UNIVERSITAT DE BARCELONA

Phase Combination and its Application to the Solution of Macromolecular Structures: Developing ALIXE and SHREDDER

Claudia Lucía Millán Nebot

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L’ALIMENTACIÓ

PHASE COMBINATION AND ITS APPLICATION TO THE SOLUTION OF MACROMOLECULAR STRUCTURES: DEVELOPING ALIXE AND SHREDDER

CLAUDIA LUCÍA MILLÁN NEBOT

2018

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L’ALIMENTACIÓ

PROGRAMA DE DOCTORAT EN BIOTECNOLOGIA

PHASE COMBINATION AND ITS APPLICATION TO THE SOLUTION OF MACROMOLECULAR STRUCTURES: DEVELOPING ALIXE AND SHREDDER

Memòria presentada per Claudia Lucía Millán Nebot per optar al títol de doctor per la Universitat de Barcelona

Directora: Prof. Dr. Isabel Usón Finkenzeller

Doctoranda: Claudia Lucía Millán Nebot

Tutora: Prof. Dr. Josefa Badía Balacín

CLAUDIA LUCÍA MILLÁN NEBOT
2018

To those who have supported me

"… how nice it would be if we could only get through into Looking-glass House! I'm sure it's got, oh! such beautiful things! Let's pretend there's a way of getting through into it, somehow…"

Through the Looking-glass, and What Alice found there
Lewis Carroll, 1871

ACKNOWLEDGEMENTS

I would like to first acknowledge all the support and mentorship that my supervisor, Isabel Usón, has provided me with throughout these years. From the very beginning, she entrusted me with projects that have helped me to develop so many skills, that range from the scientific to the personal, that have made me grow so much. She has been capable of finding which was the right moment to push me, and which was the moment to leave me the necessary space to figure out things by myself. She has truthfully supported me and help me taking decisions about my research and my life, and I am really grateful for it. To her and her family, Agustín, Flavia, Miranda and Candela.

The next person I need to mention, is my friend and colleague Massimo. We have spent together hours and hours and learnt so much from our defeats and our triumphs. He has been my mentor in programming, pushing me to achieve my best and facilitating me the resources to keep expanding my knowledge. And we have been through many experiences together, both positive and negative, that have only strengthened our bond, both personally and in our research. The research presented in this thesis includes work we have done together, and I am grateful to him because he established the basis of the collaborative setup in which we develop our software now. I have understood what real teamwork means with him.

I am very grateful to my current colleagues at the Arcimboldo Team. To Ana and Iracema, for inspiring me and let me push them through this process of learning crystallography and programming when you come from a biotechnology background. To Giovanna, for her support and her interest in the work I do even though instead of working with her in the wet lab I hanged the lab coat and went for programming! To Nicolás and Rafa, for their questions and discussions which have helped me learn about new aspects of crystallography and its applications.

I am also grateful to all the former members of the Arcimboldo Team that are not currently working with us, but who spend their time with me on the first years of the PhD: Dayté, Iván, and Kathrin. Also to our former and current system administrators or collaborators for our hardware setup: César, Guillem, Iñaki, Alfonso and Pep. They made a wonderful environment in which we could use our own institute as a supercomputer possible. They provided a nice website, helped with the distribution, the testing and so much more. Having you there has made a huge difference, and I have learned a lot from you guys. I want to mention too all the members of the other groups at the Structural Biology Unit, as they constitute a really engaging community in which to work.

I would also like to thank my professors from the Pablo de Olavide University. In particular, Antonio Prado and Fernando Govantes, who introduced me to research and taught me not only the theory but also the practice of their disciplines. Thanks to the time I spent at their labs, I realised what was research really about, and developed skills that have been extremely useful during my PhD. I also want to acknowledge Antonio Perez, who showed me the relevance of Bioinformatics and taught us in the most engaging way. In fact, the time spent with my friends Pablo and Aida working together in his subject is one of my best recollections from my time at the University. I also want to thank them for all their support during the degree and afterwards.

It’s also worth mentioning my experience with the Master in Crystallography and Crystallization. I am thankful to its organisers, its professors and my colleagues. It was a deep immersion into crystallography. Furthermore it gave me the opportunity to meet Almudena and Rubén, and forge a friendship that has grown during my PhD.

I want to thank Randy Read for receiving me in his group in Cambridge during almost a year. I have learnt a lot from him and I have enjoyed sharing so many nice coffees and discussions with his group. I am particularly thankful to Airlie for her patience in answering all the questions I threw at her during coffee breaks, and for her speed in implementing changes that were required for us. And to Rob and Tristan, to whom one can ask almost anything about programming and if they don’t know the answer, they will point you in the right direction anyway!

I also want to thank George Sheldrick, both for his contribution in what was the initial project of my PhD, the phase combination of partial solutions, and for the continuous support of the projects our groups develop together, enhancing this collaboration by receiving me up to four times during the summer break to work in Göttingen with his group. From his team, I want to thank Jens, Anna and Claudia.

I would also like to thank all our collaborators providing data, as it has been through interesting, challenging cases that I have developed and tested our software. And to Judith Juanhuix, Lynn Tenn Eyck, and Kay Diederichs for giving us access to supercomputers and grids and establishing fruitful collaborations with us.

I also want to thank my family. My parents, Fede and Nuria, who have loved me and supported me through the years. Hearing you talking with pride of what I do just feels wonderful. My aunts/uncles/cousins, scattered all over Spain and even some other countries (my dear Cari, in UK, and Susan, in the States). Elena, Ernesto, David, Tony, Ana, Roser, Lluisa and all your families. You all have fed my curiosity and make me feel I was doing something worthwhile. You help me discover much and more about all sorts of things. My grandmother, Clara.

My family in Sevilla became extended by the friends of my parents who acted as if they were my aunts/uncles. I cannot forget Nina, telling me about how I will supposedly cure diabetes, or Fernando, who help me with maths when I was still a teenager. And Jesús, María, Ana, Nela, Alberto... My neighbourhood itself has been a real social network in which people have supported each other, including me, so I also want to thank my super neighbours!

