crystals

MDPI

Article

Refining Protein Envelopes with a Transition Region for Enhanced Direct Phasing in Protein Crystallography

Ruijiang Fu [1], Wu-Pei Su [2,*] and Hongxing He [1,*]

[1] Department of Physics, School of Physical Science and Technology, Ningbo University, Ningbo 315211, China; 2011077041@nbu.edu.cn

[2] Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, TX 77204, USA

* Correspondence: wpsu@uh.edu (W.-P.S.); hehongxing@nbu.edu.cn (H.H.)

Abstract: In protein crystallography, the determination of an accurate protein envelope is of paramount importance for *ab initio* phasing of diffraction data. In our previous work, we introduced an approach to ascertain the protein envelope by seeking an optimal cutoff value on a weighted-average density map. In this paper, we present a significant advancement in our approach by focusing on identifying a transition region that demarcates the boundary between the protein and solvent regions, rather than relying solely on a single cutoff value. Within this transition region, we conducted a meticulous search for the protein envelope using a finer map and our proposed transition hybrid input–output (THIO) algorithm. Through this improvement, we achieved a refined protein envelope even when starting from random phases, enabling us to determine protein structures with irregular envelopes and successfully phase crystals with reduced solvent contents. To validate the efficacy of our method, we conducted tests using real diffraction data from five protein crystals, each containing solvent contents ranging from 60% to 65%. Solving these structures through conventional direct methods proved difficult due to the limited solvent content. The mean phase error obtained through our proposed method was about 30°. The reconstructed model matched with the structure in the protein data bank with a root mean square deviation (r.m.s.d.) of about 1 Å. These results serve as compelling evidence that the utilization of the proposed transition region in conjunction with the THIO algorithm contributes significantly to the construction of a reliable protein envelope. This, in turn, becomes indispensable for the direct phasing of protein crystals with lower solvent contents.

Keywords: protein envelope; direct method; *ab initio* phasing; transition; hybrid input–output algorithm; protein crystallography

check for updates

Citation: Fu, R.; Su, W.-P.; He, H. Refining Protein Envelopes with a Transition Region for Enhanced Direct Phasing in Protein Crystallography. *Crystals* **2024**, *14*, 85. https://doi.org/10.3390/cryst14010085

Academic Editors: Vasundara Srinivasan and Santosh Panjikar

Received: 13 December 2023
Revised: 9 January 2024
Accepted: 12 January 2024
Published: 16 January 2024

1. Introduction

Direct methods play a crucial role in solving protein crystal structures directly from diffraction data without relying on any prior information, such as heavy-atom derivatives or homology structures [1–16]. Traditional direct methods for phasing small-molecule crystals [17–29] use the triplet relation and the tangent formula [21–23], the *vive la différence* algorithm (VLD) [28], the electron-density modification (EDM) and the difference electron-density modification (DEDM) [29], etc. There are packages for *ab initio* crystal structure solutions for small and medium molecules using direct methods, such as SHELX [27] and SIR (semi-invariants representation) [30]. The *ab initio* phase retrieval for macromolecule crystals utilizes iterative projection algorithms, such as the hybrid input–output algorithm (HIO) [1] and the difference map algorithm (DM) [5]. HIO is usually used for *ab initio* phase retrieval of macromolecules. It is cyclic and easy to implement with a large convergence radius when the crystal has a high solvent content. While the direct method for macromolecule crystals has been successfully tested in retrieving known crystal structures of proteins and viruses [8–16], it has not been applied to solve the atomic structure of unknown protein crystals.

The term "envelope" in this context refers to a boundary that demarcates the region occupied by the protein from the bulk solvent in the crystal unit cell. Researchers have used various methods to determine a support or envelope for single particles, such as employing a simple geometric shape, manual inspection of a poor electron-density map, utilizing a homologous structure or cryo-EM reconstruction, examining an averaged map, and employing autocorrelation functions, density connectivities and shrinking support techniques [7,31–37]. The envelope plays a crucial role while using direct methods to solve protein crystal structures. Protein and bulk solvent have different density constraints. In order to apply density constraints correctly, a good envelope is required. A well-designed strategy for reconstructing the envelope can limit the search space, effectively reduce degrees of freedom, and speed up phase retrieval.

When dealing with protein crystals, direct methods typically utilize a low-resolution envelope estimate. Previous studies have employed fixed and predefined protein boundaries along with the HIO algorithm to phase protein crystals [8]. Additionally, iterative projection algorithms such as DM and low-resolution envelopes have been employed to retrieve lost phases in protein and virus crystals [9,11,38]. When the protein crystal has a high solvent content, the protein envelope and the structure can be reconstructed from the diffraction amplitudes alone [10,15].

Constructing a good envelope during direct phasing can be challenging, especially when starting from random phases, as the calculated density is nearly random [39]. To address this, we propose a transition approach to refine the reconstructed envelope, thereby improving direct phasing.

Direct phasing of protein crystals demands a balance between two competing constraints: a large bulk solvent and a well-defined envelope. To achieve a unique solution for the phase problem, a large solvent fraction (>65%) is necessary [4,8,10,15,40,41]. Iterative projection algorithms, such as HIO and DM, retrieve the lost phases by modifying the calculated density outside the protein envelope [8,10,11,15]. Therefore, the envelope must encompass all protein residues in the unit cell, and a large and loose envelope is typically employed to cover the complete protein structure.

However, this approach can lead to undesirable effects where the loose protein envelope squeezes the bulk solvent outside the envelope within the unit cell, hindering direct phasing. To overcome this challenge, we propose the use of a transition region, a thin layer expected to contain a well-defined protein envelope, to differentiate between the protein and bulk solvent regions in the crystal unit cell.

The proposed transition region approach involves the construction of the transition region from scratch using a weighted-average density map with a larger averaging radius, starting from random phases and diffraction data. Inside this region, we introduce a transition hybrid input–output (THIO) algorithm to reconstruct and refine the protein envelope on a finer weighted-average density map with a smaller averaging radius. The benefits of the proposed transition region for direct phasing are elaborated in Sections 2 and 3.

In our experiments, we tested the transition HIO method on diffraction data from six protein crystals, discussed in Section 3, five of which had high-resolution data ranging from 1.5 to 1.97 Å, with solvent contents of 60% to 65%. Our results clearly demonstrate that the proposed transition region significantly increases the success rate and speed of direct phasing compared to traditional HIO methods. Additionally, we assessed the transition region's performance with low-resolution diffraction data.

The results presented in this study establish the efficacy of the proposed transition region approach for refining protein envelopes and its potential for enhancing direct phasing in protein crystallography.

2. Methodology

2.1. Direct Phasing Method of Protein Crystallography

Analyzing the uniqueness of the solution is a critical step in solving the phase problem of protein crystallography [4,8,10,40,41]. The unit cell in real space is divided into N grid points, and the goal is to determine the densities at these points. After a fast Fourier transform, N complex structure factors in reciprocal space are obtained, with $N/2$ of them being independent due to the real nature of the densities.

