International Journal of
Molecular Sciences

MDPI

Article

The Automatic Solution of Macromolecular Crystal Structures via Molecular Replacement Techniques: REMO22 and Its Pipeline

Benedetta Carrozzini, Giovanni Luca Cascarano and Carmelo Giacovazzo *

Istituto di Cristallografia, The National Research Council (CNR), Via G. Amendola 122/o, I-70126 Bari, Italy; benedetta.carrozzini@ic.cnr.it (B.C.); gianluca.cascarano@ic.cnr.it (G.L.C.)
* Correspondence: carmelo.giacovazzo@ic.cnr.it; Tel.: +39-3316202574

Abstract: A description of REMO22, a new molecular replacement program for proteins and nucleic acids, is provided. This program, as with REMO09, can use various types of prior information through appropriate conditional distribution functions. Its efficacy in model searching has been validated through several test cases involving proteins and nucleic acids. Although REMO22 can be configured with different protocols according to user directives, it has been developed primarily as an automated tool for determining the crystal structures of macromolecules. To evaluate REMO22's utility in the current crystallographic environment, its experimental results must be compared favorably with those of the most widely used Molecular Replacement (MR) programs. To accomplish this, we chose two leading tools in the field, PHASER and MOLREP. REMO22, along with MOLREP and PHASER, were included in pipelines that contain two additional steps: phase refinement (SYNERGY) and automated model building (CAB). To evaluate the effectiveness of REMO22, SYNERGY and CAB, we conducted experimental tests on numerous macromolecular structures. The results indicate that REMO22, along with its pipeline REMO22 + SYNERGY + CAB, presents a viable alternative to currently used phasing tools.

Keywords: molecular replacement; proteins; nucleic acids; automated pipeline

Citation: Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C. The Automatic Solution of Macromolecular Crystal Structures via Molecular Replacement Techniques: REMO22 and Its Pipeline. *Int. J. Mol. Sci.* **2023**, *24*, 6070. https://doi.org/10.3390/ijms24076070

Academic Editors: Giuseppe Zanotti and Zhongzhou Chen

Received: 20 January 2023
Revised: 8 March 2023
Accepted: 14 March 2023
Published: 23 March 2023

1. Introduction

The practical solution to the phase problem for small to medium-sized molecules containing up to 300 non-H atoms in the asymmetric unit has been achieved. Several well-documented computer programs that represent this accomplishment include SnB [1,2] SHELX-D [3], ACORN [4], SUPERFLIP [5], SIR2002 [6], SIR2004 [7] (the acronym SIR is associated with *semi-invariant representations* [8,9], which is a general theory that explains the role of structure invariants and semi-invariants in the phasing process using Direct Methods).

Ab initio techniques in the macromolecular field have not achieved the same level of success. To succeed, at least one of the following two strict conditions must be met: atomic or quasi-atomic data resolution for Direct Methods, or the presence of heavy atoms in the unit cell for Patterson Deconvolution Techniques. The largest unknown protein that was solved ab initio by Direct Methods prior to 2006 was cytochrome C3 (PDB code: 1gyo; [10]), with 2003 non-H atoms in the asu, solved by SHELX-D. In 2006, Mooers & Matthews [11] solved the unknown structure of the bacteriophage P22 lysozyme (PDB code: 2anv), which has 2268 non-H atoms in the asymmetric unit, using SIR2002. Patterson deconvolution techniques [12–17] were able to solve large-sized protein structures at non-atomic resolution (e.g., [18], 1e3u, with about 7890 non-H atoms in the asymmetric unit and 1.65 Å of data resolution), and also achieved success with 1buu, a protein with 1283 non-H atoms in the asymmetric unit and 1.92 Å data resolution.

Among the non-ab initio techniques for solving the phase problem in macromolecular crystallography, MR has been the most successful so far, with a higher probability of automatically determining macromolecular structures [19–21]. Researchers have attempted to search in the six-dimensional space, with efforts by Kissinger et al. [22], Jamrog et al. [23], Glykos & Kokkinidis [24], Fujinaga & Read [25], among others. However, some authors have preferred to split the expensive six-dimensional search into two steps, the rotation and translation step, as done by AMoRe [26], BEAST [27], MOLREP [28], PHASER [29], REMO09 [30], and ARCIMBOLDO [31–33].

This paper will only focus on MR approaches to the phase problem. The full automation of the crystal structure solution via MR requires four steps to be completed successfully:

(i) Finding a good enough model. If the sequence identity between the known structure and target is low or limited, the solution of the phase problem may be hindered.
(ii) An efficient MR program to orient and translate the model molecules correctly into the target asymmetric unit. This program must be able to handle cases where the model search is far from optimal, as even well-defined rotation and translation parameters can lead to a large mean phase error.
(iii) The phase extension and refinement step. The MR modulus often produces a large phase error on a limited number of reflections. This step is usually accomplished through electron density modification (EDM) techniques included in large crystallographic packages such as CNS [34], CCP4 [35], SHARP [36], PHENIX [37] and the SHELX series [38]. Burla et al. [39] described a procedure, named SYNERGY, which combines DM by Cowtan [40] with out-of-mainstream techniques such as *free lunch* [41,42], low-density Fourier transform [43], *vive la difference* [44,45], phantom derivative [46,47], and phase driven model refinement [48].
(iv) An automated model building (AMB) program to generate a model that fits the experimental data. Popular AMB programs include BUCCANEER [49] for proteins, NAUTILUS [50] for nucleic acids, ARP/wARP [51] for proteins and nucleic acids, and the PHENIX AUTOBUILD wizard [52] for proteins and nucleic acids. Recently, a new cyclic AMB procedure called CAB [53], which uses BUCCANEER for protein model building and NAUTILUS for nucleic acids building, has been developed and shown to be highly efficient in experimental applications (see Papers I–III [54–56]).

Improving step (i) can greatly enhance the success and automation of MR phasing processes. The success rate of MR is primarily dependent on the root-mean-squared distance between the atomic positions of the template and the target structure. As sequence identity (SI) between the target and the model decreases, the success rate usually increases (although there are no universal cutoffs for SI, SI < 0.30 is generally considered the lower limit for MR success). Despite the increasing number of structures providing good coverage of protein families, the difficulties in this area arise from the non-negligible percentage of proteins sequences without structural homologues. To tackle this issue, automated pipelines have been developed to discover and prepare numerous search models, which can then be processed by MR programs. For instance, for the 1w2y structure, the pipeline MrBUMP by Keegan & Winn [57] selected five models with varying degrees of usefulness.

Point (i) is beyond the scope of this paper, as we will focus solely on steps (ii), (iii) and (iv). Nevertheless, we acknowledge that integrating our techniques with homology detection programs could enhance their potential further. Additionally, the relationship between MR techniques and advanced machine-learning-based structure prediction algorithms, such as AlphaFold [58], will not be covered in this paper. While these methods can predict substantial regions of a protein structure accurately based on its amino acid sequence, experimental data must verify such predictions. MR procedures may play a central role in this area, and a two-way relationship is expected. For instance, the information contained in an MR density map may improve the accuracy of the AlphaFold modelling [59]. Likewise, the predicted models could facilitate a more straightforward application of MR techniques.

This is the fourth paper in a series dedicated to the automatic crystal structure solution of macromolecular structures. Our goal is to achieve, as in the case of small molecules, a

high percentage of practical MR cases solved automatically through a pipeline that requires minimal input, allowing users more time to focus on final model refinement. However, it is important to note that REMO22 is not a completely directive-free program, as directives may be necessary to change, e.g., the MR model or define the estimated number of model copies in the target asymmetric unit. To understand how this paper builds on previous work, we need to revisit Papers I–III. Paper I demonstrated the effectiveness of phase refinement by SYNERGY and the high quality of the CAB automated model building for proteins, while revealing the inadequacy of REMO09 and AMB programs for nucleic acid structures. Papers II and III were dedicated to extending CAB to nucleic acids, and now we return to the MR step to present REMO22, which is an effective successor to REMO09.

To evaluate whether REMO22 can truly be a viable alternative to the most used MR programs, we compared its performance to that of MOLREP (version 11.7.03) and PHASER (version 2.8.3) using the same set of test structures. We chose these programs because they have been used to solve a high percentage of published structures (approximately 61,000 solved by PHASER and 24,000 solved by MOLREP out of over 200,000 structures in the PDB). Additionally, the two programs use different theoretical approaches: PHASER operates in reciprocal space and relies heavily on maximum likelihood techniques, while MOLREP orients model copies using the Patterson space. In contrast, REMO22 employs joint probability distribution function methods.

The phases obtained from any MR procedure may not be of sufficient quality to confirm with certainty that the target structure has been solved. Therefore, it is common practice to refine the MR phases and use them as a starting point for AMB programs. To ensure a thorough and automated phasing process without any unanswered questions, we have developed three pipelines, REMO22 + SYNERGY + CAB, PHASER + SYNERGY + CAB and MOLREP + SYNERGY + CAB, which automate the MR process, phase refinement, and model building. This allows for a more efficient and complete phasing process, providing a more accurate view of the effectiveness of Molecular Replacement techniques for solving macromolecular crystal structures. Furthermore, we will also investigate the role of SYNERGY and CAB in the pipelines to determine whether they contribute significantly to the success of the phasing process or are simply trivial tools for refining phases and building models.

2. Results

In Section 2.1, we present the experimental results that demonstrate the effectiveness of REMO22 in solving MR problems. To assess its performance, we compared its results with those obtained by MOLREP and PHASER on the same set of test structures. In Section 2.2, we discuss the role of SYNERGY and CAB in the REMO22 + SYNERGY + CAB pipeline and compare its effectiveness with that of other pipelines.

2.1. About REMO22

A total of 157 macromolecular structures were used as test cases, comprising 101 proteins and 56 nucleic acids. The PDB codes of the test cases are listed in Table 1, and they are divided into five subsets: PH, PD, PG, DNA and RNA. The first three subsets contain proteins, while the last two contain nucleic acids. Further details can be found in the Section 4.

To keep Table 1 concise, we have not listed the molecular models used in the MR step. However, it is important to note that we used the models that were adopted for the original crystal structure solution, whenever possible. This was done to address the same problems that were encountered during the original solution process. It is possible that better models may be available today, which could make the solution easier. Unfortunately, for 37 of the 46 structures in the PG subset, we were unable to use this approach since they were solved using SAD-MAD techniques. Instead, we utilized search models obtained by Bond [60] by aligning the target and homologue sequence and by using the sequence alignment to trim and mutate the homologous chain with CHAINSAW [61]. To assist

interested readers with their own reviews, all of the models used in this study are included in the Supplementary Materials (Table S1).

Table 1. PDB codes of test structures for MR applications, organized by set: PH, PD, PG, DNA and RNA.

