Acta Cryst A
FOUNDATIONS
ADVANCES

ISSN 2053-2733

Phase seeding may provide a gateway to structure solution by deep learning

Anders Østergaard Madsen*

Department of Pharmacy, University of Copenhagen, Copenhagen, Denmark. *Correspondence e-mail: a.madsen@sund.ku.dk

Keywords: phase seeding; artificial intelligence; crystallographic methods.

The phase-seeding method proposed by Carrozzini *et al.* [(2025), *Acta Cryst.* A**81**, 188–201] introduces a strategy for integrating artificial intelligence (AI) with established *ab initio* phasing techniques. Rather than presenting an AI-based phasing solution itself, the authors demonstrate how traditional crystallographic methods can be significantly enhanced if provided with a small subset of approximate phase values – a 'phase seed' – that could, in principle, be generated by a machine learning model. By discretizing phase values into a few angular bins, the method transforms the continuous phase problem into a classification task, thereby reducing the computational burden on AI training. This hybrid approach shows promise for improving structure solution, particularly for large and complex non-centrosymmetric crystals, and opens a pathway for future AI-assisted crystallographic workflows.

1. The phase problem

The crystallographic phase problem – whereby only the amplitudes $|F(\mathbf{h})|$ and not the phases $\phi(\mathbf{h})$ of the complex structure factors $F(\mathbf{h}) = |F(\mathbf{h})| \exp[i\phi(\mathbf{h})]$ of each reflection $\mathbf{h}$ can be retrieved from a diffraction experiment – is perhaps the most fundamental of all crystallographic challenges. It is also one of the most studied, because overcoming the phase problem of a given set of single-crystal X-ray diffraction data allows one to determine the electron density $\rho(\mathbf{r})$ by Fourier summation over all reflections,

$$\rho(\mathbf{r}) = 1/V \sum_{\mathbf{h}} F(\mathbf{h}) \exp(-2\pi i \mathbf{h} \cdot \mathbf{r}), \qquad (1)$$

where V is the volume of the unit cell. From the electron density it is normally possible to infer the atomic positions. However, arriving at the correct set of phases is not a trivial task.

In the early days of modern crystallography, structures were 'guessed' and further refined by Fourier refinement, which consists of correcting the model by inspecting approximate electron-density maps generated using the phases from the previous guess – basically the same technique used in crystal structure refinements today, but without the least-squares method, and without the aid of computers to generate density maps. This cumbersome approach was used until the invention of the Patterson maps (Patterson, 1934), from which heavy-atom positions (if any) could be inferred, and the remaining atoms derived using the Fourier refinement approach.

However, all these approaches were inherently slow, because they required substantial computational efforts in a time where digital computers were not yet available.

In the years after World War II, and in particular in the 1950s and 1960s, computers started to become available. At

https://doi.org/10.1107/S2053273325004814

the same time, new ideas for solving the phase problem were published, in particular the work by Harker & Kasper (1948) who showed that there must exist inequality relationships between some of the **F**'s and others. In other words, it was shown that phases are not entirely lost in the diffraction experiment but can to some degree be deduced from the magnitudes.

This work inspired efforts by Karle & Hauptman (1950), and simultaneously Sayre (1952) to derive approximate and probabilistic mathematical relationships between phases of strong reflections. These relationships are based on basic physical constraints of the electron density in crystals, namely that a physically meaningful density must be positive and consists of atoms, *i.e.* resolved peaks of high density in a volume of lower densities.

The concept of constraining the density to be physically meaningful can also be performed in direct space. This is the idea behind modern *ab initio* phasing approaches, such as the charge-flipping algorithm (Oszlányi & Sütő, 2004), where trial phase-sets are used to produce electron-density maps that are then amended to be physically plausible (*e.g.* negative density is made into positive density). Many modern methods operate iteratively in direct and reciprocal space and are implemented in programs such as *SHELXT* (Sheldrick, 2015) and *SIR2014* (Burla *et al.*, 2015).

What all these modern *ab initio* methods have in common is that they require a level of computational power that was substantial in the middle of the 20th century, but not by today's standards – these highly automated algorithms can often lead to a structure solution within a few seconds given a quite complete dataset at atomic resolution.

In recent years there has been a substantial increase in the computational power and, in particular, development of algorithms and software for performing supervised machine learning (ML). There is a long tradition of exploring ML approaches in crystallography, as demonstrated by the recent 'Machine learning in crystallography' virtual collection of *Acta Cryst.* A (Billinge & Proffen, 2024). It is timely to investigate whether further improvements in phasing can be performed using such tools.

