–CRUI-CARE agreement.

Data availability

The data supporting the results reported in this article, taken from the Crystallography Open Database, are available upon request from the authors.

Funding information

Financial support by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by the European Union – NextGenerationEU – PNRR, Missione 4 Componente 2 Investimento 1.4 (grant No. CN00000013) to F. Fedele, A. Moliterni and C. Cuocci, MUR PRIN (project 20223B4JWC) (Valorization of carbon oxides by sequential catalysis: combining the reverse water gas shift reaction with catalytic carbonylation for the synthesis of high value added compounds – COXSECAT)

to A. Altomare, and MENDELEEV PRIN (project 2022KMS84P) (Green revolution by merging metal–organic frameworks with deep eutectic solvents for the development of sustainable technologies and artificial nitrogen fixation) to R. Caliandro is acknowledged.

References

- Altomare, A., Cuocci, C., Giacomazzo, C., Moliterni, A., Rizzi, R., Corriero, N. & Falcicchio, A. (2013). *J. Appl. Cryst.* **46**, 1231–1235.
- Breiman, L. (2001). *Mach. Learn.* **45**, 5–32.
- Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., Cuocci, C., Giacomazzo, C., Mallamo, M., Mazzone, A. & Polidori, G. (2015). *J. Appl. Cryst.* **48**, 306–309.
- Burla, M. C., Caliandro, R., Giacomazzo, C. & Polidori, G. (2010). *Acta Cryst.* **A66**, 347–361.
- Carrozzini, B., De Caro, L., Giannini, C., Altomare, A. & Caliandro, R. (2025). *Acta Cryst.* **A81**, 188–201.
- Downs, R. T. & Hall-Wallace, M. (2003). *Am. Mineral.* **88**, 247–250.
- Fawcett, T. (2006). *Pattern Recognit. Lett.* **27**, 861–874.
- Giacomazzo, C. (1998). *Direct phasing in crystallography: fundamentals and applications*. Oxford University Press.
- Giacomazzo, C. (2013). *Phasing in crystallography: a modern perspective*. Oxford University Press.
- Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. F. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.
- Hartigan, J. A. & Wong, M. A. (1979). *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature* **596**, 583–589.
- Karle, J. & Hauptman, H. (1956). *Acta Cryst.* **9**, 635–651.
- Kaufman, L. & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. Chichester: Wiley.
- Larsen, A. S., Rekis, T. & Madsen, A. Ø. (2024). *Science* **385**, 522–528.
- Madsen, A. Ø. (2025). *Acta Cryst.* **A81**, 251–253.
- Oszlányi, G. & Sütő, A. (2011). *Acta Cryst.* **A67**, 284–291.
- Palatinus, L. & Chapuis, G. (2007). *J. Appl. Cryst.* **40**, 786–790.
- Rius, J. (2011). *Acta Cryst.* **A67**, 63–67.
- Rius, J. (2014). *IUCrJ* **1**, 291–304.
- Sheldrick, G. M. (1990). *Acta Cryst.* **A46**, 467–473.