SET	PDB	PDB	PDB	PDB	PDB	PDB	PDB	PDB	PDB	PDB
PH	1a6m	1aki	1bxo	1dy5	1kf4	1kqw	1lat	1lys	1na7	1s31
	1tgx	1tp3	1xyg	1ycn	1yxa	1zs0	2a03	2a46	2a4k	2ah8
	2ayv	2b5o	2f53	2f84	2fc3	2gq3	2h8q	2hyw	2i3p	2iff
	2o3k	2oka	2omt	2otb	2p0g	2pby	2qu5	2sar	6ebx	6rhn
PD	3nng	3npg	3nr6	3o8s	3on5	3q6o	3tx8	3zyt	4e2t	4fqd
	1cgn	1cgo	1e8a	2f8m	5ww0					
PG	1vkf	1vki	1vl2	1vl7	1vlc	2wu6	2x7h	3e49	3gp0	3h9e
	3h9r	3khu	3l23	3llx	3m7a	3mbj	3mcq	3mdo	3mz2	3nyy
	3obi	3oz2	3p94	3ufi	3us5	4e2e	4ef2	4ezg	4fvs	4gbs
	4gcm	4ler	4mru	4ogz	4ouq	4q1v	4q34	4q53	4q6k	4q9a
	4qjr	4qni	4r0k	4rvo	4rwv	4yod				
DNA	1s45	1s47	2b1d	2htt	3ce5	3eil	3gom	3goo	3n4o	3tok
	4gsg	4l24	4ltl	4ms5	4wo3	4xqz	4zym	5cv2	5i4s	5ihd
	5j0e	5ju4	5lj4	5mvt	5nt5	5t4w	5tgp	5ua3	6f3c	6h5r
	6tzq									
RNA	1iha	1lc4	1mwl	1q96	1z7f	2a0p	2fd0	2pn4	3d2v	3fs0
	3owi	3oxm	3s49	3td1	4enc	4jab	5fj0	5kvj	5l4o	5nz6
	5ux3	5uz6	5zeg	6az4	6cab					

The extensive set of test structures listed in Table 1 was used to evaluate the effectiveness of REMO22 in default conditions for both proteins and nucleic acids across a wide range of scenarios. However, this evaluation cannot be considered complete without a comparison to the most popular MR programs available. Indeed, REMO22 should not be considered effective if it succeeds only in cases where another popular MR program succeeds and fails in cases where the same program fails. Therefore, we applied two of the most widely used and effective MR programs, PHASER and MOLREP, to the test structures listed in Table 1. We then compared the results obtained from these programs with those obtained by REMO22. In all our tests, we ensured that the same prior information was provided to all three programs, including measured reflections, space group, unit cell parameters, and MR models for the rotation and translation steps.

Given our focus on automating crystal structure solutions via MR techniques, we chose to apply each program using the default conditions, as suggested by the manuals. We recognize, however, that default procedures may not always be the optimal way to apply the software. Supplementary directives can alter default approaches and potentially increase the chances of a successful crystal structure solution. Despite this, default approaches are widely used by users and are typically the first choice. As the MOLREP default mode includes 20 restrained cycles of REFMAC to reduce the average phase error at the end of the MR step, we decided to add 20 REFMAC cycles to the PHASER automated MR mode. It is important to note that REFMAC cycles are already part of the REMO22 algorithms (see Section 4). Specifically, the PHASER run consisted of seven distinct steps: anisotropy correction, model generation, rotation function, translation function, packing function, rigid-body refinement and restrained REFMAC cycles. For MOLREP, we used its automatic mode, which represents an optimal balance between reducing CPU time and maintaining effectiveness. In this mode, anisotropy correction is not performed (like REMO22), but a packing function is included.

To evaluate the quality of the phases provided by REMO22, PHASER and MOLREP at the end of the MR procedure, some initial remarks are necessary.

Firstly, the procedure for locating copies of the model in the target asu is specific to each program. This includes the estimated number of model copies to locate, the rotation and translation search algorithms, and the FOMs used to rank the solutions. Secondly, a program may choose to simplify the MR techniques to save CPU time, while another program may invest in CPU time-consuming algorithms to improve the quality of the MR models. If two programs simultaneously fail or succeed, the program that saves CPU time is usually the preferred choice. However, if the program that requires more CPU time can solve more MR problems than the faster program, its procedure may be preferred. Therefore, any comparison between programs should consider the computer resources required by each algorithm.

The first figure of merit to evaluate the quality of the MR models, in the absence of any prior information on the target structure, is the final crystallographic residual R, which represents the accuracy of the model and influences the user's trust in it. However, different programs define different resolution limits and subsets of phased reflections, so a fair comparison of R values can only be made when calculated over all observed reflections. Therefore, R values are calculated for each structure using the available MR models. To save time, we do not provide individual R values for each structure in this report, but they can be found in Table S2 of the Supplementary Material section. It should be noted, however, that PHASER stops prematurely in five cases (2htt, 4gsg, 5i4s, 5lj4, 3fs0), when attempting to estimate the number of chains in the target asymmetric unit and produces an error message about the mismatch between composition and unit cell volume. Although user intervention can solve the problem, we treat these cases as failures of the automatic PHASER procedure for statistical purposes. For these cases, we assume an average phase error of 90°, and set the R value to 0.59, the expected R value for acentric random structures.

Table 2 presents a statistical analysis of the R values obtained by REMO22, PHASER, and MOLREP, based on the criteria described above. The final average R values for REMO22, PHASER and MOLREP, denoted as $<R_R>$, $<R_P>$ and $<R_M>$, respectively, were calculated for each subset of test structures.

Table 2. Performance comparison of REMO22, PHASER and MOLREP (the subscripts R, P and M represent the three MR programs) The final average R values (in %), denoted by $<R_R>$, $<R_P>$ and $<R_M>$ respectively, were calculated for each subset of test structures. $NR30_R$, $NR30_P$ and $NR30_M$ are the number of test structures for which the final R value was $\leq$0.30. $N70_R$, $N70_P$, and $N70_M$ are the number of test structures for which the final $<|\Delta\phi|>$ value of $\geq$70°.

SUBSET	$<R_R>$	$<R_P>$	$<R_M>$	$NR30_R$	$NR30_P$	$NR30_M$	$N70_R$	$N70_P$	$N70_M$
PH	30	36	31	24	16	20	2	5	3
PD	42	50	50	1	0	0	7	12	12
PG	35	43	38	14	3	7	6	20	11
DNA	34	45	46	11	3	2	3	10	16
RNA	34	46	42	11	2	7	4	8	8
OVERALL	34	43	40	61	22	36	22	55	50

The overall <R> values for PHASER and MOLREP are close to each other, at 0.43 and 0.40, respectively. In contrast, the <R> values for REMO22 are significantly smaller for each subset of test structures, indicating higher quality MR models and greater user trust in the program. Specifically, the overall <R> value of 0.34 for REMO22 suggests higher model quality compared to PHASER and MOLREP.

Interesting details in Table 2 are the $NR30_R$, $NR30_P$ and $NR30_M$ entries, which indicate the number of cases in which each program produced an R value smaller than 0.30. Such cases represent high-quality MR models that do not require further refinement before being submitted to an AMB program. REMO22 produced models meeting this criterion in 61 cases, while PHASER and MOLREP did so in 22 and 36 cases, respectively.

The high-quality phases produced by REMO22 can also be demonstrated by the average phase errors, $<|\Delta\phi|>_R$, $<|\Delta\phi|>_P$ and $<|\Delta\phi|>_M$, which represent the average deviation of the calculated phases from the published phases at the end of the MR step, for REMO22, PHASER and MOLREP, respectively. As with the R values, each $<|\Delta\phi|>$ value does not refer to the reflection subset actively used in the MR step, due to the different MR resolution limits employed by the three programs. Instead, it relates to all the measured reflections and can therefore be regarded as an absolute, meaningful a posteriori figure of merit. In Figure 1, we present $<|\Delta\phi|>_R$, $<|\Delta\phi|>_P$ and $<|\Delta\phi|>_M$, structure by structure, for each subset of test cases (i.e., PH, PD, PG, DNA and RNA). The structures are arranged in ascending order of $<|\Delta\phi|>_R$ to facilitate readability. For interested readers' numerical reference, we report $<|\Delta\phi|>_R$, $<|\Delta\phi|>_P$ and $<|\Delta\phi>_M$ for each structure in Table S2 of the Supplementary Materials.

Insight into the overall quality of the REMO22, PHASER and MOLREP phases can be gained by examining the global average phase error calculated over all 157 test structures. Table 3 shows that REMO22 has the lowest average phase error, with $<|\Delta\phi|>_R = 45°$, followed by MOLREP with $<|\Delta\phi|>_M = 56°$ and PHASER with $<|\Delta\phi|>_P = 58°$. These values are in good correlation with the corresponding <R> values presented in Table 2.

An additional criterion that may help readers in interpreting the experimental results presented in Figure 1 and in our tables is the use of the following rules of thumb.

If $<|\Delta\phi|>$ is greater than or equal to 70° at the end of the MR step, it is highly likely that the MR model is either misplaced or inaccurate. In such cases, subsequent model refinement is likely to be unsuccessful or result in incomplete or rough models.

On the other hand, if $<|\Delta\phi|>$ is less than 70°, the corresponding model is probably suitable for refinement, and the final AMB programs have a high probability of generating satisfactory structural models.