2. Use of ML for phasing

Whereas it requires effort to find the phases and thereby derive an electron density from a set of measured structure factors, it is very easy to generate a set of structure factors – and thus, a synthetic dataset – from a structural model. Thanks to the efforts put in to making algorithms for crystal structure prediction, it is possible to generate millions of realistic, virtual structures, and therefore it is also possible to generate millions of virtual datasets, ideal for supervised ML approaches. Artificial data have shown their usefulness in many of the papers in the virtual collection mentioned above.

In the realm of protein structure determination, Pan *et al.* (2023) have shown how structures of small proteins can be inferred from Patterson maps by the use of a convolutional neural network.

However, since virtual datasets contain both the amplitudes and phases of the structure factors, it is also possible to perform ML for phasing in reciprocal space, as we have recently demonstrated for centrosymmetric crystals with small unit cells (Larsen *et al.*, 2024). The actual phasing algorithm is hidden within the layers of such a deep learning approach and is hard to unravel; however, the capability to phase below atomic resolution – as low as 2 Å – might imply that phase relationships beyond the positivity and atomicity inherent in established *ab initio* approaches could be encoded in the network.

3. Bottlenecks for ML approaches

Two bottlenecks appear in applying ML techniques for phasing.

The first problem is a matter of scalability. At a given resolution, the amount of data in a single-crystal diffraction dataset scales linearly with the size of each unit-cell dimension. This implies that going from small unit cells (10 Å cell dimensions) to medium (20 Å) increases the input to the neural network by a factor of 8. In a network architecture where the full dataset is input, this increase penalizes the training speed and puts high demands on the hardware used during training. In a similar vein, reducing the crystallographic symmetry implies a substantial increase in symmetry-independent reflections.

The second problem is that, in general, the phases can take any value from 0 to 2π, and the general phasing problem is therefore a regression problem. In a centrosymmetric crystal the phases are constrained to either 0 or π, and it is therefore a somewhat less complicated classification problem.

Fortunately, Carrozzini *et al.* (2025) have a good idea that sets a more achievable benchmark for making a successful ML-based phasing approach.

4. The phase-seeding method

In the paper by Carrozzini *et al.* a new idea is proposed: to create a hybrid approach between established *ab initio* phasing techniques and ML. They show that, if a neural network could provide very crude estimates of the phases, it would be possible to use these estimates as 'seeds' for further phase estimation using established dual-space methods.

In practice, Carrozzini *et al.* show that if the phases of a random selection (as small as 10%) of intense reflections of a dataset can be correctly binned into a few regions of phase space, *e.g.* the four quadrants ($0, \pi/2, \pi, 3\pi/4$), it is possible to use these phases as seeds for further phasing in established phasing approaches (electron-density modification and phase-extension procedures), both for small and large unit cells. This approach is seen to be moderately better than classical phasing approaches for the small-molecule cases, and shows a significant increase in solved structures for larger structures; thus, it shows promise in complementing existing techniques.

However, to make this hybrid approach work, there is a need for a type of ML algorithm that can approximate phases on a subset of a dataset. In previous work (Larsen *et al.*, 2024), the Miller indices of a reflection were implicitly encoded by the position of the reflection in the input array. If only a subset of reflections is used, this information must become explicit, which increases the size of the input: there is a trade-off to consider here.

The phase-seeding approach was tested on complete data at atomic resolution. Further work to investigate the potential of applying this approach to low-resolution data or to incomplete data will be very interesting; and most of all it will be interesting to explore ML approaches that can provide the necessary set of 10% approximately correct phases.

The work by Carrozzini and co-workers is inspiring and may indicate that, as the use of ML approaches is steadily increasing in crystallography, we may not see ML algorithms fully replacing existing workflows, but rather have hybrid approaches appearing, where ML ideas can complement and improve established techniques.

References

Billinge, S. J. L. & Proffen, Th. (2024). *Acta Cryst.* A**80**, 139–145.

Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., Cuocci, C., Giacovazzo, C., Mallamo, M., Mazzone, A. & Polidori, G. (2015). *J. Appl. Cryst.* **48**, 306–309.

Carrozzini, B., De Caro, L., Giannini, C., Altomare, A. & Caliandro, R. (2025). *Acta Cryst.* A**81**, 188–201.

Harker, D. & Kasper, J. S. (1948). *Acta Cryst.* **1**, 70–75.

Karle, J. & Hauptman, H. (1950). *Acta Cryst.* **3**, 181–187.

Larsen, A. S., Rekis, T. & Madsen, A. Ø. (2024). *Science* **385**, 522–528.

Oszlányi, G. & Sütő, A. (2004). *Acta Cryst.* A**60**, 134–141.

Pan, T., Jin, S., Miller, M. D., Kyrillidis, A. & Phillips, G. N. (2023). *IUCrJ* **10**, 487–496.

Patterson, A. L. (1934). *Phys. Rev.* **46**, 372–376.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Sheldrick, G. M. (2015). *Acta Cryst.* F**71**, 3–8.