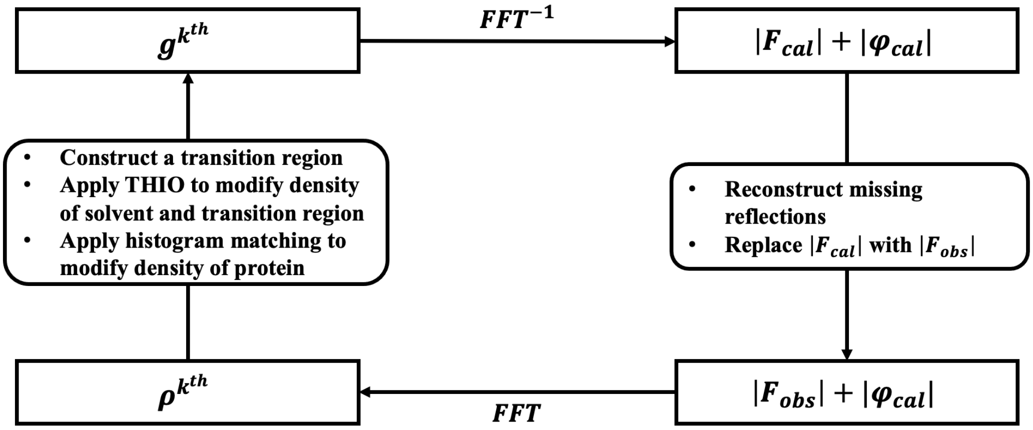
In experimental setups, phases of diffracted beams are lost, and $N/2$ diffraction intensities of the independent structure factors are recorded. This results in an underdetermined phase problem since we need to retrieve N densities from $N/2$ diffraction intensities. To address this issue and ensure a unique solution, density constraints such as high solvent content or non-crystallographic symmetry (NCS) are essential to increase the redundancy of the phase problem. If the bulk solvent in the crystal unit cell occupies more than half of the unit cell, the phase problem becomes overdetermined, but in practice, due to factors like missing reflections, measurement errors, and an inaccurate protein envelope, a high solvent content is still required for successful direct phasing of protein crystals.

To find the unique solution, a direct method with the HIO algorithm is employed [1,8,10,12–14,16]. The process starts from random numbers, and lost phases are iteratively retrieved through thousands of iterations. In each iteration, the calculated density is modified based on density constraints in the protein crystal. A weighted-average density ρ^{avg} is computed from the calculated density using Equation (1).

$$\rho_i^{avg} = \sum_j exp[-d_{ij}^2/(2\sigma^2)]\rho_j \tag{1}$$

where ρ_j is the calculated density at the jth grid point. d_{ij} is the distance between two grid points i and j. σ is a constant which controls the averaging radius. A cutoff value for ρ^{avg} is searched in accordance with the estimated solvent content of the unit cell. This cutoff value is used to distinguish between protein and bulk solvent regions.

The calculated density in the protein region is improved using traditional histogram matching [42], while the density in the solvent region is modified using the HIO algorithm [1] or solvent flattening [43]. The HIO algorithm's negative feedback pushes the solvent density toward zero, leading the calculated density closer to a solution. After density modification, the complex structure factors in reciprocal space are recalculated using an inverse Fourier transform. The observed diffraction data are then incorporated by replacing the calculated magnitudes with the observed values while keeping the calculated phases intact. A flowchart is shown in Figure 1.



$g^{k^{th}}$
FFT^{-1}
$|F_{cal}| + |\varphi_{cal}|$
• Construct a transition region
• Apply THIO to modify density of solvent and transition region
• Apply histogram matching to modify density of protein
• Reconstruct missing reflections
• Replace $|F_{cal}|$ with $|F_{obs}|$
$\rho^{k^{th}}$
FFT
$|F_{obs}| + |\varphi_{cal}|$

Figure 1. Flowchart of the k^{th} iteration of the direct phasing method.

To achieve a solution, several hundred independent runs are usually performed in parallel on a multi-thread server. The iterative cycle involves the computation of four error metrics that characterize the retrieved phases and the calculated structure factors. To avoid overfitting, 5% of the diffraction data are reserved as a free data set, are not involved in phase retrieval, and are used to compute R_{free} using Equation (2). The remaining diffraction data are used to compute R_{work} using Equation (3), which represents the difference between the calculated magnitudes and the diffraction data.

$$R_{free} = \frac{\sum_{\mathbf{h}\in free} ||F_{obs}(\mathbf{h})| - \lambda|F_{cal}(\mathbf{h})||}{\sum_{\mathbf{h}\in free} |F_{obs}(\mathbf{h})|} \tag{2}$$

$$R_{work} = \frac{\sum_{\mathbf{h}\in work} ||F_{obs}(\mathbf{h})| - \lambda|F_{cal}(\mathbf{h})||}{\sum_{\mathbf{h}\in work} |F_{obs}(\mathbf{h})|} \tag{3}$$

where $|F_{obs}(\mathbf{h})|$ and $|F_{cal}(\mathbf{h})|$ are the observed and calculated magnitudes. λ is a scale factor.

Additionally, mean phase error ($\Delta\varphi$) and correlation coefficient (CC) are computed to assess the retrieved phase's quality according to Equations (4) and (5). These metrics are utilized for testing the direct method and are unavailable for unknown structures.

$$\Delta\varphi = \frac{\sum_{\mathbf{h}\in work} arccos\{cos[\varphi_{true}(\mathbf{h}) - \varphi_{cal}(\mathbf{h})]\}}{\sum_{\mathbf{h}\in work} 1} \tag{4}$$

$$CC = \frac{\sum_{\mathbf{h}\in work} |F_{obs}(\mathbf{h})||F_{cal}(\mathbf{h})|cos[\varphi_{true}(\mathbf{h}) - \varphi_{cal}(\mathbf{h})]}{[\sum_{\mathbf{h}\in work} |F_{obs}(\mathbf{h})|^2 \sum_{\mathbf{h}\in work} |F_{cal}(\mathbf{h})|^2]^{1/2}} \tag{5}$$

where $\varphi_{cal}(\mathbf{h})$ are the retrieved phases and $\varphi_{true}(\mathbf{h})$ are the true phases computed from the structure posted in the Protein Data Bank.