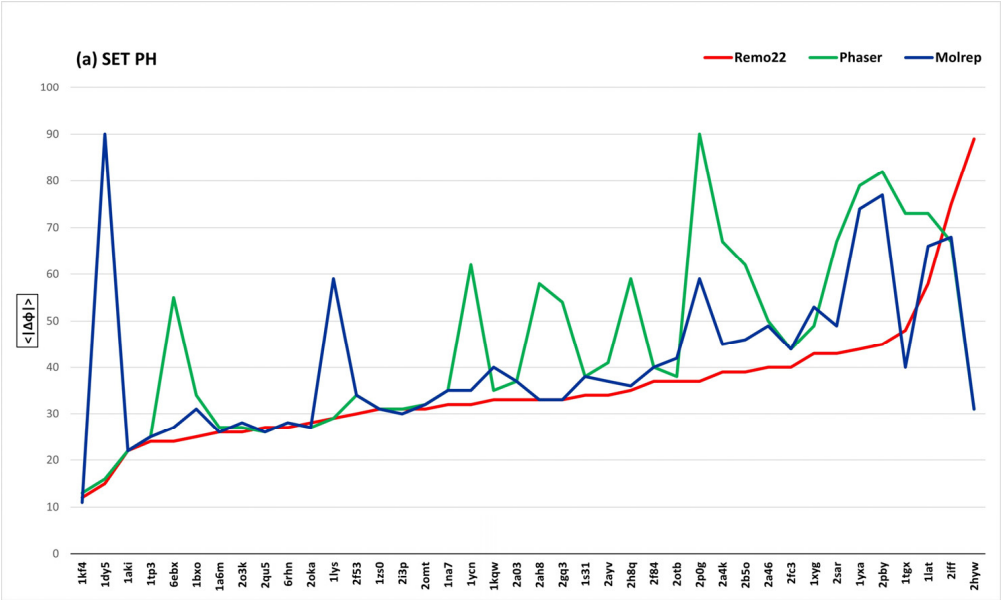

Figure 1. *Cont.*

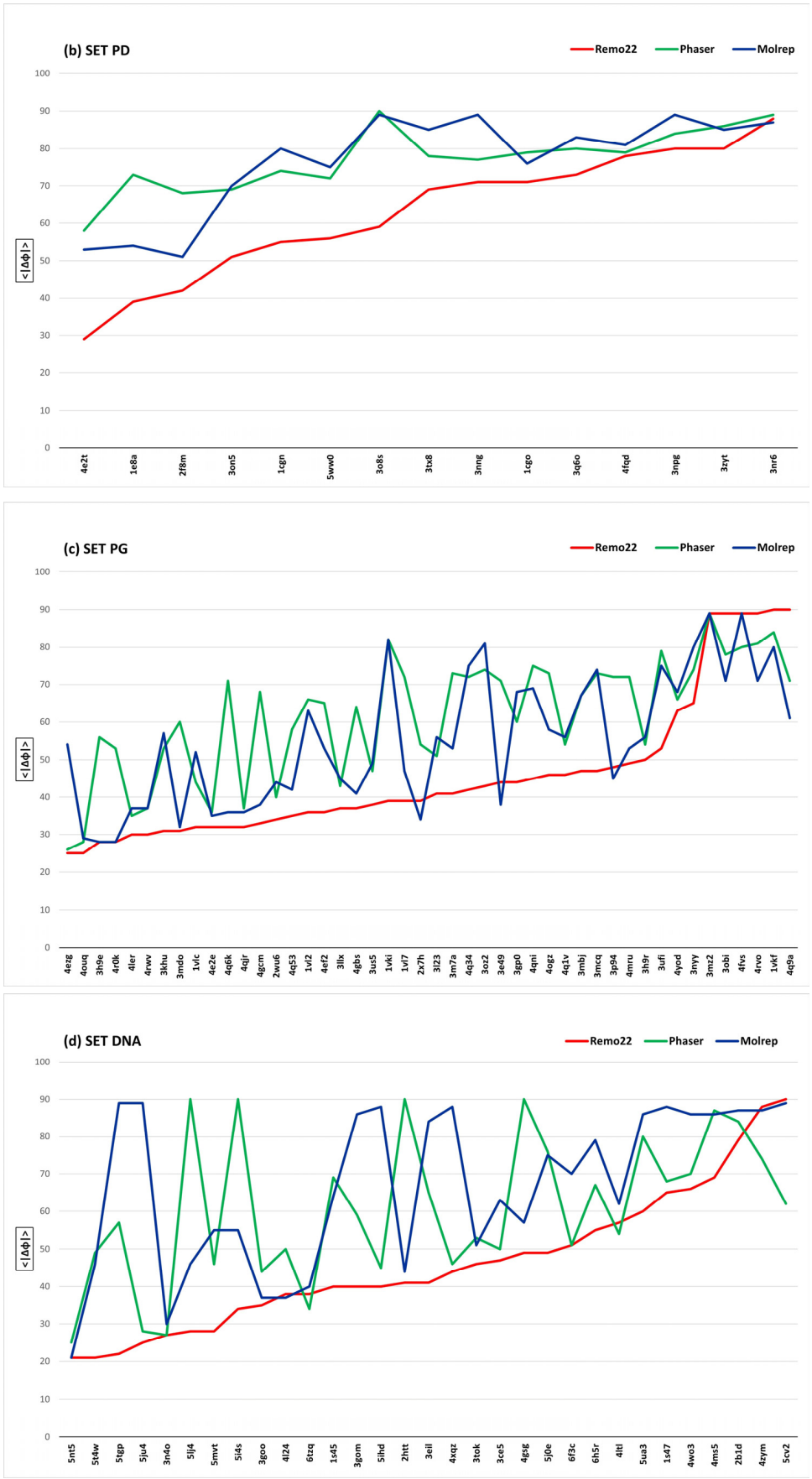

Figure 1. *Cont.*

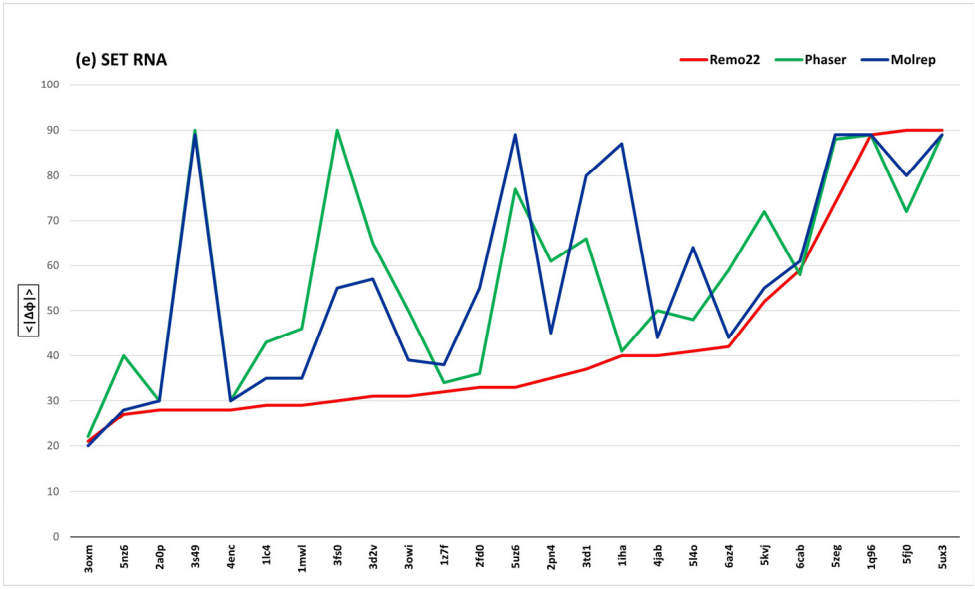

Figure 1. The average phase errors $<|\Delta\phi|>$ (in degrees) obtained by REMO22 ($<|\Delta\phi|>_R$; red line), PHASER ($<|\Delta\phi|>_P$; green line) and MOLREP ($<|\Delta\phi|>_M$; blue line) at the end of the MR steps ((**a**) SET PH; (**b**) SET PD; (**c**) SET PG; (**d**) SET DNA; (**e**) SET RNA). In cases where an MR program declares a failure before the standard ending, we assume $<|\Delta\phi|> = 90^\circ$ (in five DNA-RNA cases for PHASER). The structures are ordered in increasing values of ($<|\Delta\phi|>_R$) for clarity.

Table 3. The results for each pipeline segment are quoted: (i) the global average phase error $\langle|\Delta\phi|\rangle_{MR}$ calculated over all the test structures at the end of the MR step via REMO22, PHASER and MOLREP, and the corresponding $\langle|\Delta\phi|\rangle_{REF}$ calculated after the phase refinement step using either SYNERGY or RESOLVE. All phase errors are in degrees; (ii) the number of proteins for which SYNERGY or RESOLVE improves the MR average phase error by at least 10° (N_P10) and by at least 20° (N_P20); (iii) the number of nucleic acids for which SYNERGY or RESOLVE improves the MR average phase error by at least 10° ($N_{NA}10$) and by at least 20° ($N_{NA}20$).

Pipeline Segment	$\langle\|\Delta\phi\|\rangle_{MR}$	$\langle\|\Delta\phi\|\rangle_{REF}$	N_P10	N_P20	$N_{NA}10$	$N_{NA}20$
REMO22 + SYNERGY	45	41	12	6	0	0
PHASER + SYNERGY	58	53	16	6	17	4
MOLREP + SYNERGY	56	46	44	23	16	10
PHASER + RESOLVE	58	56	1	0	0	0
MOLREP + RESOLVE	56	54	1	0	1	0

The cut-with-ridge criteria mentioned above are not absolute, as the success of model refinement depends on various factors such as data quality (e.g., $|F|/\sigma(|F|)$ statistics, observed data resolution, percentage of solvent, and the effectiveness of the program used for phase refinement). However, using these criteria simplifies the analysis. Table 2 displays the number of test structures for which REMO22, PHASER and MOLREP exhibit a $<|\Delta\phi|> \geq 70^\circ$ ($N70_R$, $N70_P$ and $N70_M$, respectively). It is noteworthy that $N70_R$ is significantly smaller than $N70_P$ and $N70_M$ for each subset of the test structures (22 against 55 and 50, respectively). For proteins, MOLREP seems to be more effective than PHASER, while PHASER appears to be more effective than MOLREP for nucleic acids (18 cases with $<|\Delta\phi|> \geq 70^\circ$ against 24). The correlation of $<|\Delta\phi|>_P$ and $<|\Delta\phi|>_M$ with the $<R_P>$ and $<R_M>$ values presented in Table 2 suggests that the overall qualities of the structural models provided by MOLREP and PHASER are quite similar. Furthermore, the REMO22 structural models are of superior quality compared to those provided by PHASER and MOLREP.

Let us review Table 2 on the structure subsets. The subset PD presents the greatest difficulties due to the small SI values. In this case, REMO22 appears to be more effective than PHASER and MOLREP in limiting the adverse effects of SI, with only 7 cases of

$<|\Delta\phi|> \geq 70^\circ$ compared to 12 cases for PHASER and MOLREP. The PH subset, on the other hand, is generally easy to solve for all programs. However, the MR techniques are less effective for the PG subset than for the PH subset. The difficulty in PG is not due to smaller SI values, but rather to the number of model copies that need to be accommodated in the target asu, which is equal to or greater than 2 for 55% of the PG structures. MOLREP is particularly challenged in nucleic acids, with $N70_M$ corresponding to approximately 43% of the nucleic acid test structures.

It is important to note that the better performance of REMO22 compared to PHASER and MOLREP is primarily due to the implementation of new algorithms (see Section 4) that require larger computer resources. REMO22 is the most demanding program in terms of CPU time, with PHASER and REMO22 requiring approximately 3 min and 4 h, respectively, if the CPU time for MOLREP is set to 1 min. This significant difference in CPU time is due to our decision to include a significant part of the phase refinement process in REMO22, which helps to identify the correct MR solution and also save CPU time in subsequent steps of the crystal structure solution process.

2.2. About the SIR22 Pipeline

As the title and content of this paper suggest, we aimed to develop an automated pipeline for solving crystal structures of macromolecules through MR techniques. However, our analysis of the experimental results obtained using REMO22, PHASER and MOLREP cannot be considered conclusive as the MR models were not subjected to model refinement and AMB, two essential steps in the crystal structure solution process.

To address this, we decided to submit the phases and weights obtained by these programs to the same refinement and AMB procedure using SYNERGY and CAB, respectively. SYNERGY's efficacy was demonstrated by Burla et al. [39], while the ability of CAB was verified in a Paper III by Cascarano & Giacovazzo [56]. We implemented the three pipelines, REMO22 + SYNERGY + CAB, PHASER + SYNERGY + CAB and MOLREP +SYNERGY + CAB, into SIR22, a modified version of SIR2014 [62], for checking the automatic crystal structure solution via different MR techniques. The question we sought to answer was whether the SYNERGY and CAB modules add value to the MR programs or if most of the work was already done at the MR step, making SYNERGY + CAB a trivial bimodule for ending the phasing process.

Let us start with SYNERGY refinement. To simplify the analysis of our experimental results, we need to establish some criteria given the large number of test cases. The first criterion is to compare the average phase error $<|\Delta\phi|>_{MR}$, calculated over all the test structures at the end of the MR step with the corresponding $<|\Delta\phi|>_{REF}$, calculated after the SYNERGY phase refinement (see Table 3). The second criterion focuses on the number of cases where SYNERGY improves the MR average phase error by at least 10° (N_P10 for proteins and $N_{NA}10$ for nucleic acids), or by at least 20° (N_P20 for proteins and $N_{NA}20$ for nucleic acids).

Table 3 summarizes the statistical results for the segments REMO22 + SYNERGY, PHASER + SYNERGY, and MOLREP + SYNERGY, based on various criteria. We observe:

i $<|\Delta\phi|>_{REF}$ is consistently smaller than $<|\Delta\phi|>_{MR}$, irrespective of whether SYNERGY is applied to the REMO22, PHASER, or MOLREP phases.

ii REMO22 + SYNERGY provides the phases with the smallest average error (41°), while PHASER + SYNERGY and MOLREP + SYNERGY have average errors of 53° and 46°, respectively.

iii The effectiveness of SYNERGY varies depending on the MR program. When applied to PHASER phases, SYNERGY provides an average phase improvement of 5°, whereas for MOLREP phases, it provides an improvement of 10°. However, for REMO22 phases, the improvement is only 4°. This is not surprising, as REMO22 phases are already refined phases (with an average phase error of 45°, compared to 58° and 56° for PHASER and MOLREP, respectively), making further refinement more challenging.

iv The number of test structures with a phase error reduction of more than 10° (N_P10, $N_{NA}10$) or 20° (N_P20, $N_{NA}20$) is much higher for the PHASER and MOLREP phases when SYNERGY is applied. Specifically, a reduction of more than 10° is observed for 10% of the test structures for the PHASER phases and 28% of the test structures for the MOLREP phases.

v We note that the larger effectiveness of SYNERGY for MOLREP phases compared to PHASER phases is not completely understood at this point.

It is possible that other refinement programs could yield better results than SYNERGY in improving the MR phases. To further investigate this issue, we decided to apply RESOLVE [63,64] as an alternative refinement program. RESOLVE is a highly respected package based on maximum-likelihood approaches [65–67] that expresses the experimental phase and amplitude information for a given structure factor in terms of a log-likelihood function and calculates the log-likelihood of the resulting electron-density map. Unlike SYNERGY, which employs traditional EDM techniques, RESOLVE assigns more realistic weights to the phases, thereby enhancing their effectiveness. If RESOLVE proves to be more effective than SYNERGY in improving MR phases, it could replace SYNERGY in the SIR22 pipeline, resulting in obvious benefits for the subsequent AMB step.