My friends. Naming all of those who I consider my friends will take just too much long. Some of them might be closer than others to me, but there is a lot of people out there that I feel with their friendship they have contributed to this work. Nonetheless, I would particularly like to thank Aldara, because we have shared so much and she has taught me much more than she thinks. Thanks also to Rocío, who even in the distance kept checking on me during this time, and to Laura, Francisco, and Santi, for their comprehension and their company in Barcelona.

Last but not least, during more than 10 years I have been sharing my life with Javi, who is a great friend and a wonderful partner. He has given me support beyond measure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

ASU	Asymmetric unit
CC	Correlation Coefficient
CCP4	Collaborative Computational Project Nº4
eLLG	Expected Log Likelihood Gain
FOM	Figure of merit
LLG	Log Likelihood Gain
ML	Maximum Likelihood
MPE	Mean Phase Error
MPD	Mean Phase Difference
MR	Molecular Replacement
NCS	Non-Crystallographic Symmetry
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
SNR	Signal-to-Noise Ratio

SUMMARY

Phasing X-ray data within the frame of the ARCIMBOLDO programs requires very accurate models and a sophisticated evaluation of the possible hypotheses. ARCIMBOLDO uses small fragments, that are placed with the maximum likelihood molecular replacement program *Phaser*, and are subject to density modification and autotracing with the program SHELXE. The software receives its name from the Italian painter Giuseppe Arcimboldo, who used to compose portraits out of common objects such as vegetables or flowers. Out of most possible arrangements of such objects, only a still-life will result, and just a few ones will truly produce a portrait. In a similar way, from all possible placements with small protein fragments, only a few will be correct and will allow to get the full "protein's portrait".

Figure 1 Galatea of the Spheres, by Dalí

Salvador Dalí, 1952. Dalí Theatre and Museum, Figueres, Girona, Spain. From Wikipedia

The work presented in this thesis has explored new ways to exploit partial information and increase the signal in the process of phasing with fragments. This has been achieved through two main pieces of software, ALIXE and SHREDDER. With the spherical mode in ARCIMBOLDO_SHREDDER, the aim is to derive compact fragments starting from a distant homolog to our unknown protein of interest. Then, locations for these fragments are searched with *Phaser*. These include strategies for refining the fragments against the experimental data and giving them more degrees of freedom. With ALIXE, the aim is to combine information in reciprocal space from partial solutions, such as the ones produced by SHREDDER, and use the coherence between them to guide their merging and to increase the information content, so that the step of density modification and autotracing starts from a more complete solution. Even if partial solutions contain both correct and incorrect information, the combination of solutions that share some similarity will allow to get a better approximation to the correct structure. A beautiful analogy to our approach can be found in the picture shown in Figure 1, presenting a piece from Dalí, the "Galatea of the Spheres". Inside it, one sees how spheres similar in colour and size provide enough information to intuitively reconstruct Gala's portrait.

Both ARCIMBOLDO_SHREDDER and ALIXE have been used on test data for development and optimisation but also on datasets from previously unknown structures, which have been solved thanks to these programs. These programs are distributed through the website of the group but also through software suites of general use in the crystallographic community such as CCP4 and SBGrid.

This thesis is organized into the following sections:

OBJECTIVES: Including an enumeration of the main questions aimed to address in this doctoral work.

INTRODUCTION (CHAPTER 1): Including the basic theoretical background required to understand the results of this work. The first concept presented is that of signal to noise ratio and its relevance to science in general and crystallography in particular (section 1.1). Afterwards, section 1.2. discusses the phase problem, which is the major bottleneck in macromolecular crystallographic determination. In section 1.3, an overview of different solutions to the phase problem is presented, but the major focus is then in molecular replacement in section 1.4. Section 1.5 discusses how to improve noisy electron density maps and interpret them by density modification. Lastly, section 1.6 introduces the ARCIMBOLDO framework for phasing with small fragments.

MATERIALS AND METHODS (CHAPTER 2): Covering a description of all the external software and algorithms that have been used in this work. In section 2.1, the figures of merit that are used to compare solutions within the thesis are described in detail. Section 2.2 presents the external software used by our methods, mainly *Phaser* and SHELXE. Section 2.3 describes the software environment in which our programs are developed, and section 2.4 our main distribution systems. In section 2.5, the hardware used for the different experiments in the thesis is characterised, and in section 2.6 the test data are presented.

RESULTS AND DISCUSSION (CHAPTERS 3, 4, 5): Including three chapters with the results obtained during this doctoral work.

Chapter 3 is dedicated to ALIXE, and after a brief introduction (section 3.1), phase comparison measures are defined (section 3.2). In section 3.3, the proof of principle on using phase combination from partial solutions is established by using perfect fragments and studying their maps. Section 3.4 is devoted to the use of phase combination with real solutions out of ARCIMBOLDO_LITE runs. In section 3.5, the use of larger, related fragments is discussed in the context of ARCIMBOLDO_SHREDDER solutions. Then, section 3.6 explores the landscape of possible origin shifts for polar space groups with both overlapping and independent fragments. In Section 3.7, phase combination with solutions from ARCIMBOLDO_BORGES runs with libraries of local folds is described. Section 3.8 presents the main conclusions about phase combination in the different scenarios of the ARCIMBOLDO programs and fragments. Lastly, section 3.9 describes in detail the current implementation of ALIXE.

Chapter 4 deals with the spherical mode in ARCIMBOLDO_SHREDDER, a phasing method that uses distant homologs, extracts compact fragments and refines them. After a short discussion on distant homologs and their application in molecular replacement (section 4.1), the external methods used for model improvement are described (section 4.2). Section 4.3 presents the proof of principle that established the idea to implement the spherical mode, while section 4.4 introduces the possible strategies for finding a suitable homolog. The spherical mode implementation is discussed in detail in section 4.5 and its use and parameterization is illustrated in section 4.6.

Chapter 5 shows the practical impact that the methods described in chapters 3 and 4 have had for structure solution of previously unknown structures. Test cases are also analyzed. After a brief introduction (section 5.1), the cases of PPAD (section 5.2), LTG (section 5.3), HHED2 (section 5.4), and SYCP1 (section 5.5) are discussed. Section 5.6 comments on a few other previously unknown structures that have been solved with these methods but are not yet published. Finally, section 5.7 includes results of the combined use of ARCIMBOLDO_SHREDDER's spherical mode and ALIXE in a set of 43 structures, all in different space groups.