The reconstruction of missing reflections resulting from the beam stop [8] and the treatment of observed reflections with significant measurement errors are also important aspects of the process with Equation (6). Furthermore, the free data used to compute R_{free} are also subjected to reconstruction to ensure accurate results.

$$|F_{missing}(\mathbf{h})| = \frac{\sum_{\mathbf{h}\in work} |F_{obs}(\mathbf{h})|}{\sum_{\mathbf{h}\in work} |F_{cal}(\mathbf{h})|} |F_{cal}(\mathbf{h})| \tag{6}$$

2.2. *Introducing a Transition Region to Refine Protein Envelopes*

The direct method employs distinct constraints for protein and solvent regions within a crystal unit cell. The reconstruction of a reliable protein envelope, starting from random phases and diffraction data, is critical for its success. In our previous works [10,12–14,16], we sought a cutoff value on a weighted-average density map based on the estimated solvent content of the crystal. Grid points with a weighted-average density above this cutoff were designated as the protein region, while the remaining points constituted the solvent region. Instead of relying on a single cutoff value, we proposed a transition region. When starting from random phases, the calculated density is almost random. The envelope is reconstructed from the calculated density. The reconstructed envelope deviates from the true envelope, which implies that some protein structure is outside the envelope and some bulk solvent is inside the envelope. The proposed transition region corresponds to that ambiguous zone. Figure 2 illustrates the transition region concept.

The transition region is defined by two cutoff values on the weighted-average density map. Let V_{sol} be the estimated solvent content. By sorting the grid points in the unit cell based on their weighted-average density, the grid points at the bottom $V_{sol} - 5\%$ form the solvent region, while the ones at the top $1 - V_{sol} - 5\%$ form the protein region. The remaining 10% of grid points constitute the transition region. It is important to strike a balance for the size of the transition region. If it is too large, it diminishes the volume occupied by the bulk solvent, which is not preferred for the direct phasing method. On the other hand, if it is too small, it cannot adequately cover and refine the complete protein envelope, rendering it equivalent to a single cutoff value as used in our previous work. In practice, the transition region empirically occupies approximately 10% of the unit cell volume at the outset of phase retrieval. We also tested slightly larger or smaller transition regions, yielding similar results. However, the volume of the transition region is not fixed for all iterations; instead, it linearly shrinks to zero towards the end of the iterations.

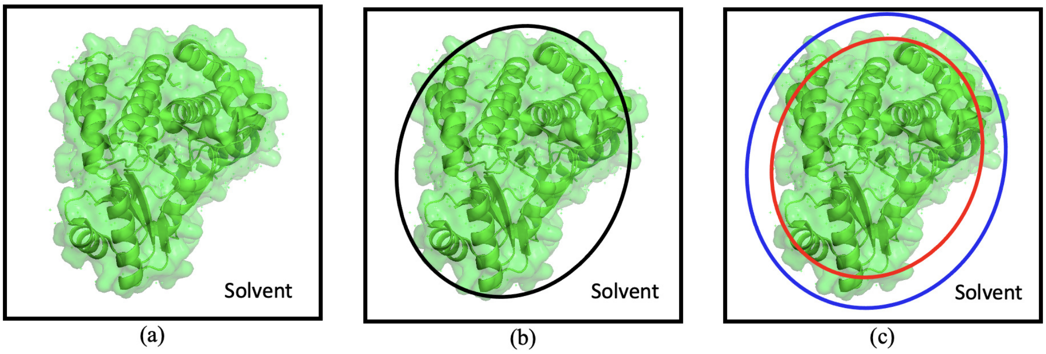


Figure 2. An illustration to show (**a**) the true protein envelope of a molecule in the unit cell, (**b**) the boundary determined by a single cutoff value in HIO direct phasing, and (**c**) the transition region determined by two cutoff values represented by the red and blue circles. The unit cell is represented by the black box. Panel (**b**) shows the boundary determined by a single cutoff value used in HIO direct phasing. The single cutoff value separates the unit cell into protein and solvent regions, but it may not accurately capture the complete protein envelope. In contrast, a more complete picture of the boundary is located within the transition region in panel (**c**). In real calculations, the shape of the transition region is much more complicated than a spherical shell, as shown in this illustrative representation.

Starting from random phases, the calculated density provides limited information about the protein structure. To obtain a broad protein region, a larger averaging radius is employed to compute a weighted-average density map. Using a single cutoff value to discriminate between protein and solvent in this map can lead to the mislabeling of some protein residues as solvents. To avoid the mislabelling, another average density using a smaller radius is computed within the transition region. This density ρ^{avg} is employed to assess the probability α that a grid point is inside the protein. α is defined by Equation (7).

$$\alpha = \frac{\rho^{avg} - \rho^{avg}_{min}}{\rho^{avg}_{max} - \rho^{avg}_{min}} \qquad \textit{in the transition region} \tag{7}$$

where ρ^{avg}_{max} and ρ^{avg}_{min} represent the max and min of ρ^{avg} within the transition region. A grid point with α close to 1 is likely to be located within the protein.

2.3. *Introducing the Transition Hybrid Input–Output Algorithm for Refined Protein Envelope Reconstruction*

Prior to reaching a final solution, the calculated density may not be entirely accurate, leading to a less precise reconstructed protein envelope. Grid points within the transition region exhibit a mixed state of protein and solvent, with a high weighted-average density indicating a higher probability of being proteins and vice versa. HIO primarily modifies the calculated density in the solvent region through negative feedback, leaving the protein region unaffected [1]. As grid points in the transition region exist in a hybrid state, a novel transition hybrid input–output (THIO) algorithm is introduced in Equation (8).

$$g_i^{(k+1)^{th}} = \begin{cases} \rho_i^{k^{th}} & \textit{in the protein region} \\ \alpha_i \rho_i^{k^{th}} + (1-\alpha_i)(g_i^{k^{th}} - \beta \rho_i^{k^{th}}) & \textit{in the transition region} \\ g_i^{k^{th}} - \beta \rho_i^{k^{th}} & \textit{in the solvent region} \end{cases} \tag{8}$$

where $\rho_i^{k^{th}}$ represents the calculated density of the i^{th} grid point in the k^{th} iteration, while $g_i^{k^{th}}$ signifies the modified density of the same grid point in the same iteration. A constant β is utilized, typically set to a value ranging from 0.7 to 0.9. The parameter α_i was defined in Equation (7).

The solvent region provides the constraint needed to solve the phase problem. In the conventional HIO algorithm, due to inaccuracy of the boundary, incorrect density modification is applied near the boundary, effectively weakening the constraint. The introduction of THIO is a remedy for that. It is particularly important for phasing crystals with low solvent contents.

The proposed transition HIO algorithm distinguishes itself from conventional HIO by not solely modifying the density in the solvent region. Instead, it introduces continuity in the modified density on the boundary of protein and solvent. In real protein crystals, the boundary of the protein is diffused rather than sharp. The introduction of the transition region aims to achieve this continuous modification of density.

3. Results

3.1. Enhancing Unit Cell Selection with the Transition Region: Breaking Crystallographic Degeneracy

Protein crystals often possess several equivalent representations due to crystal symmetry, leading to allowed origin translations or origin choices for the unit cell. Additionally, the crystal and its enantiomorph exhibit identical diffraction patterns. For example, in the $P1$ space group, the crystal remains invariant under arbitrary translations, resulting in an infinite number of equivalent origin choices for the unit cell. In the case of crystals in the $P2_12_12_1$ space group, there are eight origin choices, as well as eight enantiomorphs, due to the allowed origin translations. Those equivalent representations sometimes make direct phasing difficult, as the protein envelope lacks the required precision to differentiate between them.