The results of the combination of PHASER + RESOLVE and MOLREP + RESOLVE are presented in the last two rows of Table 3. The following observations can be made:

i RESOLVE leads to a 2° improvement in the PHASER and MOLREP phases, as compared to the 5° and 10° improvement obtained by SYNERGY, respectively.

ii The values of N_P10, N_P20, $N_{NA}10$, $N_{NA}20$ corresponding to RESOLVE phases are almost always close to zero. This means that RESOLVE is not able to improve the average phase errors by at least 10°, regardless of whether the phases were originally obtained by MOLREP or PHASER.

iii The phases obtained by PHASER + RESOLVE are similar to those obtained by MOLREP + RESOLVE, making them an almost equivalent starting point for the application of the AMB programs.

Based on these observations, SYNERGY seems to be a more promising alternative to RESOLVE. Its significant phase improvements can be even more appreciated if one considers that there are cases in which the tested MR programs are not able to correctly locate the model and there are other cases in which the MR phase errors are already quite small. In both the above cases, it is unrealistic to hope for an improvement in the phase refinement step.

Let us now consider the role of CAB. Its potential was previously discussed in Papers II and III, where it was compared to BUCCANEER, NAUTILUS, ARP/wARP and PHENIX.AUTOBUILD, all run in their default settings. The results showed that the cyclic approach of CAB significantly enhances the effectiveness of BUCCANEER and NAUTILUS, and it is highly competitive with ARP/wARP and PHENIX.AUTOBUILD. With the larger set of protein structures analyzed in this paper, we can perform more meaningful tests.

One algorithm included in the current version of CAB is worth mentioning. In Paper III, we expanded the NAUTILUS library by adding representative structures of the A-DNA, B-DNA, Z-DNA, and four-stranded DNA forms. We also included the MR model because it was selected from structures with the highest sequence identity to the target structure and, by its nature, it deserves to be part of the library. In this version of CAB, we also added the MR model to the BUCCANEER library.

Let us begin by examining the three pipelines: REMO22 + SYNERGY + CAB, PHASER + SYNERGY + CAB and MOLREP + SYNERGY + CAB, to determine their success rate. Table S3 quotes the MA values (MA represents the percentage of non-H atoms within 0.6 Å of the published coordinates) obtained at the end of each pipeline for each test structure. However, for the sake of brevity and clarity, the user may be more interested in a shorter and more comprehensible statistical summary of the results. To accomplish this task, we adopted the following three criteria:

i If 65% or more of non-H atoms are within 0.6 Å of the published coordinates at the end of the CAB procedure, then the automatic crystal structure solution is considered successful. While some readers may find this percentage too lenient, and others too strict, we believe it to be practical, since refinement and completion of the model structure may be easily performed once this percentage is exceeded.

ii If less than or equal to 40% of non-H atoms are within 0.6 Å of the published coordinates at the end of the CAB procedure, then the automatic crystal structure solution fails.

iii Partial success occurs when a percentage smaller than 65% and larger than 40% is obtained.

Table 4 reports the number of structures with MA values lying in each interval (INT_{MA}) for each pipeline.

Table 4. MA denotes the percentage of non-hydrogen atoms within 0.6 Å of the published atomic coordinates, represented by the metric MA. The number of structures (NRSC, NPSC, NMSC, NPRC, NMRC, NRSBN) with MA belonging to each MA interval (INTMA) are shown *.

INT_{MA}	N_{RSC}	N_{PSC}	N_{MSC}	N_{PRC}	N_{MRC}	N_{RSBN}
MA > 65	122	93	108	80	94	98
40 < MA $\leq$ 65	12	12	14	8	11	13
MA $\leq$ 40	23	52	35	69	52	46

* The entries in the table are generated by six pipelines, namely REMO22 + SYNERGY + CAB (N_{RSC}), PHASER + SYNERGY + CAB (N_{PSC}), MOLREP + SYNERGY + CAB (N_{MSC}), PHASER + RESOLVE + CAB (N_{PRC}), MOLREP + RESOLVE + CAB (N_{MRC}) and REMO22 + SYNERGY + (BUCCANEER or NAUTILUS) (N_{RSBN}).

We found that:

i The number of test structures for which the automatic crystal structure solution procedure succeeds, as per the criteria specified earlier, are: 122 for REMO22 + SYNERGY + CAB (N_{RSC}), 93 for PHASER + SYNERGY + CAB (N_{PSC}) and 108 for MOLREP + SYNERGY + CAB (N_{MSC}). The failure cases, as per the same criteria, are 23 for REMO22 + SYNERGY + CAB, 52 for PHASER + SYNERGY + CAB, and 35 for MOLREP + SYNERGY + CAB.

ii MOLREP phases resulted in a smaller number of CAB failures and a larger number of successes compared to PHASER. It is important to note that part of this bias is due to five cases where PHASER stops prematurely while trying to estimate the number of chains in the target asu. User intervention can solve this problem, leading to a statistical improvement in the PHASER results.

The quality of the molecular models provided by PHASER + RESOLVE + CAB (N_{PRC}) and MOLREP + RESOLVE + CAB (N_{MRC}) pipelines was also analyzed. Using RESOLVE instead of SYNERGY as the phase refinement program implies that:

- 13 structures are no longer automatically solved with PHASER data, while the number of failures increased by 17 (compare the columns N_{PSC} and N_{PRC}).
- 14 structures are no longer automatically solved with MOLREP data, while the number of failures increased by 17 (compare the columns N_{MSC} and N_{MRC}).

The results obtained indicate that the use of SYNERGY in the pipeline REMO22 + SYNERGY + CAB is not only effective, but it may also be beneficial in other pipelines that rely on different MR programs. However, it is important to note that the benefits of SYNERGY come at the cost of increased computing resources required by its algorithms.

Furthermore, the individual contribution of CAB to the success of the REMO22 + SYNERGY + CAB pipeline can be assessed by replacing CAB with BUCCANEER for proteins and NAUTILUS for nucleic acids. It is worth noting that CAB essentially uses the same algorithms as BUCCANEER or NAUTILUS, but in a cyclic manner. To obtain the BUCCANEER or NAUTILUS results, the REMO22 + SYNERGY + CAB pipeline was stopped at the first cycle of the CAB procedure (as shown in Table 4). The number of struc-

tures automatically solved by the REMO22 + SYNERGY + (BUCCANEER or NAUTILUS) pipeline is 98 (NRS_{BN} in Table 4), which is 25 less than the number solved by the REMO22 + SYNERGY + CAB pipeline. However, the number of failures increases from 22 to 46. This demonstrates the significant contribution of CAB to the success of the pipeline.

In conclusion, the pipeline REMO22 + SYNERGY + CAB appears to be the most promising option among the tested pipelines. However, it also requires significant computer resources. To enable users of the pipeline REMO22 + SYNERGY + CAB to visually inspect the final structural models, a graphical program (JAV [68]) can be launched by the user. We are planning to automate this step in an upcoming release of SIR22. Figures 2 and 3 show the JAV images of two structures, 3zyt with MA = 0.81, SI = 0.22, and 2i3p with MA = 0.63, SI = 0.99. We superimposed the CAB chains (in red) onto the chains corresponding to the published structures (in blue).

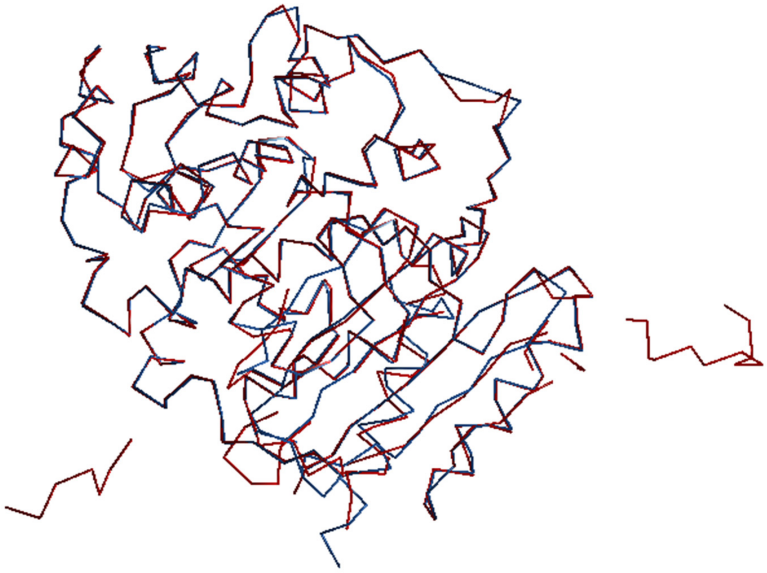

Figure 2. 3zyt, MA = 0.81, SI = 0.22. CAB chain-trace in red, published chain-trace in blue. CAB and the published backbones coincide in most of the target asu.

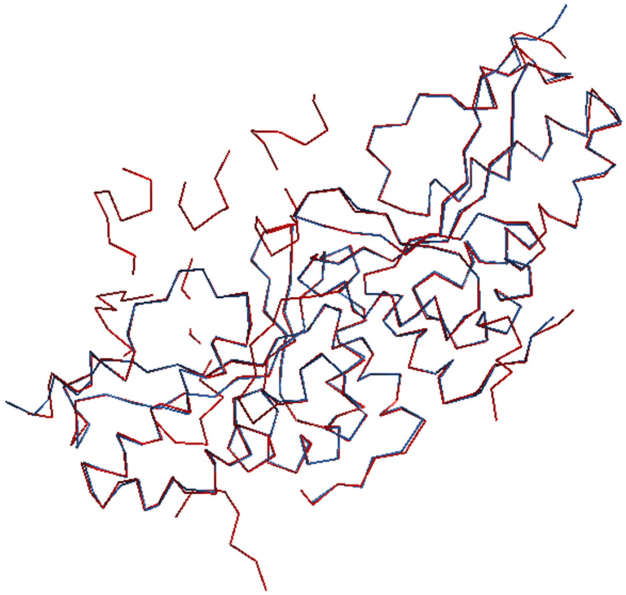

Figure 3. 2i3p, MA = 0.63. CAB chain-trace in red, published chain-trace in blue of the protein component. There are regions of the target asu in which CAB and the published backbones do not coincide.

The two examples presented in Figures 2 and 3 demonstrate that the CAB model can produce good overlap of chains with the published structures in some cases, while in others there may be significant deviations, even with a high SI value (as in the case of 2i3p) when parts of the structure are missing. Additional tests not included in this report indicate that an MA value of 0.65 is a reasonable threshold for a successful automatic crystal structure solution.

3. Discussion

The REMO09 algorithms for the MR step underwent significant modifications, and new algorithms were designed to create REMO22. This program is particularly suitable for the automatic crystal structure solution of biomolecules using MR techniques for both proteins and nucleic acids. To test the usefulness of REMO22 for the crystallographic community, we selected various proteins and nucleic acids and compared REMO22 results with those obtained by using PHASER and MOLREP. We chose the automatic approach recommended by the corresponding manuals for all three programs. The comparison of experimental results clearly indicates that the larger investment in terms of computing resources required by REMO22 is justified by a higher success rate when automatic approaches are used. Therefore, REMO22 can be considered a valuable alternative to the most used MR programs.

REMO22 is the first step in the REMO22 + SYNERGY + CAB pipeline, designed for automatic phasing using MR techniques. To understand the role of SYNERGY, we submitted the MR phases obtained from PHASER and MOLREP to RESOLVE, a popular phase refinement program. The results show that SYNERGY plays a crucial role in the success of automatic phasing procedures. Additionally, we tested the effectiveness of CAB by comparing it with BUCCANEER and NAUTILUS, and found that CAB significantly contributes to the success of the pipeline. Our findings suggest that investing more computer resources into the automatic crystal structure solution using MR techniques has led to the development of the REMO22 + SYNERGY + CAB pipeline, which is a valuable alternative to existing pipelines for solving the phase problem using MR techniques.