CONCLUSIONS: Detail the main outcome of the work performed in this doctorate.

OUTLOOK: Providing a short discussion of possible developments and questions to address within the context of the work described in the thesis.

REFERENCES: listing all the bibliography cited in the main text.

CURRICULUM VITAE: including scientific production and participation in scientific events.

OBJECTIVES

The overall objective of this thesis is to push the limits of fragment-based phasing methods in macromolecular crystallography. This end has been pursued using two main methodologies. The first one was to generate and improve fragments from distant homologs in order to use them for structure solution. The second one was to design a procedure for the combination of correct but very partial solutions. Within this overall goal, specific scientific and technical objectives can be distinguished.

- To recognise correct partial solutions within ARCIMBOLDO and characterise them in both real and reciprocal space.
- To develop a phase combination method for such correct partial solutions:
    - To make the procedure applicable to any kind of symmetry, including approximations for cases where the origin is not constrained.
    - To use a phase comparison measure that allows weighting the contributions of the different sets.
    - To devise a scoring function that breaks ambiguity in borderline cases and allows the recognition of correct combinations of phase sets.
- To implement the phase combination in the software ALIXE. This program should give the possibility for users both to combine solutions within the ARCIMBOLDO programs as well as used as a standalone tool.
- To extend the efficiency and applicability of ALIXE by testing the implementation on both test and unknown structures to establish an optimal parameterisation of the algorithms.
- To study how to use protein fragments derived from distant homologs for structure solution.
    - To devise a successful strategy to choose the homolog or set of homologs to use.
    - To design an efficient algorithm to cut the template homolog into pieces that can be used as models.
    - To use partial solution refinement to improve the fragments.
    - To combine the partial solutions.
- To implement the procedure in the ARCIMBOLDO_SHREDDER software.
- To extend the efficiency and applicability of ARCIMBOLDO_SHREDDER by testing the implementation on test and unknown structures to establish an optimal parameterisation of the algorithms.
- The solution of previously unknown structures of biological relevance by ARCIMBOLDO_SHREDDER and ALIXE.
- To distribute the developed software both through our website and through other platforms such as CCP4, making it available to the crystallographic community
- To test the performance of the distributed versions and produce online documentation.

1 INTRODUCTION

1.1 Signal to noise ratio and its scientific relevance

Extracting information from large amounts of data with high levels of error and correlation is critical in many sciences, as for example, astrophysics (Bosh et al., 1986), nanotechnology (Heller et al., 2009) or medicine (Chakraborty & Das, 2012). More closely related with the topic of this thesis, biophysical techniques such as Nuclear Magnetic Resonance (NMR) (Hyberts et al., 2013) are also subject to the problem of maximizing the information derived from noisy data. The signal-to-noise ratio (SNR) is a quantitative measure that can be described in terms of the relation between the amount of information present (signal), and the entropy of the system (noise) (Zhanabaev et al., 2016). To estimate the SNR, it is fundamental to determine the power of the noise level experimentally. This can be done by using autocorrelation functions between the signal and the noise but can prove to be very difficult in certain situations.

In order to reduce the noise, different procedures can be applied depending on its nature. In the case of random Gaussian noise, signal averaging can reduce it successfully and increase the signal (Umer & Muhammad Sabieh, 2010). Filters of different nature can also be used, such as the Kalman filter (Kalman, 1960), which is commonly used in spacecraft navigation and other signal processing applications.

In the field of macromolecular crystallography, recent work has shown promising algorithms and metrics to distinguish, in the case of composite datasets (made of experimental data measured on different objects), which differences are genuine and which are due to the systematic and random error (Diederichs, 2017). The method is general and it has been applied successfully for merging crystallographic datasets coming from microcrystals (Yamashita et al., 2018, Gildea & Winter, 2018) or different regions of a crystal, allowing to study its polymorphism (Thompson et al., 2018).

Another example of the relevance of SNR in crystallography can be found within macromolecular phasing methods with partial models, which generate many non-independent partial solutions. Filtering and combining the information available can lead to success in difficult cases. This requires statistical treatment and the application of prior knowledge about the system while maintaining a computationally feasible sampling. An example of this strategy is the combination of partial Molecular Replacement (MR) solutions (Buehler et al., 2009). The genesis of these solutions makes them non-independent, and coherence can be exploited and weighted positively in the hope that it will contribute to discriminate which solutions have improved and set them apart from the rest.

1.2 X-ray crystallography and the phase problem

Structural biology techniques can be considered as ‘the eyes of biology’, because they provide three-dimensional models of molecules. In particular, crystallography allows to obtain them with the highest level of detail. It started more than 100 years ago when Laue (Friedrich *et al.*, 1913, Laue, 1913) and the Braggs (Bragg & Bragg, 1913) characterized the diffraction of X-rays and deduced how to determine the atomic structure of molecules in crystals from their diffraction patterns. Since then, it has evolved to a very effective technique that, together with other experimental approaches, allows scientists to get accurate models that serve as frameworks for the understanding of the actors and the mechanisms in biological processes.

In any experiment involving electromagnetic radiation, the level of detail that will be distinguishable will depend on the wavelength used. This can be understood within the frame of the Nyquist-Shannon sampling theorem (Shannon, 1949). The detection of a signal, even if the signal itself is continuous, is limited by the fact that detectors are sampling the signal in discrete steps. Due to the interaction of the radiation with objects, the size of the wavelength must be proportional to the minimal distances that we aim to distinguish. Consequently, in order to 'see' atoms in molecules, radiation of a much shorter wavelength than that of visible light is required. This is typically achieved with X-rays, and occasionally neutrons or electrons. X-ray optics do not allow direct imaging of molecules, something that can be done, for example, with electron microscopy. With X-rays, another physical phenomenon between matter and electromagnetic radiation must be exploited, which is diffraction.