A significant improvement to overcome this challenge is the introduction of the transition region. It proves to be vital in breaking apart crystallographic equivalents, as seen in Figure 3 with pairs of unit cells (a,c) and (b,d). The transition region allows for the refined reconstruction of the protein envelope, which aids in making the final choice of the unit cell. As the phase retrieval progresses, the calculated envelope becomes more accurate, ultimately differentiating between the crystallographic equivalent pairs (a,c) and (b,d). This accurate envelope selection leads to consistent evolution and brings the calculated envelope closer to the true envelope, ensuring a successful solution.

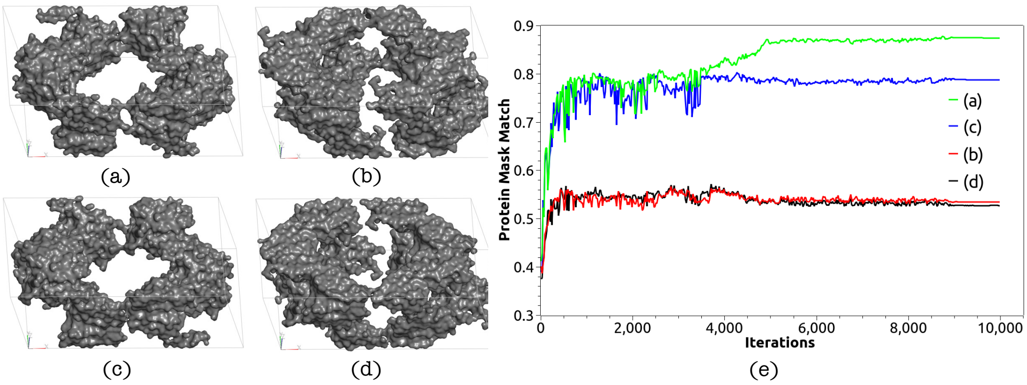


Figure 3. Panels (**a**,**b**) are two different unit cell representations arising from allowed origin translations. The unit cells are computed from the atomic model of 4iqk from the protein data bank. Panels (**c**,**d**) display the enantiomorphs of (**a**,**b**). Notably, there is little difference between representations (**a**,**c**), as well as between (**b**,**d**). Conventional direct phasing with HIO fails to find a solution due to the reconstructed envelope's lack of accuracy in distinguishing (**a**) from (**c**) or (**b**) from (**d**). When a transition region is introduced, direct phasing refines the protein envelope and ultimately converges to representation (**a**), thus resolving the crystallographic degeneracy and achieving a solution, depicted as the green line in panel (**e**).

We used protein 4iqk as an example to illustrate the construction of the transition region. With a data completeness of 99.62% and diffracting at 1.97 Å [44], the crystal has a solvent content (V_{sol}) of 63.77%, set to 63% during phase retrieval. Starting from random phases and diffraction data, the density was computed using a fast Fourier transform. A weighted-average density ρ^{avg} was derived from the calculated density, employing Equation (1), where σ controlled the averaging radius, set at 2.5 Å.

To construct the transition region, two values, v_1 and v_2, were searched for on the weighted-average density map. Shrinkage inward by 5% from the expected protein envelope and expansion outward by 5% defined the transition region. Grid points with $\rho_i^{avg} < v_1$ corresponded to the bulk solvent region, occupying a volume of $V_{sol} - 5\%$, while those with $\rho_i^{avg} > v_2$ corresponded to the protein region, occupying a volume of $1 - V_{sol} - 5\%$. Grid points with $v_1 \leq \rho_i^{avg} \leq v_2$ constituted the transition region, occupying 10% of the unit cell volume. Within this transition region, another weighted-average density ρ^{avg} was computed using a shorter averaging radius ($\sigma = 1.5$ Å). α, calculated using Equation (7), represented the probability of a grid point to be located inside the protein region.

Error metrics were computed in each iteration cycle to monitor the phase retrieval progress. As depicted in Figure 4, R_{free} and R_{work} exhibited an evident drop when a solution was reached. In 100 independent runs starting from random phases but without the transition region, all runs failed to reach a solution due to insufficient precision in reconstructing the protein envelope. However, with the inclusion of the transition region, the calculated protein envelope significantly improved, leading to 7 out of 100 runs converging to a solution. Successful runs were distinguished from failed ones based on R values. The retrieved density map was interpretable, facilitating direct model building. The rebuilt model aligned well with the structure in the protein data bank, with an r.m.s.d. of about 1 Å.

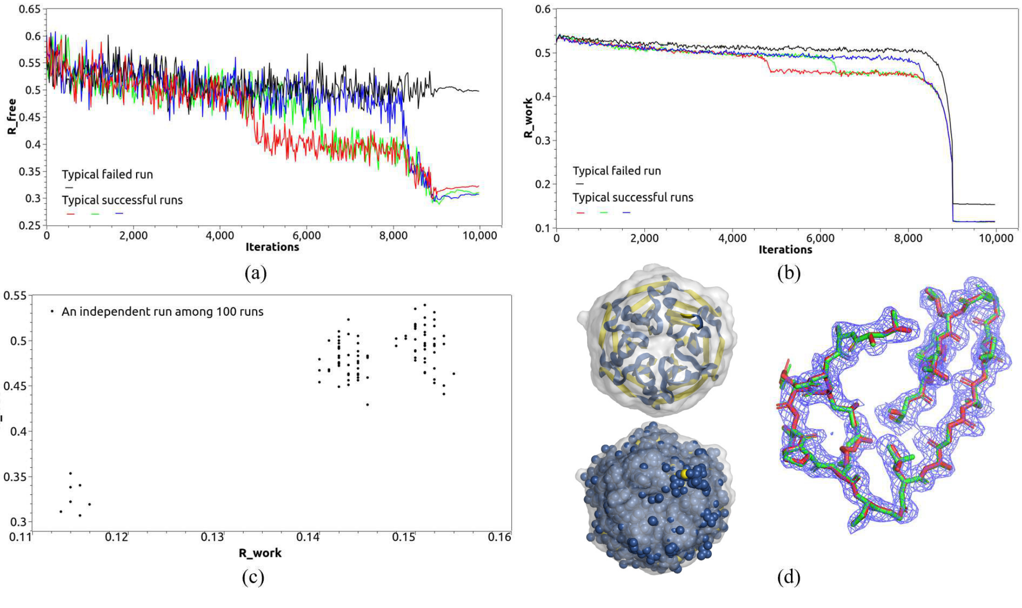


Figure 4. Direct phasing of 4iqk with a transition region. (**a**,**b**) show the evolution of R_{free} and R_{work} during the phase retrieval. A sudden drop in R value indicates a successful run. (**c**) depicts the results of 100 independent runs starting from random phases. The successful runs exhibit small values for both R_{free} and R_{work} simultaneously. (**d**) shows the reconstructed protein envelope and density map of 4iqk. The calculated density map is ready for model building. The rebuilt model is depicted in red, while the model from the protein data bank is displayed in green. The root mean square deviation (r.m.s.d.) between the two models is approximately 1 Å.