The comparison between the pipeline REMO22 + SYNERGY + CAB and the procedures for small-medium size molecules is instructive and raises the question of whether our pipeline can be considered an automatic crystal structure solution procedure similar to those available for small-medium sized molecules. While the main limits for small-medium molecules are the number of non-H atoms per asu and data resolution (300 non-H atoms per asu at 1.1 Å resolution are a hard limit for success unless enough heavy atoms are present), in the pipeline REMO22 + SYNERGY + CAB these parameters are not critical. More critical parameters are the sequence identity between model and target, the number of model copies to accommodate into the target asu, the presence of non-crystallographic symmetry, and the unknown crystal-chemical nature of the target chains.

Let us briefly discuss some of the reasons for failure in the protein structure determination process:

i The SI = 0.3 threshold presents a significant challenge, as evidenced by the fact that 4 out of 10 attempts failed (3nng, 3npg, 3nr6, 3tx8);

ii The inadequacy of the model used for protein complexes containing hetero-oligomers can lead to failure. For instance, the 1lat structure comprises two polypeptide chains of 71 and 74 residues, respectively, as well as two identical nucleic acid chains, each with 19 nucleotides. However, the model only corresponds to the polypeptide chains of the 1glu structure. Similarly, in the case of the 2iff structure, which is a complex of a monoclonal antibody (two chains of 212 and 214 residues), and a lysozyme (one chain of 129 residues), the model only contains the lysozyme chain of the 1hem structure. Even if the models are correctly positioned, recovering the full structure for these cases is a challenge;

iii Inaccurate or incomplete prior information on the crystal-chemical nature of the target can also contribute to failure. For instance, DNA molecules are flexible and can adopt

various structures, including G-quadruplex structures formed by nucleic acids rich in guanine. These structures are helical in shape but may be challenging to locate if the model is not a four-stranded DNA structure. Examples of G-quadruplex structures include 1s45, 1s47, 4wo3, and 5ua3;

iv Disorder can also pose a challenge in determining protein structures. For example, in the cases of 3tok and 4gsg, each chain exhibits two distinct configurations, with most of the phosphorus atoms being common to both configurations. The relatively small MA values (0.45 and 0.41, respectively) are calculated with respect to the total number of atoms in the asymmetric unit, including the disordered pairs.

All of the aforementioned reasons, combined with the inherent statistical limitations of MR FOMs, caused REMO22 to completely fail in 23 cases. Despite this, REMO22 + SYNERGY + CAB must still be considered as a reliable and effective automated pipeline for solving crystal structures using MR techniques, as evidenced by its high success rate (122 out of 157). However, one significant drawback of the pipeline is its high CPU time requirement, which can be attributed to our implementation of new algorithms and the insufficient attention paid to computing times when connecting the three segments of the pipeline. Nonetheless, we are actively working on ways to significantly reduce the CPU time requirement in the near future.

4. Material and Methods

Burla et al. [39] used 24 protein structures out of 157 test cases to evaluate the SYNERGY refinement process of the phases obtained by REMO09. To increase the size of the test sample, this set was expanded to 40 (SET PH). The SET PD comprises 10 of the 13 structures investigated by DiMaio et al. [69] (which have experimental data available), characterized by an SI value smaller than 0.30. These structures were originally solved by combining PHENIX with ROSETTA, a suite [70] that uses physically realistic all-atom potential functions for predicting protein structures based on their amino-acid sequence. One of these structures (4e2t) has an SI of 1 and was used by DiMaio et al. to verify the method. Four test structures from SET PH (1cgn, 1cgo, 1e8a, 2f8m), for which SI < 0.40, were moved to SET PD. Additionally, we included 5ww0 in SET PD, a structure that was originally solved by a working version of REMO22 and has an SI of 0.23.

The SET PG consists of the remaining 46 protein test structures, which were deposited in the PDB by the Joint Centre for Structural Genomics, Wilson Laboratory, Scripps Institute. These structures are commonly used as a test case for MR studies.

For the nucleic acid structures, we selected 56 structures deposited in the PDB database (solved using MR techniques), thereby having observed diffraction data, unit cell information, space group symmetry, published sequences, and MR models available. Among these, 46 were used by Cascarano & Giacovazzo [56] as test cases to assess the effectiveness of the CAB approach for nucleic acid structures. The first 31 structures are DNA (SET DNA), and the remaining 25 structures are RNA fragments (SET RNA).

REMO09 utilized the method of joint probability distribution functions, which was adapted to different types of prior information. The same approach is maintained in REMO22, but several new algorithms have been incorporated to enhance the program's robustness.

4.1. Extension to Nucleic Acids

REMO09 was originally designed to work only with protein structures. REMO22 has been extended to work with both DNA and RNA structures.

4.2. Estimation of the Number of Chains Per Asu and of the Number of Model Copies for MR

The current technique for estimating the number of chains in the target asymmetric unit (asu) is based on biochemical analysis, which establishes the size and sequence of macromolecular chains present in the target crystal structure. However, the actual number of chains per target asu is unknown. The most popular technique for estimating this

number is the Matthews method [71], which is occasionally supplemented by considerations by Kantardjieff & Rupp [72], who found a correlation between solvent content and diffraction limits.

The Matthews method assumes implicitly that the protein chains in the unit cell have the same size and that the density of the protein (δ_{prot}) is usually around 1.35 g/cm^3, which is independent of the protein's nature and molecular weight [73]. However, these assumptions are not always valid in practice. Fischer et al. [74] conducted tests that suggest that δ_{prot} = 1.41 g/cm^3 is a suitable estimate for proteins with high molecular weight (i.e., M > 30 kDa). However, the protein density increases with decreasing molecular weight and reaches its maximum value of δ_{prot} = 1.50 g/cm^3 for the smallest proteins (i.e., M ≈ 7 kDa).

Matthews' survey of 116 different proteins suggests that the protein typically occupies 57% of the crystal volume, with occupancy values ranging from 75% to 35%. While the Matthews method works well in many cases, it can lead to ambiguity, especially for higher assembly numbers. A popular criterion for estimating the number of chains per target asymmetric unit (NCHT) is to choose the value that makes the protein volume fraction (PROTFRAC) closest to 0.50, a value estimated heuristically based on a large number of observations. This criterion is commonly used in PHASER, among other software tools.

While the early estimation of the target composition is not crucial for the success of MR, a more accurate early estimate can be beneficial, particularly when using an automatic approach. In REMO22, an algorithm is used to estimate the number of chains per asu and the number of model copies to accommodate in the target asu. The algorithm involves the following steps:

i In small molecule crystallography, the expected number of molecules per asu is based on the volume per non-H atom (VOLAT), which is usually assumed to be between 16 and 18 Å^3. For macromolecules, the sizes and sequences of the molecular chains present in a target crystal are typically known beforehand. However, the volume of the surrounding solvent remains unknown, making it challenging to estimate the number of chains per target asu. We have modified this rule based on a survey of a wide range of proteins and DNA-RNA structures. For proteins, the expected number of chains per target asu (NCHT) is that for which VOLAT is closest to 38 Å^3, and not smaller than 22 Å^3. For DNA structures, NCHT is that for which VOLAT is closest to 34.5 Å^3, and not smaller than 22 Å^3. For RNA structures, NCHT is that for which VOLAT is closest to 44 Å^3, and not smaller than 22 Å^3. The numerical values were established empirically.

ii The second step of the algorithm is aimed at estimating the number of model copies to accommodate in the target asu (NMOD). This information is typically sought after by the MR user. While not critical for the success of the MR procedure, a good early estimate of NMOD can simplify the automatic approach. Furthermore, this step can correct any incorrect NCHT estimate made in the first step of the algorithm. In cases where the model includes n identical chains, the NCHT value needs to be searched among multiples of n. However, there are scenarios where the target composition is made up of NCHT copies of two different sequence chains (one large and one small), while the model comprises only a single large chain. In such cases, confirming the experimental NCHT value is clearly incorrect, while NCHT/2 is a more accurate choice. Our algorithm can identify and address such situations, especially when the size of the smaller chain is insignificant compared to the larger chain (for example, less than 50% of the long chain). In such cases, the smaller chains are disregarded. The algorithm is designed to be flexible and can be applied to situations where the model and/or target consist of copies of chains of varying sizes. To assess the effectiveness of the choices mentioned above, we compared the number of incorrect estimates using the PROTFRAC criterion (50% solvent) versus the VOLAT criterion. Out of a total of 157 test cases, we discovered that the PROTFRAC criterion led to 30 erroneous NCHT estimates, whereas the VOLAT criterion resulted in only 15 incorrect estimates. These findings provide a promising foundation for the complete automation of the

MR procedure. In addition, the NMOD value can be rectified in the third step of the algorithm, as described in the main text.

iii During the third step, it is possible to correct the number of model chains to be placed in the target asu through post-estimation. Let us assume that the orientation and location of the nth model have already been determined by the MR procedure, and that the figure of merit (FOM_n) has been calculated to assess the reliability of the model's position and orientation. The FOM_n value is expected to increase with the accuracy of the model, which corresponds to the number of accurately located model copies. If FOM_{n+1} is found to be less than FOM_n, then the (n + 1)th copy of the model is rejected, the MR procedure is stopped, and the phase refinement step is started.

To evaluate the arrangement of the located chains, including symmetry-related copies, a second figure of merit, $CLASH_n$, is calculated. For proteins, $CLASH_n$ estimates the fraction of Cα atoms that overlap (within 3.0 Å) once the nth model has been located. For nucleic acids, it estimates the overlapping fraction of the phosphate and C atoms in the ribose-phosphate backbone and the N atoms of the bases.

Suppose we are assessing whether the (n + 1)th model copy should be accepted after the rotation and translation step. In that case, R(n) represents the crystallographic R-factor corresponding to the n located and accepted model copies, while R(n + 1) corresponds to the value related to the (n + 1) located copies. If $R(n) - R(n + 1) > 0.02$, the clash FOM is not checked and the (n + 1)th model copy is accepted. If $CLASH_{n+1} > 35\%$ or

$$R(n + 1) - R(n) > 0.15 \quad (1)$$

then the (n + 1)th model copy is rejected.

The meaning of the above conditions is clear. However, we have a supplementary condition: if

$$[R(n) - R(n + 1)]/CLASH_{n+1} > 0.10 \quad (2)$$

the (n + 1)th model copy is accepted, otherwise, it is excluded.

Let us examine the purposes of Conditions (1) and (2). If the (n + 1)th model copy is incorrectly oriented and/or located, Equation (1) is expected to be satisfied, and the rejection of the (n + 1)th model copy is warranted. In cases where $R(n + 1) - R(n)$ is positive but very small, and $CLASH_{n+1}$ is sufficiently large, it may be risky to include the (n + 1)th model copy in the current model. Conversely, if $CLASH_{n+1}$ is very small, and $R(n + 1) - R(n)$ is also sufficiently small to meet Condition (2), accepting the (n + 1)th model copy appears to be a reasonable decision. To avoid numerical divergence in Equation (2), we consider a CLASH value below 0.10 to be insignificant. Therefore, if CLASH < 0.05, we set CLASH to 0.05 in Equation (2). This algorithm is applied identically to both proteins and nucleic acids.

4.3. Resolution Limits

The subsets of reflections used in the rotation and translation steps are chosen automatically. Reflections with a resolution of up to 7 Å are excluded from calculations, except in situations where SI is less than 0.5. The maximum accepted resolution for active reflections is 2.5 Å, and reflections with very high or very low normalized structure factor moduli are also disregarded. The SI value is not considered for nucleic acids, mainly because nucleic acid helices can assume comparable conformations, even when their sequences are substantially different.