Diffraction occurs when waves encounter obstacles and are scattered by them. The waves can interfere in a constructive or destructive manner. Crystals are ordered solids that produce a sharp diffraction pattern. They are constituted by a large number of identical, ordered molecules, and the interference between the X-rays scattered by this three-dimensional grid results in empty regions together with sharp peaks of diffraction. The mathematical description of the geometry of diffraction of X-rays by crystals, known as Bragg's Law (Bragg, 1913a) is formulated (equation 1) in terms of the constructive interference of a radiation of wavelength λ after the reflection by planes in the crystal separated by a distance *d*, at a glancing angle θ.

$$\mathbf{n}\,\boldsymbol{\lambda} = \mathbf{2\,d\,sin\,\theta} \qquad \textbf{(1)}$$

Once diffraction data have been recorded and processed, the aim is to retrieve the electron density in the crystal. *d* is called the maximal resolution, and it approximately matches the resolvability of structural features in the resulting electron density map. This resolution limit is conditioned by the crystal properties: disorder lowers the resolution to which a crystal can diffract. The relation between the experimental data and the electron density is given by the Fourier Transform (Cooley & Tukey, 1965) of the individual structure factors (complex numbers bearing the amplitude and phase of each diffracted X-ray that gave rise to a peak). This mathematical tool allows, starting from a function that is a composite signal made up of many contributions, to isolate the individual ones. In general, most of the well-known periodic functions can be represented in terms of a Fourier series, as a sum of several sinusoidals of the appropriate frequency, amplitude and phase. In the diffraction case, the individual contributions are the structure factors. A *structure factor* (equation 2) for a given reflection (which implies a certain scattering angle) is the quantity that expresses both the amplitude and the phase of that reflection. Given a fixed origin, it is independent of the method and the conditions of observation of the reflection, but it will depend on the position and the electron density of each atom in the structure. Consequently, each structure factor has information about all atoms in the structure.

$$\mathbf{F}_{\boldsymbol{hkl}} \;=\; \boldsymbol{\Sigma}_{\boldsymbol{j}}\, \boldsymbol{f}_{\boldsymbol{j}}\, \mathbf{e}^{[2\pi i(hx_j+ky_j+lz_j)]} \qquad \mathbf{(2)}$$

f_j is a factor that expresses the scattering of X-rays by an atom in terms of the scattering of a single electron. In (*International Tables for Crystallography Volume C: Mathematical, Physical and Chemical Tables*, 2004), *f* values and the parameters used to approximate them are tabulated.

h, *k*, *l* are the Miller indices (Miller, 1839), that describe the imaginary lattice planes throughout the crystals on which incoming X-rays would reflect to give rise to the observed diffraction pattern (hence the term 'reflection'). Each set of *h*, *k*, *l* indexes represent a reflection from the diffraction pattern.

The relation between the individual structure factors and the electron density is given by equation 3.

$$\boldsymbol{\rho}\ (\mathbf{xyz}) = \frac{\mathbf{1}}{\boldsymbol{V}} \sum_{-\infty}^{+\infty} |\,\boldsymbol{F}(\boldsymbol{hkl})|\ \ \boldsymbol{e}^{[-2\pi i(hx+ky+lz\,-\,\phi(hkl))]} \qquad \mathbf{(3)}$$

The computation of the electron density **ρ** from the individual structure factors is not straightforward, because the measurement misses the phases ϕ. During the experiment, only the intensities from the diffracted beams can be recorded, which are proportional to the square of the modulus of the structure factors (amplitudes) $|\boldsymbol{F}(\boldsymbol{hkl})|$ (Figure 2). This is known as the *phase problem* in crystallography and it needs to be overcome in order to reconstitute the structure in each crystallographic structure determination.

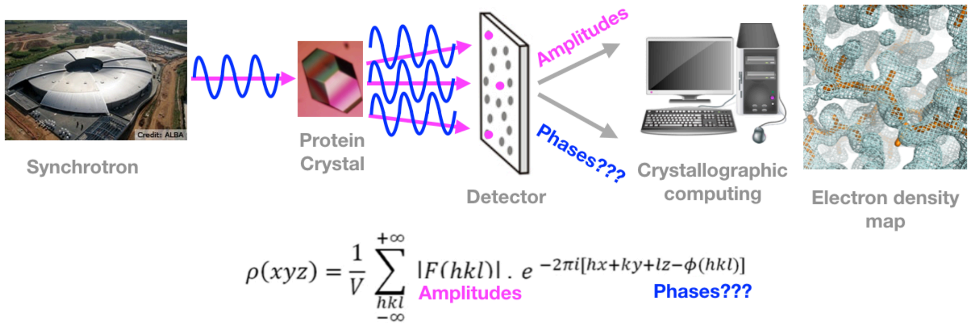

Credits: ALBA
Synchrotron
Protein Crystal
Detector
Amplitudes
Phases???
Crystallographic computing
Electron density map
$\rho(xyz) = \frac{1}{V}\sum_{hkl\,-\infty}^{+\infty} |F(hkl)|\,.\,e^{-2\pi i[hx+ky+lz-\phi(hkl)]}$
Amplitudes
Phases???

Figure 2 The diffraction experiment and the phase problem

X-rays produced at bright sources such as synchrotrons are used to perform diffraction experiments with protein crystals. However, in the process of recording the experimental data, the phase information is lost. Phases must be retrieved in order to obtain the electron density map of the structure that produced the diffraction pattern.

1.3 Phasing methods

Given a set of experimentally derived amplitudes, only one chemically plausible structure in the crystal will be consistent with the data (Patterson, 1944, Giacovazzo *et al.*, 2011). Such was the assumption behind the first structure solution attempts. Some of these first structures corresponded to inorganic compounds, including those of sodium chloride, potassium chloride, and potassium bromide (Bragg, 1913).

It took many more years, until the early 50s, to apply the same principle to macromolecules. Since proteins were composed of peptide building blocks, the first studies on peptides, such as those on keratin (Astbury & Marwick, 1932), led Pauling and other colleagues to propose a model for the structures of the α-helix (Pauling *et al.*, 1951) and the β-sheets (Pauling & Corey, 1951). These first analyses were performed by fibre diffraction. Such experiments (Franklin & Gosling, 1953) allowed to propose the structure of the DNA double helix (Watson & Crick, 1953). Beyond the cases described, where it was possible to establish a model susceptible of being validated by the data, most phasing problems had to start from the Fourier analysis of the experimental intensity data.