Although the structure of 4iqk has non-crystallographical symmetry (NCS), NCS density averaging was not applied here in *ab initio* phasing since the NCS operators were unknown, starting from random phases.

3.2. Enhancing Protein Envelope Accuracy with the Transition Region: Reconstructing Protein Envelopes with Rough Surfaces

The construction of an accurate protein envelope is a critical factor in the success of direct phase retrieval. Typically, the calculated density does not converge to the correct density until the calculated protein envelope covers approximately 90% of the true protein structure. However, achieving an accurate envelope can be challenging, especially for proteins with residues buried deep in bulk solvent after crystallization. In such cases, the surface of the protein envelope within a unit cell appears rough and uneven. Outlier residues buried in the solvent often evade inclusion in the reconstructed protein envelope, leading to failed direct phasing attempts.

Using a single cutoff value on the weighted-average density map to reconstruct the protein envelope with a rough surface in the unit cell proved inadequate. For instance, four protein structures (3tqe [45], 4q82 [46], 2fg0, and 2evr [47], shown in Figure 5) demonstrated outlier residues buried in the solvent after crystallization, resulting in non-smooth surfaces for the true protein envelope within the unit cell. Employing a single cutoff value approach in reconstructing the protein envelope from the calculated density led to a low success rate. In 100 independent runs starting from random phases for each of the four diffraction data sets, only a few successful runs were obtained, as shown in Table 1. The reconstructed envelopes in these failed runs often missed the outlier residues crucial for the accurate representation of the protein structure.

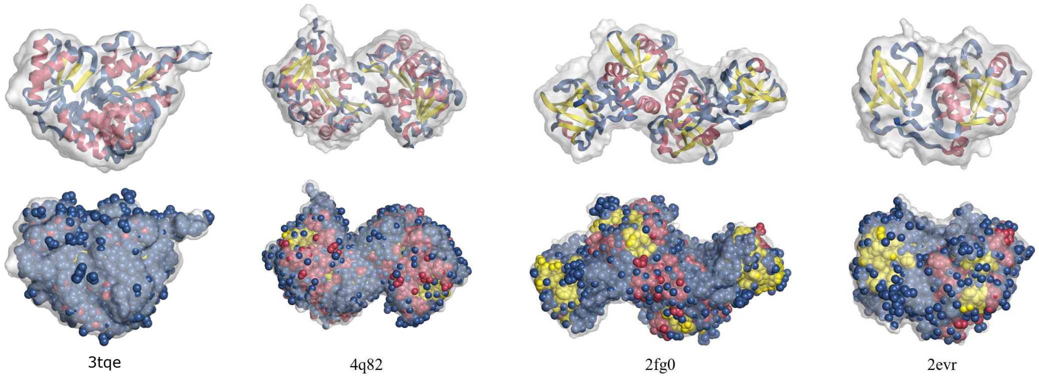


Figure 5. The transition region is important in achieving an accurate protein envelope that encompasses more outlier residues of the protein. The reconstructed envelopes are compared to the true protein structures, presented in both cartoon and sphere representations.

The introduction of a transition region proved to be valuable when dealing with rough surface envelopes. Instead of searching for a single cutoff value on the weighted-average density map, the transition region is identified first, with the expectation that the true protein envelope lies within this region. A smaller σ is used to compute a more detailed weighted-average density based on Equation (1). This detailed weighted-average density aids in identifying an accurate protein envelope that can encompass as many outlier residues as possible. Consequently, the reconstructed protein envelope becomes more accurate, significantly benefiting the phase retrieval process. In our trial calculations, the success rate nearly doubled, as shown in Table 1. The number of successful runs increased significantly, indicating the effectiveness of the transition region in refining the calculated protein envelope, especially when dealing with rough surfaces.

Table 1. The calculated results of protein structures.

				Success Rate (%) **		Speed of Convergence (Iterations) ***		Final $\Delta\varphi$ (°)	
PDB Code	Space Group	Resolution (Å)	Solvent Content (%) *	Without Transition	With Transition	Without Transition	With Transition	Without Transition	With Transition
4iqk	$C121$	1.97	63.77	0	7	-	6565	-	36.2
3tqe	$C121$	1.50	63.07	18	36	6857	4858	32.3	32.5
4q82	$P2_12_12$	1.85	62.49	3	6	6841	6152	29.3	29.4
2fg0	$P4_12_12$	1.79	63.86	3	9	4943	6200	33.3	33.0
2evr	$P4_122$	1.60	61.16	17	19	4626	2660	30.1	30.4
6c4z	$P6_122$	3.30	81.22	11	0	250	-	59.7	-

* The solvent content we used in phase retrieval was computed from the PDB model with CCP4 SFCHECK [48]. ** We conducted 100 runs for each structure, starting from independent random phase sets. When comparing the two cases with and without a transition region, we used the same random phase sets. *** It took about 3 h to complete 10,000 iterations for 100 runs on a Dell R740 server with 52 cores at 2.1 GHz.

As seen in Table 1, we conducted 100 runs for each structure, starting from independent random phase sets. When comparing the two cases with and without a transition region, we used the same random phase sets. The introduction of a transition region generally led to both an increased success rate and improved phasing speed. However, it did not contribute to reducing the final mean phase error, which remained closely associated with the quality of the measured diffraction data.

3.3. Enhancing Direct Phasing with the Transition Region for Protein Crystals with Limited Solvent Contents

Phasing protein crystals with lower solvent contents proved to be a challenge for direct phasing methods that prefer higher solvent contents, typically above 65% [8–16]. As most protein crystals exhibit solvent contents below this threshold, we explored the use of a transition region to maximize the utilization of limited solvent during direct phasing.

The introduction of the transition region played a crucial role in direct phasing, particularly when using the THIO algorithm described in Equation (8). Unlike the conventional HIO method, where a single cutoff value separates the unit cell into protein and solvent regions, the THIO algorithm introduces a transition region that occupies 10% of the unit cell, comprising 5% solvent and 5% protein. This modification ensures that all grid points within the transition region contribute to phase retrieval, making it possible for the 5% protein content to aid in the phasing process. This proves advantageous when dealing with crystals with a solvent content less than 65%.

The evolution of the transition region during iterations is depicted in Figure 6. Initially occupying 10% of the unit cell, the transition region linearly shrunk to zero as the iterations progressed. We tested the THIO algorithm on protein crystals with solvent contents ranging from 60% to 65%, including five crystal structures: 4iqk, 3tqe, 4q82, 2fg0, and 2evr.