4.4. Search Algorithm for the Rotation Step

The orientation space is based on the asymmetric region of the rotation group [75]. First, the atomic coordinates of the model are orthonormalized, and the maximum molecular dimension is calculated. Then, an orthogonal reciprocal lattice grid is generated, with the direct space dimensions chosen to be four times the maximum molecular dimension.

The model is rotated by rotating the observed reciprocal lattice with respect to the model lattice, and the structure factors of the molecular model are calculated only once.

To rotate the model, an angular grid dθ is used, with n_1dθ, n_2dθ, n_3dθ being the Euler angles corresponding to the cubic primitive lattice. There is at least one point in the unit cell of such a lattice that is approximately 0.87dθ from the lattice points (i.e., the center of the cubic cell). To reduce sampling errors, the angular grid can be lowered to dθ/2, but this results in eight times more lattice points. An alternative approach is to explore the orientation space using a body-centered lattice, which doubles the number of lattice points but ensures that no point in the body-centered cubic cell is farther than 0.56dθ from any lattice point. The body-centered cubic lattice is obtained by first exploring the orientation space using a primitive lattice and then exploring the same angular space using the same primitive cubic lattice, but starting from (dθ/2, dθ/2, dθ/2).

4.5. Anisotropy Correction

Anisotropy in diffraction data refers to the fact that diffraction intensities decrease at different rates in different directions of the reciprocal lattice. As a result, the FOM criteria used to select the correct solution in MR may fail. The reason for this is that the observed diffraction intensities are often anisotropic, while the calculated intensities, particularly in the early stages of the process, are usually assumed to be isotropic. To overcome this limitation, it is necessary to make the calculated and observed structure factors thermally homogeneous. This can be achieved by renormalizing the normalized structure factors according to their direction before calculating the FOMs. To estimate the degree of anisotropy, one can examine how the overall principal components of the anisotropic atomic displacement parameters vary in different directions of reciprocal space. In REMO22, a mathematical approach based on previous work on the preferred orientation of crystallites in a powder [76] is applied to account for anisotropy in the diffraction data.

Let us consider a scenario where the normalized structure factor moduli, $|E|$, have been calculated, and n reciprocal lattice points (with n being approximately 30) have been selected, which correspond to n directions [**h**] = [hkl]. If these points are chosen at very low resolution (e.g., [100], [010], [001], [110], [101], etc.), they will represent all directions in reciprocal space and will be referred to as polar directions. For each polar direction [**h**], the following steps are executed:

(1) The reciprocal space is divided into cones, all with the same axis as the polar direction. The cones are arranged so that each one is fully contained in the next. The shells (i.e., the regions of reciprocal space between adjacent cones) have approximately equal volumes and therefore contain approximately the same number of lattice points. For each shell, α is the average angle (**k**, **h**), where **k** is the generic lattice point in the shell.

(2) For each shell, $<|E_{\mathbf{k}}|^2>$ is calculated, and the corresponding values are plotted against α.

(3) The von Mises distribution

$$M = \exp(G \cos 2\alpha)$$

is found, where G is the parameter best fitting the experimental $<|E_{\mathbf{k}}|^2>$ distribution. If G is large and positive, then $<|E_{\mathbf{k}}|^2> > 1$ along the **h** direction, if G is large and negative then $<|E_{\mathbf{k}}|^2> < 1$ along the **h** direction.

Assuming that steps 1–3 have been applied to all n polar axes, if the values of all the G's are close to zero, then the reciprocal space is nearly isotropic. However, if some G's (either positive or negative) are significantly large, then the reciprocal space is mainly anisotropic.

To correct for anisotropy, one can easily represent the overall anisotropy of the reciprocal space with an ellipsoid. The geometrical shape of the ellipsoid depends on the crystal system being studied: it is spherical for the cubic system, a two-axis ellipsoid for the trigonal, hexagonal and tetragonal systems, and a three-axis ellipsoid for the orthorhombic, monoclinic and triclinic systems.

The orientation of an ellipsoid in reciprocal space is influenced by its underlying symmetry. In trigonal, hexagonal, and tetragonal systems, one of the two ellipsoid axes must align with **c***. In contrast, the orthorhombic system requires all three ellipsoid axes to be parallel to **a***, **b*** and **c***. In monoclinic systems, one of the three ellipsoid axes aligns with the unique two-fold axis **b***. In the absence of symmetry constraints, the ellipsoid orientation in triclinic systems is not predetermined.

To illustrate why these constraints exist, consider the orthorhombic system. The directions **a***, **b*** and **c*** are unrelated by symmetry elements, and the ellipsoid must have three axes to account for all possible anisotropy values against the crystal symmetry. To avoid discrepancies, the three ellipsoid axes necessarily align with **a***, **b*** and **c*** because otherwise the [hkl], [-h-kl], [-hk-l], [h-k-l] directions should have different anisotropy values with respect to the crystal symmetry.

Let us examine how to correct the anisotropy of the reciprocal space. If the G value is sufficiently large for certain polar directions, a correction parameter *O* can be calculated for each reflection, considering the crystal symmetry.

For hexagonal-trigonal systems:

$$O(\mathrm{hkl}) = \langle E^2\rangle_{[100]}(\cos^2\vartheta_1 + \cos^2\vartheta_2) + \langle E^2\rangle_{[001]}\cos^2\vartheta_3$$

where ϑ_1 is the angle between the direction [hkl] and the direction [100], $\vartheta 2$ is the angle between [hkl] and [1−20], $\vartheta 3$ is the angle between [hkl] and [001]. E denotes the normalized structure factor corresponding to *F*.

For the tetragonal system:

$$O(\mathrm{hkl}) = \langle E^2\rangle_{[100]}(\cos^2\vartheta_1 + \cos^2\vartheta_2) + \langle E^2\rangle_{[001]}\cos^2\vartheta_3$$

where ϑ_1, ϑ_2, ϑ_3 are the angles between [hkl] and [100], [010], [001] respectively.

For the orthorhombic system:

$$O(\mathrm{hkl}) = \langle E^2\rangle_{[100]}\cos^2\vartheta_1 + \langle E^2\rangle_{[010]}\cos^2\vartheta_2 + \langle E^2\rangle_{[001]}\cos^2\vartheta_3$$

where ϑ_1, ϑ_2, ϑ_3 are the angles between [hkl] and [100], [010] and [001] respectively.

Two or three measurements (depending on the system) are enough to define the ellipsoid.

The monoclinic system requires additional calculations due to its unique symmetry. One of the three ellipsoid axes aligns with the direction [010], while the other two must be selected in the plane defined by **a*** and **c***. As a result, these axes coincide with the directions [h0l]. To correct for anisotropy in the monoclinic system, the polar direction with the largest G value, denoted as $[h_1 0 l_1]$, is identified. Next, a direction $[h_2 0 l_2]$ that is approximately or exactly perpendicular to $[h_1 0 l_1]$ is sought. This direction will be used to correct the anisotropy of the reciprocal space:

$$O(\mathrm{hkl}) = \left\langle E^2\right\rangle_{[h_1 0\ l_1]}\cos^2\vartheta_1 + \left\langle E^2\right\rangle_{[010\ ]}\cos^2\vartheta_2 + \left\langle E^2\right\rangle_{[h_2 0\ l_2]}\cos^2\vartheta_3$$

where ϑ_1, ϑ_2, ϑ_3 are the angles between [hkl] and $[h_1 0 l_1]$, [010] and $[h_2 0 l_2]$, respectively.

To correct for anisotropy in the triclinic system, the following procedure is applied. First, the polar direction with the largest G value, denoted as $[h_1 k_1 l_1]$, is identified. Next, a direction $[h_2 k_2 l_2]$ with the largest G value is found in the plane that is approximately or exactly perpendicular to $[h_1 k_1 l_1]$. Finally, a direction $[h_3 k_3 l_3]$ is identified that is perpendicular to both $[h_1 k_1 l_1]$ and $[h_2 k_2 l_2]$, either exactly or approximately. These directions will be used to correct the anisotropy of the reciprocal space in the triclinic system:

$$O(\mathrm{hkl}) = \left\langle E^2\right\rangle_{[h_1 k_1 l_1]}\cos^2\vartheta_1 + \left\langle E^2\right\rangle_{[h_2 k_2 l_2]}\cos^2\vartheta_2 + \left\langle E^2\right\rangle_{[h_3 k_3 l_3]}\cos^2\vartheta_3$$

where ϑ_1, ϑ_2, ϑ_3 are the angles between [hkl] and $[h_1 k_1 l_1]$, $[h_2 k_2 l_2]$ and $[h_3 k_3 l_3]$ respectively.

The anisotropy is then corrected by calculating the renormalized structure factors according to

$$|E'|^2_{obs} = |E|^2_{obs}/O$$

which replace the $|E|^2_{obs}$ in the RFOM calculations.

4.6. Figures of Merit for the Rotation Step

Giacovazzo [77] developed a method to directly derive the conditional probability distribution of a structure factor based on different types of prior information without calculating the joint probability distribution functions.

Once n model copies have been oriented and placed, the orientation of the (n + 1)th model copy in the target asu can be determined using the *RFOM* figure of merit, where

$$RFOM = CORR\left(|F|^2, \langle|F|^2\rangle\right) \tag{3}$$

RFOM is the correlation between $|F|^2$ and the expected value

$$\langle F^2\rangle = |F_{p1} + F_{p2} + \ldots F_{pn}|^2 + \sum_{s=1}^{m} |F_{ps}|^2 \tag{4}$$

where $F_{p1}, F_{p2}, \ldots, F_{pn}$ are the structure factors corresponding to the first, second, . . . , nth located model copy, m is the number of symmetry operators for the given space group and $\sum_{s=1}^{m}|F_{ps}|^2$ refers to the $(n + 1)$th model copy, for which we are searching the correct orientation. When $n = 0$, meaning that the first model copy is being rotated, Equation (4) simplifies to:

$$\langle|F|^2\rangle = \sum_{s=1}^{m} |F_{ps}|^2 \tag{5}$$

The *RFOM* figure is designed to identify the orientation of the (n + 1)th model copy that maximizes the *RFOM* value, which is expected to correspond to the correct solution. When searching for the orientation of the first model copy, the 200 orientations that correspond to the highest *RFOM* values are selected and passed to the translation step.

4.7. Figures of Merit for the Translation Step

A preliminary selection of the most promising translation vectors is made by using the criterion

$$\sum_h |F_h|^2 |F_{ph}|^2 = max \tag{6}$$

The left-hand side of Equation (6) is calculated via Fast Fourier Transform techniques according to Vagin & Teplyakov [78]. For each selected rotation, up to two translation vectors are accepted. However, the final ranking of the translation vectors is not determined immediately, as Criterion (6) may fail due to the imperfect orientation of the molecule, the presence of intermolecular vectors mixed with intramolecular ones, and the small sequence identity between the model and target. As a result, the determination of the final ranking of the translation vectors is postponed until these issues can be addressed.

Supplementary steps are taken to improve the ranking before making the final selection. First, the selected translations are scored [77] based on the following criterion

$$TFOM = CORR\left(|F|, |F_p|\right) \tag{7}$$

where

$$|F_p| = |F_{p1} + F_{p2} + \ldots F_{pn}| \tag{8}$$

F_{pi} represents the structure factor of the model, which is calculated based on the position of the i-th previously located copy of the model. When $n = 0$, meaning that the first model copy is being rotated, Equation (8) simplifies to:

$$|F_p| = |F_{p1}| \tag{9}$$

Criterion (9) benefits from a typical statistical behavior: when one or more model copies have already been placed, the variance representing the uncertainty in locating the next components is reduced. This results in an increase in the ratio of signal-to-noise for the next components.

However, Equation (7) may not work effectively when high-resolution data have been measured. In this case, the molecular model, which is defined by rotation and translation parameters based on low-resolution reflections (usually between 3 and 4 Å), may not be of sufficient quality for the high-resolution reflections. Small errors in these parameters may lead to large errors in the calculated amplitudes and phases of the high-resolution reflections. In this case, we still rely on Equation (7), but TFOM is calculated only on the reflections which are actively used in the MR step.