One of the functions that can be directly computed without phase information is the Patterson function (Patterson, 1934). This corresponds to a Fourier transform calculated using as coefficients the square values of the structure factors and setting all phases to zero. The Patterson function corresponds to the product of the superimposition of two copies of the electron density in the unit cell, shifted by a variable translation. In other words it is a map of interatomic vectors. The function will have a trivial maximum for a translation value of [0,0,0], as this corresponds to the case in which every atom is superposed onto itself. However, for non-zero translations, its value will be significantly higher when it results from the superimposition of heavy atoms, as the Patterson function is proportional to the product of the atomic numbers of the correlated atoms. The use of the Patterson function allowed the solution of small structures containing one or few markedly heavier atoms, as their positions could be directly calculated, taking symmetry into account (Harker, 1936). From the initial heavy atom positions, approximate phases could be derived and in addition to the heavy atoms, the remaining atoms could be found in the electron density maps, iterating and improving them and the phases in the process. A prominent example of the use of heavy atoms for structure determination is Dorothy Hodgkin's determination of the structure of penicillin through the sodium and rubidium salts of benzylpenicillin (Hodgkin, 1949).

Small molecule crystals are composed of relatively few atoms and tend to be well ordered so that they diffract to atomic resolution (better than 1.2 Å) (Sheldrick, 1990). Consequently, the amount of independent diffraction data that can be measured is much higher than the number of parameters required to describe the positions of all atoms in the molecule to be determined. This overdetermination can be exploited as the possible sets of phases are not independent. The mathematical foundations for the phase relationships used in *direct methods* were provided in the early 50s (Harker & Kasper, 1948, Karle & Hauptman, 1950, Sayre, 1952). Computer implementations of direct methods, such as SHELXS (Sheldrick, 2008) dominate phasing in small molecule crystallography. Recently, charge flipping algorithms have been introduced, which, also easily solve equal atom structures up to 250 atoms, even when the resolution is not quite atomic (Oszlanyi & Suto, 2004, Sheldrick, 2015).

When the crystallized molecules are proteins, nucleic acids or their complexes, the situation is very different from that of small molecules. The determination of macromolecular structures imposes a series of intrinsic difficulties:

- Protein and nucleic acids are more complex, requiring a larger number of atoms to be simultaneously found. Also, frequently, several copies of them are found in the asymmetric unit (ASU). Consequently, a larger set of parameters is unknown.
- Macromolecular crystals have a high solvent content. The disordered solvent does scatter the X-rays, but not in the same way as the ordered atoms, because it does not share the periodicity, and this must be taken into account appropriately. Moreover, this makes macromolecular crystals more fragile and prone to radiation damage and mechanical damage upon handling, which can further reduce their diffraction properties.
- Because of this high solvent content, mobile parts of the molecule, such as side chains or loops, can present differences among equivalent copies in the crystal. This causes weak diffraction data to be less distinguishable from background scattering from the solvent and breaks the diffraction at a given resolution, so macromolecular crystals tend to diffract to a lower resolution.

One of the possible approaches to solve protein structures is to take advantage of their high solvent content, by including heavy atom salts (platinum, gold, mercury, uranium cations or complex ions) in the solution in which the crystals are kept. Diffusion into the crystals through the solvent channels leads to selective incorporation of these species in particular positions of the macromolecule. If diffraction patterns are recorded with and without this *soaking*, the local changes in the structure and consequently the differences in their structure factors can be used to solve the heavy atom substructure. Then, phase information can be derived from the native macromolecule trough trigonometric relations relating the recorded data and the heavy atom structure factors. This technique is called Multiple Isomorphous Replacement (MIR) (Harker, 1956), and usually required several derivatized datasets apart from the native. MIR was used to determine the first protein structures at the LMB-MRC in Cambridge, those of Myoglobin and Haemoglobin (Green *et al.*, 1954, Kendrew *et al.*, 1958). In some situations, a single derivative can be enough, and then the technique is known as Single Isomorphous Replacement (SIR). SIR can be emulated by radiation damaged induced-changes in RIP phasing (Ravelli & Garman, 2006).

Another way to exploit differences in the data is to produce various diffraction patterns out of the same crystal but collecting the data at different wavelengths. In this case, it is the difference between the experiments that can be exploited to determine the substructure of the anomalous scatterers and to establish the relationships to the phases of the macromolecular structure. This method is named *Multiple wavelength Anomalous Diffraction* (MAD) (Hendrickson, 1991). It also has an analogous *Single wavelength Anomalous Diffraction* (SAD). The elements that can be exploited for SAD or MAD must also be electron rich. A number of proteins exist that contain such elements already in their native form: iron, zinc or molybdenum being the most frequent. The most common approach, however, implies substituting methionine by seleno-methionine in recombinant proteins. The advantage versus soaking with heavy atoms is that it will preserve more the isomorphism. Sulphur has a weak anomalous signal at most accessible wavelengths but is naturally present in cysteine and methionine. Its presence can be used in what is called *Native Phasing* (Hendrickson & Teeter, 1981*)*. Lastly, anomalous scattering can be combined with isomorphous replacement, either in SIRAS (*single isomorphous replacement with anomalous scattering*) or MIRAS (*multiple isomorphous replacement with anomalous scattering*).

The advances in synchrotron beamlines, including more sensible detectors, optimisation for long wavelengths, tuneability of their setting, and goniometers with large range of orientations have made experimental phasing in general, and SAD in particular, a popular option for *de novo* structural determination of proteins (Rose *et al.*, 2015, Hendrickson, 2014).

1.4 Molecular replacement

The fact that similar sequences lead to similar structures (Clothia & Lesk, 1986), opened the door to the use of models to solve new unknown structures. MR was generalised by Michael Rossman (Rossmann & Blow, 1962, Rossmann, 1972, Rossmann, 2001), and originally, it was the term used to refer both to the use of non-crystallographic symmetry within one crystal (NCS averaging), between different crystals (cross-crystal averaging, CCA) and to the use of a model structure to compute approximated phases (what we actually call MR).

MR is based on placing in the unit cell an atomic model of a homologous protein to the target structure, to provide a set of starting phases to get an approximate electron density map for the new structure.