The THIO algorithm works well for low-solvent-content protein crystals, particularly when the structure exhibits non-crystallographical symmetry (NCS). NCS-related copies share the same density, reducing the number of unknown variables and increasing the phase problem's redundancy. This overdetermined state proves beneficial for low-solvent-content crystals with NCS, as described in our previous works [14,16]. However, without NCS, both THIO and HIO methods may not work effectively. When NCS is absent, the phase problem of protein crystallography generally becomes underdetermined for solvent contents below 60%.

The retrieved phase typically exhibits a mean phase error of around 30° for high-resolution diffraction data. The calculated density is interpretable, facilitating model building with tools like ARP/*w*ARP [49] or Phenix AutoBuild [50]. The rebuilt models demonstrate a close match with the structures posted in the protein data bank, with an r.m.s.d. of approximately 1 Å, as shown in Figure 7. In the Section 4, we will talk about structures with low-resolution diffraction data, such as 6c4z.

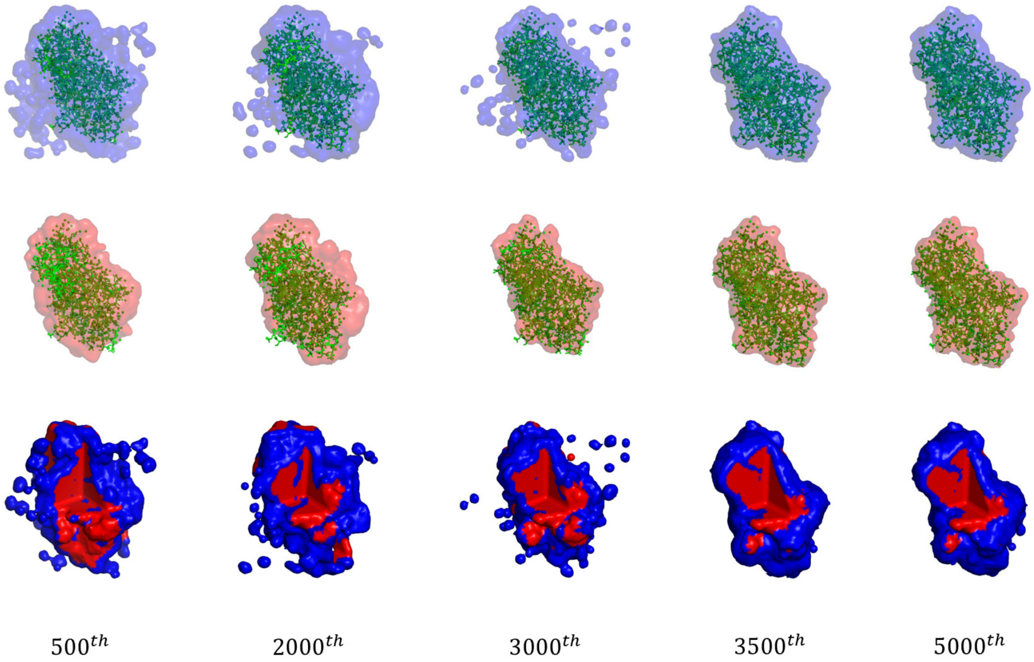


Figure 6. The evolution of the transition region throughout the iteration cycles of a successful run. The first row displays the outer surface of the transition region, highlighted in blue. The second row depicts the inner surface of the transition region, colored in red. The last row represents the transition region, which forms a thin layer located between the red and blue surfaces. To reveal its internal structure, a corner was removed.

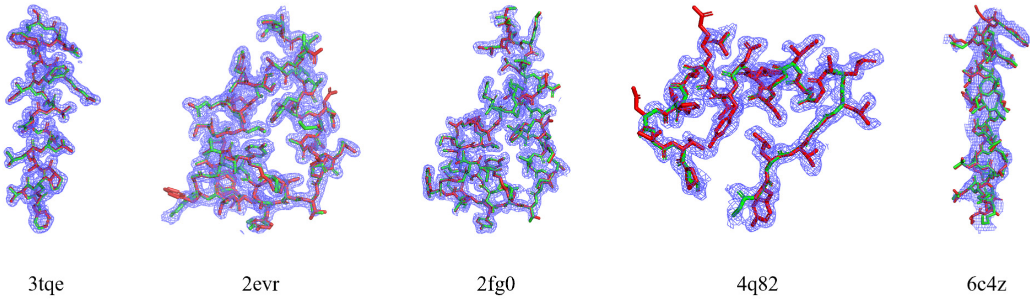


Figure 7. A comparison between the calculated density map (represented by the blue mesh), the rebuilt model (depicted as red sticks), and the true structure from the PDB (shown as green sticks).

4. Discussion

Our approach to finding an accurate protein envelope involved the introduction of a transition region between the protein and solvent within the unit cell. We computed a weighted-average density map from the calculated density, enabling us to identify the inner and outer surfaces of the transition region, which provided an approximate location of the protein envelope. To refine the envelope, we employed a smaller radius to compute a finer weighted-average density within the transition region, and this was utilized by the THIO algorithm. We wondered if using only the smaller radius would yield accurate results, but our trial calculations indicated that the density was fragmented and not interpretable, rendering it unsuitable for determining a complete envelope and bulk solvent. Additionally, experimenting with multiple layers of transition regions did not improve the results.

Our trial calculations primarily focused on high-resolution diffraction data ranging from 1.50 to 1.98 Å. However, we sought to explore whether the transition region could prove beneficial for low-resolution diffraction data as well. We tested the 6c4z crystal, a human-designed amyloid-like structure with a solvent content of 81.22% and diffraction at 3.30 Å [51]. In this case, the transition region failed to aid phase retrieval due to the low-resolution data. The calculated protein density extended significantly into the solvent, blurring the boundary between protein and solvent and making the search for a clear protein envelope impractical. For both HIO and THIO, direct phasing low-resolution diffraction data is still a challenge. Trail calculations on several structures with low-resolution diffraction data failed for both HIO and THIO. A more effective approach to deal with low-resolution data is under research.

Empirically, we set the volume of the transition region to 10% of the unit cell in our trial calculations. We experimented with larger and smaller transition regions but found that 10% was an appropriate choice. During the initial iterations, the calculated density map is nearly random, with grid points having high or low weighted-average density located deep within the protein or bulk solvent. Other grid points remain undetermined, with the probability of being either a protein or solvent. That prompted us to assign a probability (Equation (7)) to such grid points. The volume of the transition region can vary with iteration cycles, and in our tests, it linearly shrunk from a 10% volume at the beginning to complete shutdown at the end of the iterations. A balance was maintained to ensure the transition region was not too large, as it should not overshadow the unit cell, or too small, as it would be indistinguishable from a single cutoff value used in the conventional HIO phasing method.