A situation where Equation (7) may not work effectively is when there is a pseudo-translational symmetry present. This type of symmetry generates a group of reflections with high intensities and another group with low intensities. To address this issue, Equation (10) is employed, where

$$TFOM = 1 - < |E_p|^2 > = max \tag{10}$$

$\langle |E_p|_2 \rangle$ is determined by computing it for reflections where the normalized observed structure factor $|E|$ is less than 0.3. Here, E_p is the normalized structure factor of F_p.

In our experience, both Criteria (7) and (10) are effective scoring functions. However, in some cases, the model may not be accurate enough, or the data may have limitations, making it difficult to identify good solutions based solely on the score values. To overcome such challenges, we select a variable number of the most promising solutions, selected by using either Criterion (7) or (10): they are further refined by using a rigid body refinement technique called SIMPLEX (see Section 4.8).

It is worth noting that RFOM and TFOM are unweighted FOMs, and we have not found any meaningful weights that can make them more effective.

4.8. Rigid Body Refinement by SIMPLEX

The solutions identified by the FOMs described in Section 4.7 are refined using the SIMPLEX method [79], which is an unconstrained optimization technique related to the downhill method. Here, the SIMPLEX method is applied to a six-dimensional parameter space, with three dimensions for rotation and three for translation. The refinement process typically results in a smaller average phase error, and it also facilitates the clustering of closely related solutions.

4.9. Selection of the Correct Solutions

The solutions refined using the SIMPLEX method undergo a cyclic procedure that combines applications of EDM and REFMAC [80]. This procedure is primarily focused on phase extension and refinement, and is crucial for the success of the crystal structure determination process. During this step (referred to as PRESYN to indicate that it precedes the SYNERGY step), the rigid body model obtained from the MR step is transformed into a model where individual atoms can shift to new positions under the control of REFMAC restraints. This cyclic procedure typically lowers the average phase errors of correct solutions while leaving the errors of false solutions unchanged. This makes it easier to distinguish correct solutions from false ones. The best solution is then identified based on the minimum REFMAC R value.

If only one copy of the model needs to be placed, the best solution is passed to the SYNERGY step for final phase refinement, and then to CAB. If multiple copies of the model

need to be located, the top five solutions are selected, and, one at a time, each is used as prior information to locate the second copy. The same practice is used to locate additional copies of the model (see Section 4.10).

4.10. About the Location of the Second and Further Model Copies

Let us consider a scenario where the first copy of the model has been oriented and located, and the search for the second model copy's rotation has started. REMO09 recovers the three Euler angles and corresponding three shift vectors that define the orientation and translation of the first model copy, and the Fp_1 values are calculated to start the search for the second model copy's roto-translation. However, this approach has a potential pitfall because Fp_1 arises from a rigid body model, and inaccuracies in the model orientation and location, as well as structural differences between the model and target, may create a systematic bias that can affect the FOMs effectiveness. As a result, it can be challenging to recognize the correct roto-translation parameters for the second model copy.

To overcome this issue, REMO22 refines for the first model copy using REFMAC, causing it to lose its original rigidity. Accordingly, structure factors corresponding to the first model copy are calculated from appropriate coordinates and used as prior information to locate the second model copy. The resulting phase improvement makes FOMs more effective at identifying the correct orientation and position of the second model copy. The same approach is used to locate additional copies. When all model copies are located, only the best solution is submitted to SYNERGY.

4.11. Automatic Restart

The success rate of MR may decrease when it is applied to models with lower scattering power compared to the target asu or when the root mean square deviation between the model and target structures is large. Let us assume that the final R value at the end of CAB is too high for proteins with SI < 0.4. In such cases, REMO22 is automatically restarted using a different strategy. According to Chothia & Lesk [81], when SI = 0.4, the root mean square deviation from the correct positions is 1.22 Å, which is likely an underestimate. This value makes it challenging to identify the correct rotation and translation. In these circumstances, it is expected that a high value of the crystallographic residual R will be observed between the calculated and observed structure factors, even when the model is correctly located. In REMO22, as is already the case in REMO09, when the SI < 0.4, up to 80% of the residues with the largest isotropic temperature factor are routinely treated as alanine during the SYNERGY step. This is done in the hope of removing atoms that are too far from their correct positions in the model. If the final R value at the end of CAB is greater than 0.50 and the SI < 0.4, a fully "alaninized" model is resubmitted to the REMO22 procedure.

4.12. Essential Directives

The full REMO22 + SYNERGY + CAB pipeline can be run automatically with very few directives. As an example, we will use 1aki structure:

```
%cab buccaneer
%structure 1aki
%job ORTHORHOMBIC FORM OF HEN EGG-WHITE LYSOZYME AT 1.5 Å RESOLUTION
%data
mtz 1aki.mtz
label H K L F SIGF
sequence 1aki.seq
%remo
fragment 2ihl.pdb
%end
```

If the users prefer to use PHASER or MOLREP as an MR program, they will need to provide a few additional directives to process their data through the segments SYNERGY + CAB (see the Supplementary Material section).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24076070/s1.

Author Contributions: Conceptualization, B.C., G.L.C. and C.G.; methodology, C.G.; software, B.C. and G.L.C.; validation, B.C. and G.L.C.; formal analysis, C.G.; investigation, B.C., G.L.C. and C.G.; data curation, B.C. and G.L.C.; writing—original draft preparation, C.G.; writing—review and editing, B.C., G.L.C. and C.G.; visualization, G.L.C.; All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are present within the article.

Acknowledgments: We thank Rocco Caliandro and Annamaria Mazzone for their fruitful discussions and helpful advice. We thank Maria Cristina Burla and Giampiero Polidori for the graphical section.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

asu	asymmetric unit
m	number of symmetry operators for a given space group
t, t_p	number of non-H atoms in the asymmetric units of the target and model structure, respectively
N = mt, $N_p = mt_p$	number of non-H atoms in the unit cells of the target and model structure, respectively. To simplify, all of the atoms are assumed to be in general position.
$F = \sum_{s=1}^{m} F_s$	structure factor of the target structure, where $F_s = \sum_{j=1}^{t} f_j \exp[2\pi i \boldsymbol{h}(\boldsymbol{R}_s \boldsymbol{r}_j + \boldsymbol{T}_s)]$, $\boldsymbol{r}_j$ are the atomic positions of the model structure
$F_p = \sum_{s=1}^{m} F_{ps}$	structure factor of the model structure, where $F_{ps} = \sum_{j=1}^{t_p} f_j \exp[2\pi i \boldsymbol{h}(\boldsymbol{R}_s \boldsymbol{r}_{pj} + \boldsymbol{T}_s)]$
E, E_p	normalized structure factors of F, F_p respectively
$\boldsymbol{r}_{pj}$	are the atomic positions of the model structure
EDM	electron density modification techniques
SI	sequence identity between target and model structure
AMB	automated model building
R	crystallographic R residual

References

1. Weeks, C.M.; DeTitta, G.T.; Hauptman, H.A.; Thuman, P.; Miller, R. Structure solution by minimal-function phase refinement and Fourier filtering. II. Implementation and applications. *Acta Crystallogr. A* **1994**, *50*, 210–220. [CrossRef]
2. Rappleye, J.; Innus, M.; Weeks, C.M.; Miller, R. SnB version 2.2: An example of crystallographic multiprocessing. *J. Appl. Crystallogr.* **2002**, *35*, 374–376. [CrossRef]
3. Sheldrick, G.M. SHELX Applications to Macromolecules. In *Direct Methods for Solving Macromolecular Structures*; Fortier, S., Ed.; Springer: Dordrecht, The Netherlands, 1998; pp. 401–411, ISBN 978-94-015-9093-8.
4. Foadi, J.; Woolfson, M.M.; Dodson, E.J.; Wilson, K.S.; Jia-xing, Y.; Chao-de, Z. A flexible and efficient procedure for the solution and phase refinement of protein structures. *Acta Cryst. D Biol. Crystallogr.* **2000**, *56*, 1137–1147. [CrossRef]
5. Palatinus, L. Ab initio determination of incommensurately modulated structures by charge flipping in superspace. *Acta Crystallogr. A* **2004**, *60*, 604–610. [CrossRef] [PubMed]
6. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. More power for direct methods: SIR2002. *Z. Krist.* **2002**, *217*, 629–635. [CrossRef]
7. Burla, M.C.; Caliandro, R.; Camalli, M.; Carrozzini, B.; Cascarano, G.L.; Caro, L.D.; Giacovazzo, C.; Polidori, G.; Spagna, R. SIR2004: An improved tool for crystal structure determination and refinement. *J. Appl. Crystallogr.* **2005**, *38*, 381–388. [CrossRef]