The selection of an appropriate model is a crucial step in MR. Experimentally determined structures in the PDB databank (Berman, 2008, Berman *et al.*, 2013, Berman *et al.*, 2000) provide a source of templates for MR. A good model should represent accurately the target structure (have a low root mean square deviation, r.m.s.d.). To find potential models, sequence-based searches can be performed. These comparisons can also guide further modification in the model to remove divergent regions.

Once a suitable model is available, MR comprises two main aspects: (1) the search procedure, in which orientations and translations of the model are sampled within the ASU, and (2) the scoring procedure, in which the best match between model and target has to be discriminated by comparing the computed structure factors from the placed model with the experimental ones. The MR search is commonly performed as separate 3D rotation and 3D translation searches. Since MR entails a search procedure, its success is subject to the signal to noise ratio of the correct placement, which in its turn, depends on the quality of the model and the data.

While the original scoring functions were Patterson-based (Huber, 1965), currently the most successful targets are formulated upon a Maximum Likelihood (ML) basis. As described previously, the Patterson function can be computed for the observed data and for a model, allowing the comparison between them. For a correct orientation and position of the model in the unit cell, the two Patterson maps should correlate optimally. This is the basis for Patterson-based MR, which works very well when an accurate model for the whole structure is available. It is implemented in software such as AMoRe (Navaza, 1994) and MolRep (Vagin & Teplyakov, 1997). Numerous targets can be defined within the Patterson MR, both in real and reciprocal space (Rossmann & Arnold, 2001). The most important difference between Patterson and ML methods is that ML accounts in its formulations for the prior knowledge about the system as well as the estimated experimental errors and model differences. When the search model shows larger deviations to the target structure, it is not complete, or if data presents pathologies, ML methods perform better because all the information about these situations is handled in its target functions (Read, 2001).

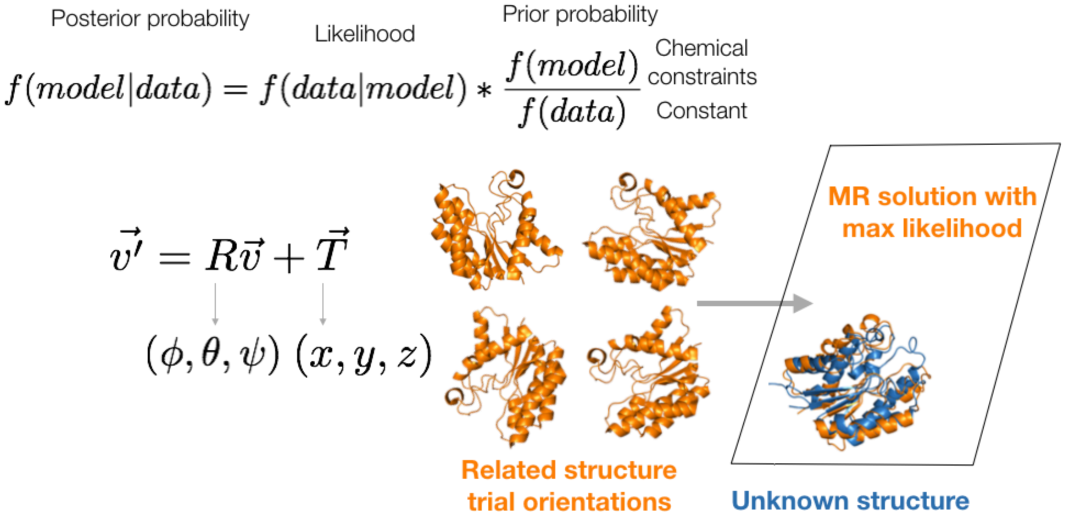

Posterior probability
Likelihood
Prior probability
Chemical constraints
Constant
$f(model|data) = f(data|model) * \frac{f(model)}{f(data)}$
$\vec{v}' = R\vec{v} + \vec{T}$
(ϕ, θ, ψ) (x, y, z)
MR solution with max likelihood
Related structure trial orientations
Unknown structure

Figure 3 Maximum Likelihood Molecular Replacement

In MLMR, the goal is to use the likelihood to compare the trial placements. The best model will be the one that best explains the data. The probability of the experimental data *f(data)* is constant and when comparing probabilities can be ignored. The probability of the model without having any data *f(model)*, is constant for all models as all proteins are subjected to the same chemical constraints.

Nowadays, the most widespread software for MR (Berman *et al.*, 2013, Scapin, 2013) is *Phaser* (McCoy et al., 2007), which is based on ML. Figure 3 shows a scheme of MLMR. The MLMR hypothesis in *Phaser* is the current orientation and placement or only orientation of the search component, within the background of the placement of the other components. *Phaser* compares the likelihood associated to a candidate placement to the one associated to the null hypothesis: a Wilson distribution of intensities coming from a random distribution of isotropically scattering atoms. The difference between the log of the likelihood of each hypothesis is known as the LLG (Log-Likelihood-Gain) and is the main indicator for recognizing a correct solution. The hypotheses are expressed in terms of intensities (LLGi), which allows accounting for both errors in the model and in the data, a formulation that has proven to be more sensitive and to remove bias in comparison with targets formulated in terms of structure factors amplitudes (Read & McCoy, 2016).

The contribution to the LLGi from any individual reflection depends on the resolution of the reflection, the fraction of the scattering that the model represents and the error expected from the model. Since correct placements must be identified within the intermediate steps, and more importantly, need to be preserved for further steps of the search, determination of which solutions to carry along is fundamental. It has been shown in recent work (McCoy *et al.*, 2017), that the LLGi is a direct measure of the probability of a placement being correct.

1.5 Phase improvement

Often, the initial density maps are noisy and difficult to interpret, and in the case of MR maps, present a model bias towards the structure that was used for the search. At this stage, an enhancement of the correct density and a reduction of the errors is required in order to solve the structure and be able to refine it correctly.