Comparing our transition method with other related approaches, we observed that the transition region aids in refining the calculated protein envelope, resulting in increased success rates and faster phasing. Liu et al. proposed a block region outside of a fixed protein boundary, setting the density in that region to zero [8]. While that approach ignores the outliers of residues, it does not contribute to envelope refinement. In contrast, our method refines the protein envelope by updating the transition region in each iteration cycle, proving to be more effective.

The transition region not only increases the volume of density modified by phasing algorithms in bulk solvent but also includes density in the transition region itself, making it beneficial for the direct phasing of protein crystals with lower solvent contents. Utilizing as much solvent as possible is crucial for direct methods, as protein folding often results in pockets buried deep within the protein envelope, which can be occupied by solvent molecules. In future research, we plan to employ new algorithms to exploit those small pockets effectively.

The code for our approach is available online [52].

5. Conclusions

In conclusion, our introduction of the transition region and the transition hybrid input–output algorithm proved to be highly effective in refining the calculated protein envelope from random phases and diffraction data. The transition region contributes to the direct phasing method in multiple ways. Firstly, it aids in breaking crystal translation symmetry and determining the origin choice, addressing a critical challenge in phase retrieval. Secondly, it proves to be particularly valuable in reconstructing the protein envelope when dealing with crystals that have a rough surface, ensuring accuracy even in challenging cases. Lastly, the transition region significantly increases the volume of the density modified by iterative projection algorithms, benefiting phase retrieval, especially for protein crystals with lower solvent contents.

Our approach demonstrated remarkable success when applied to high-resolution diffraction data. However, for low-resolution diffraction data, where a distinct protein boundary is absent, the transition region does not provide the same benefits. Nevertheless, the transition region, in conjunction with the transition hybrid input–output algorithm, consistently results in an accurate protein envelope, leading to increased success rates and accelerated phase retrieval. This enhancement makes the direct method more adept at phasing protein crystals with limited solvent contents. The improved performance of our method signifies a promising advancement in the field of protein crystallography, making it more versatile and reliable for various crystal structures.

Author Contributions: H.H. conceived the concepts; R.F. and H.H. carried out the calculations; R.F., H.H. and W.-P.S. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The diffraction data were downloaded from the protein data bank at https://www.rcsb.org.

Acknowledgments: We thank Yuhui Dong, Yi Zhang, and Zhi Geng at the Institute of High Energy Physics for enlightening discussions and hospitality during our visit. This work was supported by the research start-up grant from Ningbo University.

Conflicts of Interest: The authors declare no conflicts of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Fienup, J.R. Phase retrieval algorithms: A comparison. *Appl. Opt.* **1982**, *21*, 2758–2769. [CrossRef] [PubMed]
2. Millane, R.P. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am. A* **1990**, *7*, 394–411. [CrossRef]
3. Komis, G.; Mistrik, M.; Šamajová, O.; Doskočilová, A.; Ovečka, M.; Illés, P.; Šamaj, J. Dynamics and organization of cortical microtubules as revealed by superresolution structured illumination microscopy. *Plant Physiol.* **2014**, *165*, 129–148. [CrossRef]
4. Miao, J.; Sayer, D.; Chapman, H.N. Phase retrieval from the magnitude of the Fourier transforms of non-periodic objects. *J. Opt. Soc. Am.* **1998**, *15*, 1662–1669. [CrossRef]
5. Elser, V. Phase retrieval by iterated projections. *J. Opt. Soc. Am. A* **2003**, *20*, 40–55. [CrossRef] [PubMed]
6. Elser, V. Solution of the crystallographic phase problem by iterated projections. *Acta Cryst. A* **2003**, *59*, 201–209. [CrossRef] [PubMed]
7. Marchesini, S.; He, H.; Chapman, H.N.; Hau-Riege, S.P.; Noy, A.; Howells, M.R.; Weierstall, U.; Spence, J.C.H. X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev. B* **2003**, *68*, 140101. [CrossRef]
8. Liu, Z.C.; Xu, R.; Dong, Y.H. Phase retrieval in protein crystallography. *Acta Cryst. A* **2012**, *68*, 256–265. [CrossRef] [PubMed]
9. Millane, R.P.; Lo, V. L. Iterative projection algorithms in protein crystallography. I. Theory *Acta Cryst. A* **2013**, *69*, 517–527. [CrossRef]
10. He, H.; Su, W.-P. Direct phasing of protein crystals with high solvent content. *Acta Cryst. A* **2015**, *71*, 92–98. [CrossRef]
11. Lo, V.L.; Kingston, R.L.; Millane, R.P. Iterative projection algorithms in protein crystallography. II. Application. *Acta Cryst. A* **2015**, *71*, 451–459. [CrossRef]
12. He, H.; Fang, H.; Miller, M.D.; Phillips, G.N.; Su, W.P. Improving the efficiency of molecular replacement by utilizing a new iterative transform phasing algorithm. *Acta Cryst. A* **2016**, *72*, 539–547. [CrossRef]
13. He, H.; Su, W.-P. Improving the convergence rate of a hybrid input-output phasing algorithm by varying the reflection data weight. *Acta Cryst. A* **2018**, *74*, 36–43. [CrossRef]
14. He, H.; Jiang, M.; Su, W.P. Direct Phasing of Protein Crystals with Non-Crystallographic Symmetry. *Crystals* **2019**, *9*, 55. [CrossRef]
15. Kingston, R.L.; Millane, R.P. A general method for directly phasing diffraction data from high-solvent-content protein crystals. *IUCrJ* **2022**, *9*, 648–665. [CrossRef]
16. Fu, R.; Su, W.P.; He, H. Direct Phasing of Coiled-Coil Protein Crystals. *Crystals* **2022**, *12*, 1674. [CrossRef]
17. Giacovazzo, C.; Siliqi, D.; Ralph, A. The *ab initio* crystal structure solution of proteins by direct methods. I. Feasibility. *Acta Cryst. A* **1994**, *50*, 503–510. [CrossRef]
18. Giacovazzo, C.; Siliqi, D.; Spagna, R. The *ab initio* crystal structure solution of proteins by direct methods. II. The procedure and its first applications. *Acta Cryst. A* **1994**, *50*, 609–621. [CrossRef]
19. Giacovazzo, C.; Siliqi, D.; Zanotti, G. The *ab initio* crystal structure solution of proteins by direct methods. III. The phase extension process. *Acta Cryst. A* **1995**, *51*, 177–188. [CrossRef]
20. Giacovazzo, C.; Siliqi, D.; Gonzalez Platas, J.; Hecht, H.J.; Zanotti, G.; York, B. The *ab initio* crystal structure solution of proteins by direct methods. VI. Complete phasing up to derivative resolution. *Acta Cryst. A* **1996**, *52*, 813–825.
21. Karle, J.; Hauptman, H. A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Cryst.* **1956**, *9*, 635–651. [CrossRef]
22. Sayre, D. The squaring method: A new method for phase determination. *Acta Cryst.* **1952**, *5*, 60–65. [CrossRef]
23. Cochran, W.T. Relations between the phases of structure factors. *Acta Cryst.* **1955**, *8*, 473–478. [CrossRef]
24. White, P.S.; Woolfson, M.M. The application of phase relationships to complex structures. VII. Magic integers. *Acta Cryst. A* **1975**, *31*, 53–56. [CrossRef]
25. Schenk, H. *An Introduction to Direct Methods: The Most Important Phase Relationships and Their Application in Solving the Phase Problem*; University College Cardiff Press: Cardiff, Wales, 1984.