8. Giacovazzo, C. A general approach to phase relationships: The method of representations. *Acta Crystallogr. A* **1977**, *33*, 933–944. [CrossRef]
9. Giacovazzo, C. Representations of structure invariants and seminvariants. In *Direct Phasing in Crystallography: Fundamentals and Applications*; Oxford Science Publications; Oxford University Press: New York, NY, USA, 1998; pp. 243–274, ISBN 978-0-19-850072-8.
10. Frazão, C.; Sieker, L.; Sheldrick, G.; Lamzin, V.; LeGall, J.; Carrondo, M.A. Ab initio structure solution of a dimeric cytochrome c3 from Desulfovibrio gigas containing disulfide bridges. *JBIC J. Biol. Inorg. Chem.* **1999**, *4*, 162–165. [CrossRef] [PubMed]
11. Mooers, B.H.M.; Matthews, B.W. Extension to 2268 atoms of direct methods in the ab initio determination of the unknown structure of bacteriophage P22 lysozyme. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 165–176. [CrossRef]
12. Buerger, M.J. Phase Determination with the Aid of Implication Theory. *Phys. Rev.* **1948**, *73*, 927–928. [CrossRef]
13. Buerger, M.J. *Vector Space*; Wiley: New York, NY, USA, 1959; Chapter 11.
14. Simpson, P.G.; Dobrott, R.D.; Lipscomb, W.N. The symmetry minimum function: High order image seeking functions in X-ray crystallography. *Acta Crystallogr.* **1965**, *18*, 169–179. [CrossRef]
15. Richardson, J.W.; Jacobson, R.A. Computer-aided analysis of multi-solution Patterson superpositions. In *Patterson and Pattersons: Fifty Years of the Patterson Function: Proceedings of a Symposium Held at the Institute for Cancer Research, the Fox Chase Cancer Center, Philadelphia, PA, USA, November 13–15, 1984*; Glusker, J.P., Patterson, B.K., Rossi, M., Eds.; International Union of Crystallography Crystallographic Symposia; Oxford University Press: New York, NY, USA, 1987; ISBN 978-0-19-855230-7.
16. Sheldrick, G.M. Tutorial on automated Patterson interpretation to find heavy atoms. In *Crystallographic Computing 5: From Chemistry to Biology*; Moras, D., Podjarny, A.D., Thierry, C., Eds.; IUCr Crystallographic Symposia; Oxford University Press: New York, NY, USA, 1991; pp. 145–157, ISBN 978-0-19-855384-7.
17. Pavelčík, F.; Kuchta, L.; Sivý, J. Patterson-oriented automatic structure determination. Utilizing Patterson peaks. *Acta Crystallogr. A* **1992**, *48*, 791–796. [CrossRef]
18. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; Caro, L.D.; Giacovazzo, C.; Mazzone, A.; Siliqi, D. Ab initio phasing of proteins with heavy atoms at non-atomic resolution: Pushing the size limit of solvable structures up to 7890 non-H atoms in the asymmetric unit. *J. Appl. Crystallogr.* **2008**, *41*, 548–553. [CrossRef]
19. Rossmann, M.G.; Blow, D.M. The Detection of Sub-Units within the Crystallographic Asymmetric Unit. *Acta Crystallogr.* **1962**, *15*, 24–31. [CrossRef]
20. Rossmann, M.G. *The Molecular Replacement Method*; Gordon & Breach: New York, NY, USA, 1972.
21. Rossmann, M.G. The Molecular Replacement Method. *Acta Crystallogr A* **1990**, *46*, 73–82. [CrossRef]
22. Kissinger, C.R.; Gehlhaar, D.K.; Fogel, D.B. Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55*, 484–491. [CrossRef]
23. Jamrog, D.C.; Zhang, Y.; Phillips, G.N.J. SOMoRe: A multi-dimensional search and optimization approach to molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59*, 304–314. [CrossRef]
24. Glykos, N.M.; Kokkinidis, M. A stochastic approach to molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 169–174. [CrossRef]
25. Fujinaga, M.; Read, R.J. Experiences with a new translation-function program. *J. Appl. Crystallogr.* **1987**, *20*, 517–521. [CrossRef]
26. Navaza, J. AMoRe: An automated package for molecular replacement. *Acta Crystallogr. A* **1994**, *50*, 157–163. [CrossRef]
27. Read, R.J. Detecting outliers in non-redundant diffraction data. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55*, 1759–1764. [CrossRef]
28. Vagin, A.; Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 22–25. [CrossRef] [PubMed]
29. McCoy, A.J.; Grosse-Kunstleve, R.W.; Adams, P.D.; Winn, M.D.; Storoni, L.C.; Read, R.J. *Phaser* crystallographic software. *J. Appl. Crystallogr.* **2007**, *40*, 658–674. [CrossRef] [PubMed]
30. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Mazzone, A.; Siliqi, D. Molecular replacement: The probabilistic approach of the program REMO09 and its applications. *Acta Crystallogr. A* **2009**, *65*, 512–527. [CrossRef] [PubMed]
31. Rigden, D.J.; Thomas, J.M.H.; Simkovic, F.; Simpkin, A.; Winn, M.D.; Mayans, O.; Keegan, R.M. Ensembles generated from crystal structures of single distant homologues solve challenging molecular-replacement cases in AMPLE. *Acta Crystallogr. D Struct. Biol.* **2018**, *74*, 183–193. [CrossRef]
32. Millán, C.; Sammito, M.; Usón, I. Macromolecular ab initio phasing enforcing secondary and tertiary structure. *IUCrJ* **2015**, *2*, 95–105. [CrossRef] [PubMed]
33. Millán, C.; Jiménez, E.; Schuster, A.; Diederichs, K.; Usón, I. ALIXE: A phase-combination tool for fragment-based molecular replacement. *Acta Crystallogr. D Struct. Biol.* **2020**, *76*, 209–220. [CrossRef] [PubMed]
34. Brünger, A.T.; Adams, P.D.; Clore, G.M.; DeLano, W.L.; Gros, P.; Grosse-Kunstleve, R.W.; Jiang, J.S.; Kuszewski, J.; Nilges, M.; Pannu, N.S.; et al. Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr. D Biol. Crystallogr.* **1998**, *54*, 905–921. [CrossRef]
35. 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.W.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 235–242. [CrossRef]
36. Bricogne, G.; Vonrhein, C.; Flensburg, C.; Schiltz, M.; Paciorek, W. Generation, representation and flow of phase information in structure determination: Recent developments in and around SHARP 2.0. *Acta Crystallogr. D Biol. Crystallogr.* **2003**, *59*, 2023–2030. [CrossRef]

37. Adams, P.D.; Afonine, P.V.; Bunkóczi, G.; Chen, V.B.; Davis, I.W.; Echols, N.; Headd, J.J.; Hung, L.-W.; Kapral, G.J.; Grosse-Kunstleve, R.W.; et al. *Phenix*: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **2010**, *66*, 213–221. [CrossRef] [PubMed]
38. Sheldrick, G.M. Crystal structure refinement with SHELXL. *Acta Crystallogr. C Struct. Chem.* **2015**, *71*, 3–8. [CrossRef] [PubMed]
39. Burla, M.C.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. Synergy among phase-refinement techniques in macromolecular crystallography. *Acta Crystallogr. D Struct. Biol.* **2017**, *73*, 877–888. [CrossRef] [PubMed]
40. Cowtan, K. Fast Fourier feature recognition. *Acta Crystallogr. D Biol. Crystallogr.* **2001**, *57*, 1435–1444. [CrossRef]
41. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Phasing at resolution higher than the experimental resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61*, 556–565. [CrossRef]
42. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Ab initio phasing at resolution higher than experimental resolution. *Acta Crystallogr. D Biol. Crystallogr.* **2005**, *61*, 1080–1087. [CrossRef]
43. Giacovazzo, C.; Siliqi, D. Improving Direct-Methods Phases by Heavy-Atom Information and Solvent Flattening. *Acta Crystallogr. A* **1997**, *53*, 789–798. [CrossRef]
44. Burla, M.C.; Caliandro, R.; Giacovazzo, C.; Polidori, G. The difference electron density: A probabilistic reformulation. *Acta Crystallogr. A* **2010**, *66*, 347–361. [CrossRef]
45. Burla, M.C.; Giacovazzo, C.; Polidori, G. From a random to the correct structure: The VLD algorithm. *J. Appl. Crystallogr.* **2010**, *43*, 825–836. [CrossRef]
46. Giacovazzo, C. Solution of the phase problem at non-atomic resolution by the phantom derivative method. *Acta Crystallogr. A Found. Adv.* **2015**, *71*, 483–512. [CrossRef]
47. Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C. Phase improvement via the Phantom Derivative technique: Ancils that are related to the target structure. *Acta Crystallogr. D Struct. Biol.* **2016**, *72*, 551–557. [CrossRef] [PubMed]
48. Giacovazzo, C. From direct-space discrepancy functions to crystallographic least squares. *Acta Crystallogr. A Found. Adv.* **2015**, *71*, 36–45. [CrossRef]
49. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62*, 1002–1011. [CrossRef]
50. Cowtan, K. Automated nucleic acid chain tracing in real time. *IUCrJ* **2014**, *1*, 387–392. [CrossRef]
51. 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*, 1171–1179. [CrossRef] [PubMed]
52. Terwilliger, T.C.; Grosse-Kunstleve, R.W.; Afonine, P.V.; Moriarty, N.W.; Zwart, P.H.; Hung, L.-W.; Read, R.J.; Adams, P.D. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 61–69. [CrossRef]
53. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Polidori, G.; Giacovazzo, C. CAB: A cyclic automatic model-building procedure. *Acta Crystallogr. D Struct. Biol.* **2018**, *74*, 1096–1104. [CrossRef] [PubMed]
54. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Polidorib, G. How far are we from automatic crystal structure solution via molecular-replacement techniques? *Acta Crystallogr. D Struct. Biol.* **2020**, *76*, 9–18. [CrossRef]
55. Burla, M.C.; Carrozzini, B.; Cascarano, G.L.; Giacovazzo, C.; Polidori, G. Cyclic automated model building (CAB) applied to nucleic acids. *Crystals* **2020**, *10*, 280. [CrossRef]
56. Cascarano, G.L.; Giacovazzo, C. Towards the automatic crystal structure solution of nucleic acids: Automated model building using the new CAB program. *Acta Crystallogr. D Struct. Biol.* **2021**, *77*, 1602–1613. [CrossRef]
57. Keegan, R.M.; Winn, M.D. MrBUMP: An automated pipeline for molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* **2008**, *64*, 119–124. [CrossRef] [PubMed]
58. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef] [PubMed]
59. Terwilliger, T.C.; Poon, B.K.; Afonine, P.V.; Schlicksup, C.J.; Croll, T.I.; Millán, C.; Richardson, J.S.; Read, R.J.; Adams, P.D. Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* **2022**, *19*, 1376–1382. [CrossRef] [PubMed]
60. Bond, P.S. Next Generation Software for Placing Atoms into Electron Density Maps. Ph.D. Thesis, University of York, York, UK, 2021.
61. Stein, N. *CHAINSAW*: A program for mutating pdb files used as templates in molecular replacement. *J. Appl. Crystallogr.* **2008**, *41*, 641–643. [CrossRef]
62. 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. Crystallogr.* **2015**, *48*, 306–309. [CrossRef]
63. Terwilliger, T.C. Reciprocal-space solvent flattening. *Acta Crystallogr. D Biol. Crystallogr.* **1999**, *55*, 1863–1871. [CrossRef]
64. Terwilliger, T.C. Maximum-likelihood density modification. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 965–972. [CrossRef]
65. Bricogne, G. Maximum entropy and the foundations of direct methods. *Acta Crystallogr. A* **1984**, *40*, 410–445. [CrossRef]
66. Bricogne, G. A Bayesian statistical theory of the phase problem. I. A multichannel maximum-entropy formalism for constructing generalized joint probability distributions of structure factors. *Acta Crystallogr. A* **1988**, *44*, 517–545. [CrossRef]
67. Lunin, V.Y. Electron-denisty histograms and the phase problem. *Acta Crystallogr. D Biol. Crystallogr.* **1993**, *49*, 90–99. [CrossRef]

68. Cascarano, G.L.; Cuocci, C.; Mallamo, M.; Carrozzini, B.; Moliterni, A. JAV (Just Another Viewer). Istituto di Cristallografia, The National Research Council (CNR), Bari, Italy. Graphic software to display and manipulate atomic models of small structures or macomolecules. Unpublished work. 2021.
69. DiMaio, F.; Terwilliger, T.C.; Read, R.J.; Wlodawer, A.; Oberdorfer, G.; Wagner, U.; Valkov, E.; Alon, A.; Fass, D.; Axelrod, H.L.; et al. Improved molecular replacement by density- and energy-guided protein structure optimization. *Nature* **2011**, *473*, 540–543. [CrossRef]
70. Das, R.; Baker, D. Prospects for de novo phasing with de novo protein models. *Acta Crystallogr. D Biol. Crystallogr.* **2009**, *65*, 169–175. [CrossRef]
71. Matthews, B.W. Solvent content of protein crystals. *J. Mol. Biol.* **1968**, *33*, 491–497. [CrossRef] [PubMed]
72. Kantardjieff, K.A.; Rupp, B. Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **2003**, *12*, 1865–1871. [CrossRef] [PubMed]
73. Quillin, M.L.; Matthews, B.W. Accurate calculation of the density of proteins. *Acta Crystallogr. D Biol. Crystallogr.* **2000**, *56*, 791–794. [CrossRef]
74. Fischer, H.; Polikarpov, I.; Craievich, A.F. Average protein density is a molecular-weight-dependent function. *Protein Sci.* **2009**, *13*, 2825–2828. [CrossRef]
75. Hirshfeld, F.L. Symmetry in the generation of trial structures. *Acta Crystallogr. A* **1968**, *24*, 301–311. [CrossRef]
76. Altomare, A.; Burla, M.C.; Cascarano, G.; Giacovazzo, C.; Guagliardi, A.; Moliterni, A.G.G.; Polidori, G. Early Finding of Preferred Orientation: Applications to Direct Methods. *J. Appl. Crystallogr.* **1996**, *29*, 341–345. [CrossRef]
77. Giacovazzo, C. Updating direct methods. *Acta Crystallogr. A Found. Adv.* **2019**, *75*, 142–157. [CrossRef]
78. Vagin, A.; Teplyakov, A. MOLREP: An Automated Program for Molecular Replacement. *J. Appl. Crystallogr.* **1997**, *30*, 1022–1025. [CrossRef]
79. Rowan, T. Functional Stability Analysis of Numerical Algorithms. Ph.D. Thesis, University of Texas, Austin, TX, USA, 1990.
80. Murshudov, G.N.; Skubák, P.; Lebedev, A.A.; Pannu, N.S.; Steiner, R.A.; Nicholls, R.A.; Winn, M.D.; Long, F.; Vagin, A.A. *Refmac* 5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **2011**, *67*, 355–367. [CrossRef] [PubMed]
81. Chothia, C.; Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823–826. [CrossRef] [PubMed]