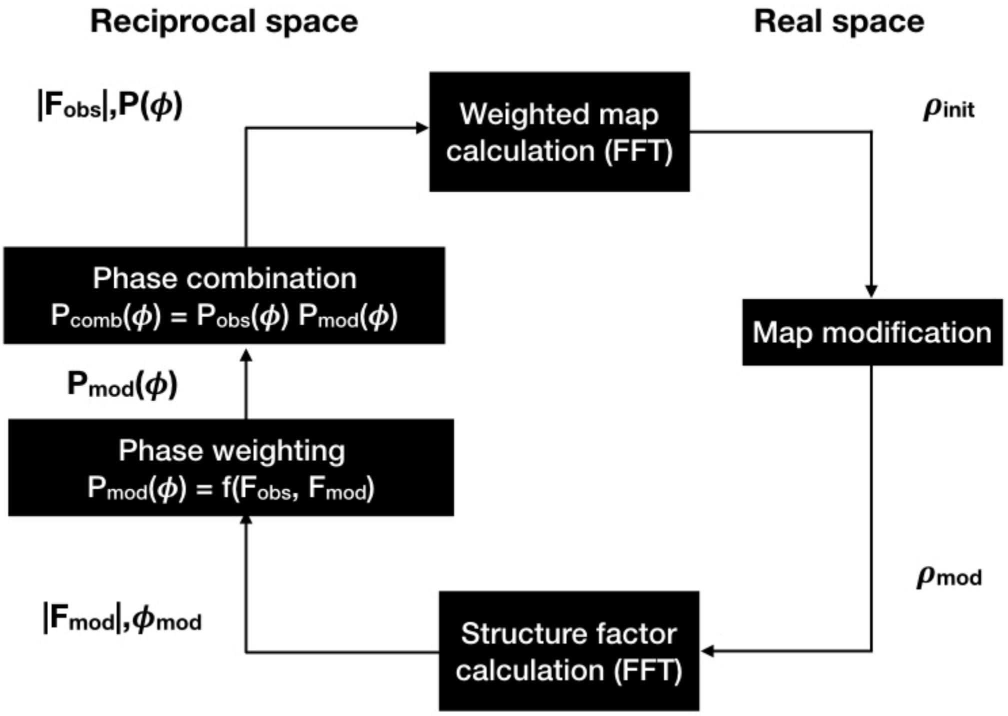
Density modification is a tool to improve the phase estimates from a starting set of experimental structure factors and phases. It encompasses methods that apply physically meaningful constraints or prior statistics to the electron density maps. Their input is the set of observed native magnitudes and the experimental phase estimates. As solvent areas in the crystal can be discriminated from very weak phase information, this is often the first step towards modifying phases. Assumptions about macromolecular structures can be used: they are composed of connected chains of atoms that pack in certain ways, and their maps present a similar electron density distribution when compared with other macromolecules. Part of this information is expressed in real space and part in reciprocal space, and consequently phase improvement calculations tend to iterate over both spaces, as shown in Figure 4.

Electron density properties, both overall and local, allow computing statistics that show patterns that can be imposed on the new structure. In traditional/classical density modification, a map is calculated from starting phases and modified, while in statistical density modification, the masks from the map are converted to a probability and it is that probability that is modified (Cowtan, 2010).

The use of *solvent flattening* (Wang, 1985) improves phases by locating solvent regions in maps and setting them to an expected mean value before combining the modified phases with the experimental ones. It takes advantage of features from the solvent regions. One of them is that the mean electron density should be lower in the solvent region than in the protein one. The other is that the variation of the density in the protein regions, where, at high resolution, sharp atomic features are present, should be much higher than in the solvent. In *solvent flipping* (Abrahams & Leslie, 1996) instead of just flattening, a relaxation factor is included that will flip the density in the solvent regions by a factor depending on the difference between the density expected in the solvent and the density being evaluated.

Protein maps, at a given resolution, have density distributions that are similar. That allows generating histograms of theoretical/expected distributions that can be compared with the map to be modified. This technique is known as *histogram matching* (Zhang & Main, 1990, Cowtan & Main, 1993). It is often performed in conjunction with solvent flattening and it can even be used at multiple resolutions (from lower to full) to improve map coefficients iteratively (Cowtan, 1994).

When several copies of a molecule are present in the ASU, they are necessarily related by some non-crystallographic symmetry (NCS) operation. NCS is a powerful constraint, which can be applied to modify the density, given that the different monomers or parts of the macromolecules and the operations relating them have been found. The local density can be modified by averaging. In practice, the more copies of NCS-related protein chains there are (and the less similar their symmetry is to the crystallographic one), the more regions of reciprocal space will be sampled, making averaging more powerful (Kleywegt & Read, 1997). This averaging can be exploited successfully in either experimental or MR maps (Vellieux *et al.*, 1995, Rossmann & Arnold, 2006).



Reciprocal space
Real space
|Fobs|,P(ϕ)
Weighted map calculation (FFT)
ρinit
Phase combination Pcomb(ϕ) = Pobs(ϕ) Pmod(ϕ)
Map modification
Pmod(ϕ)
Phase weighting Pmod(ϕ) = f(Fobs, Fmod)
ρmod
|Fmod|,ϕmod
Structure factor calculation (FFT)

Figure 4 Density modification in real and reciprocal space

The initial phase probability distribution is used to compute a centroid, obtaining the best phase and weight that can be used to compute a map. In real space, the map is modified with the chosen constraints. Then, the map is back-transformed to produce structure-factor magnitudes and phases. The agreement between the observed structure factor magnitudes and the ones from the modified map is used to estimate the error of the phases. Next, this error estimate is transformed into a probability distribution, that can be combined with the initial one. The whole process is performed iteratively. Note: image adapted from (Cowtan & Zhang, 1999)

The constraints that density modification imposes can be really powerful and allow to obtain a complete solution from only a very partial map from a small substructure (Cowtan & Main, 1993, Vellieux *et al.*, 1995, Foadi *et al.*, 2000, Sheldrick, 2010, Burla *et al.*, 2010, Usón & Sheldrick, 2018, Terwilliger *et al.*, 2009)

The automatic interpretation of the density-modified maps is an integral part of the process, the key to validate whether the map is becoming a better representation of the crystal. This interpretation can be done in the form of peptide tracing (Cowtan, 2006, Sheldrick, 2010, Perrakis *et al.*, 2001, Terwilliger *et al.*, 2008). The intercalation of cycles of density modification and map interpretation constrains phasing towards the correct solution. In each cycle, the phases of the trace are combined with the previous ones, giving rise to the next map to modify and interpret.