26. Miller, R.; DeTitta, G.T.; Jones, R.; Langs, D. A.; Weeks, C.M.; Hauptman, H. A. On the application of the minimal principle to solve unknown structures. *Science* **1993**, *259*, 1430–1433. [CrossRef]
27. Sheldrick, G.M. A short history of SHELX. *Acta Cryst. A* **2008**, *64*, 112–122. [CrossRef]
28. Burla, M. C.; Giacovazzo, C.; Polidori, G. From a random to the correct structure: The VLD algorithm. *J. Appl. Cryst.* **2010**, *43*, 825–836. [CrossRef]
29. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Mazzone, A.M.; Siliqi, D. EDM–DEDM and protein crystal structure solution. *Acta Cryst. D* **2009**, *65*, 477–484. [CrossRef]
30. 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. *J. Appl. Cryst.* **2015**, *48*, 306–309. [CrossRef]
31. Urzhumtsev, A.G.; Lunin, V.Y.; Luzyanina, T.B. Bounding a Molecule in a Noisy Synthesis. *Acta Cryst. A* **1989**, *45*, 34–39. [CrossRef]
32. Lunin, V.Y.; Urzhumtsev, A.G.; Skovoroda, T.P. Direct low-resolution phasing from electron-density histograms in protein crystallography. *Acta Cryst. A* **1990**, *46*, 540–544. [CrossRef]
33. Rossmann, M.G.; Arnold, E. *International Tables for Crystallography*; Wiley: Chichester, UK, 2006.
34. Fienup, J.R.; Crimmins, T.R.; Holsztynski, W. Reconstruction of the support of an object from the support of its autocorrelation. *J. Opt. Soc. Am.* **1982**, *72*, 610–624. [CrossRef]
35. Crimmins, T.R.; Fienup, J.R.; Thelen, B.J. Improved bounds on object support from autocorrelation support and application to phase retrieval. *J. Opt. Soc. Am. A* **1990**, *7*, 3–13. [CrossRef]
36. Lunin, V.Y.; Lunina, N.L.; Petrova, T.E. Mask-based approach in phasing and restoring of single-particle diffraction data. *Math. Biol. Bioinformatics.* **2020**, *15*, 57–72. [CrossRef]
37. Petrova, T.E.; Lunin, V.Y. Determination of the structure of biological macromolecular particles using X-ray lasers. Achievements and Prospects. *Math. Biol. Bioinform.* **2020**, *15*, 195–234. [CrossRef]
38. Millane, R. P. Phase problems for periodic images: Effects of support and symmetry. *J. Opt. Soc. Am. A* **1993**, *10*, 1037–1045. [CrossRef]
39. Su, W.-P. Retrieving low- and medium-resolution structural features of macromolecules directly from the diffraction intensities—A real-space approach to the X-ray phase problem. *Acta Cryst. A* **2008**, *64*, 625–630. [CrossRef]
40. Millane, R.P.; Elser, V. Reconstruction of an object from its symmetry—Averaged diffraction pattern. *Acta Cryst. A* **2008**, *64*, 273–279. [CrossRef]
41. Millane, R.P.; Arnal, R.D. Uniqueness of the macromolecular crystallographic phase problem. *Acta Cryst. A* **2015**, *71*, 592–598. [CrossRef]
42. Zhang, K.Y.J.; Main, P. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Cryst. A* **1990**, *46*, 41–46. [CrossRef]
43. Wang, B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **1985**, *115*, 90–112. [CrossRef]
44. Marcotte, D.; Zeng, W.; Hus, J.C.; McKenzie, A.; Hession, C.; Jin, P.; Bergeron, C.; Lugovskoy, A.; Enyedy, I.; Cuervo, H.; et al. Small molecules inhibit the interaction of Nrf2 and the Keap1 Kelch domain through a non-covalent mechanism. *Bioorganic Med. Chem.* **2013**, *21*, 4011–4019. [CrossRef]
45. Franklin, M.C.; Cheung, J.; Rudolph, M.J.; Burshteyn, F.; Cassidy, M.; Gary, E.; Hillerich, B.; Yao, Z.K.; Carlier, P.R.; Totrov, M.; et al. Structural genomics for drug design against the pathogen Coxiella burnetii. *Proteins Struct. Funct. Bioinform.* **2015**, *83*, 2124–2136. [CrossRef]
46. Chang, C.; Holowicki, J.; Clancy, S.; Joachimiak, A. Crystal structure of phospholipase/Carboxylesterase from Dyadobacter fermentans DSM 18053. **2012**, *to be published*. [CrossRef]
47. Xu, Q.; Sudek, S.; McMullan, D.; Miller, M.D.; Geierstanger, B.; Jones, D.H.; Krishna, S.S.; Spraggon, G.; Bursalay, B.; Abdubek, P.; et al. Structural basis of murein peptide specificity of a γ-D-glutamyl-L-diamino acid endopeptidase. *Structure* **2009**, *17*, 303–313. [CrossRef]
48. Winn, M.D.; Ballard, C.C.; Cowtan, K.D.; Dodson, E.J.; Emsley, P.; Evans, P.R.; Keegan, R.M.; Krissinel, E.B.; Leslie, A.G.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Cryst. D* **2011**, *67*, 235–242. [CrossRef]
49. Langer, G.; Cohen, S.X.; Lamzin, V.S.; Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **2008**, *3*, 870–875. [CrossRef]
50. Liebschner, D.; Afonine, P.V.; Baker, M.L.; Bunkóczi, G.; Chen, V.B.; Croll, T.I.; Hintze, B.; Hung, L.W.; Jain, S.; McCoy, A.J.; et al. Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in phenix. *Acta Cryst. D* **2019**, *75*, 861–877. [CrossRef]
51. Zhang, S.Q.; Huang, H.; Yang, J.; Kratochvil, H.T.; Lolicato, M.; Liu, Y.; Shu, X.; Liu, L.; DeGrado, W.F. Designed peptides that assemble into cross-α amyloid-like structures. *Nat. Chem. Biol.* **2018**, *14*, 1171–1179. [CrossRef]
52. Refining-Protein-Envelope-with-a-Transition-Region-for-Enhanced-Direct-Phasing. Available online: https://github.com/Ruijiang-Fu/Refining-Protein-Envelope-with-a-Transition-Region-for-Enhanced-Direct-Phasing (accessed on 14 January 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.