1.6 The ARCIMBOLDO framework

Extending the ease of structure solution for small molecules to proteins is an attractive objective. *Ab initio* phasing (from the native intensities alone) of macromolecular structures, with no experimental phase information or previous particular structural knowledge (a homologous model for MR), was, until recently, limited by structure size and resolution of the data. It has the advantage of having less experimental dependencies and less model bias, and it can be the most successful approach when the starting hypotheses are far from correct (for example, if the contents of the crystal are unexpected).

ARCIMBOLDO (Rodríguez *et al.*, 2009) combines location of model fragments such as polyalanine-helices with the program *Phaser* (McCoy *et al.*, 2007) and density modification (Sheldrick, 2002) and main chain autotracing (Sheldrick, 2010, Thorn & Sheldrick, 2013, Usón & Sheldrick, 2018) with the program SHELXE. It exploits assumptions about general fragments in proteins, in a similar way as dual-space recycling methods (Usón & Sheldrick, 1999, Rost, 1997, Miller *et al.*, 1994), used atomicity constraints that allowed to develop a complete solution starting from either random atoms or a very small substructure. The generality of the fragments for phasing has been also observed in other works (Glykos & Kokkinidis, 2003, Jia-xing *et al.*, 2005, Caliandro *et al.*, 2008). The program receives its name from the Italian painter Giuseppe Arcimboldo (1526-1593) (Figure 5).

Due to the difficulties in discriminating correct small substructures, many possible groups of fragments location have to be tested in parallel. Originally, massive computing was required in order to produce and keep enough hypotheses to the point in which they were recognisable (Rodríguez *et al.*, 2009). More recently, thanks to the description and the study of the expected value of the LLG (Oeffner *et al.*, 2018, McCoy *et al.*, 2017), the estimation of how difficult a case will be (given a particular model) has provided an inestimable source of information to guide fragment-based MR. Taken together with the improvements in the MLMR targets, for some cases, the computing requirements have been relaxed, and currently, single-workstation implementations of ARCIMBOLDO can routinely solve those. However, on the edge of difficult cases, massive computing is still required and supported by our software. In fact, the programs adjust their parameterisation on the number of hypotheses to follow depending on the available hardware. Difficult cases frequently need the systematic testing of ranges of parameterisations in order to succeed (Schoch *et al.*, 2015)

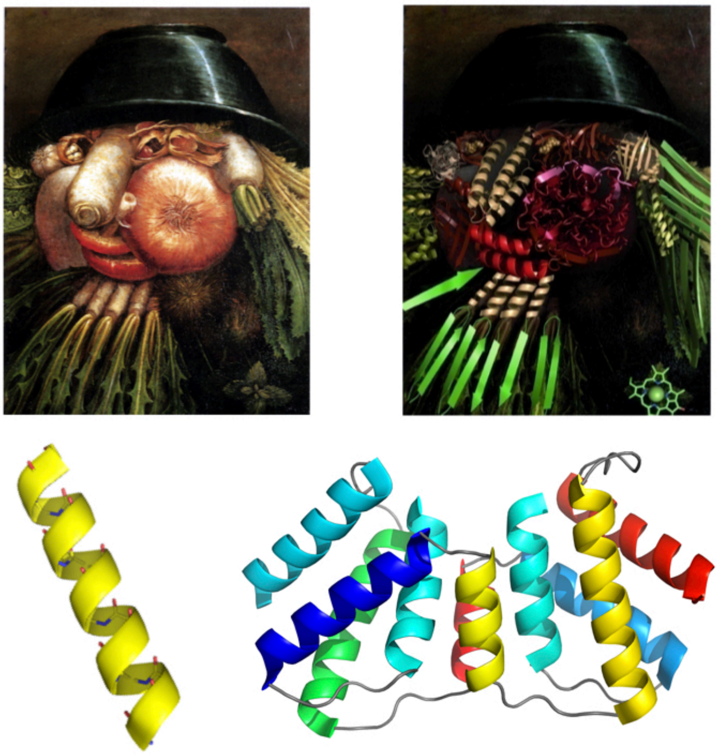


Figure 5 The ARCIMBOLDO framework

The method was named after the Italian painter Arcimboldo, who used to compose portraits out of fruit and vegetables. In a similar way, general fragments, such as ideal α-helices, can be used as approximations to true helices in structures. With ARCIMBOLDO, most collections of fragments remain a ‘still-life’, but some are accurate enough for density modification and main-chain autotracing to reveal the protein’s true portrait.

Beyond helices, other fragments can be exploited in an analogous way: libraries of helices with modelled side chains, strands, predictable fragments such as DNA-binding folds (Propper *et al.*, 2014), fragments selected from distant homologs (Millán *et al.*, 2018, Sammito *et al.*, 2014) or libraries of small local folds that are used to enforce nonspecific tertiary structure (Sammito *et al.*, 2013).

The main goal of the work presented in this thesis is to assist structure solution by the ARCIMBOLDO methods, exploiting the combination in reciprocal space of the phase information from partial solutions with ALIXE, and developing a successful approach to model generation and refinement from distant homologs with ARCIMBOLDO_SHREDDER. As each program is described in a separate chapter, and some introductory aspects are specific to them, at the beginning of those chapters a preface to the results is presented. The third chapter describes the solution of test and previously unknown structures relying on the abovementioned methods.

2 MATERIALS AND METHODS

2.1 Figures of merit

Figures of merit (FOMs) used in decision making trough the work described were *Phaser*'s LLGi (Read & McCoy, 2016) and the correlation coefficient between observed and calculated normalised intensities (CC; (Fujinaga & Read, 1987)) calculated by SHELXE (Sheldrick, 2002). Structure-amplitude-weighted mean phase-errors (wMPE; (Lunin & Woolfson, 1993)) were calculated with SHELXE against the refined models available from the PDB to assess performance.

2.2 External methods and programs

The work presented in this thesis has used the methods implemented in the program versions detailed in this section:

2.2.1 Phaser

http://www.phaser.cimr.cam.ac.uk/index.php/Phaser_Crystallographic_Software

Versions ranging from 2.5 and 2.8 from the CCP4 and PHENIX distributions.

Phaser (McCoy *et al.*, 2007) is required to perform the MR search of the fragment models. It is an MLMR software (Read, 2001). Within the work described in this thesis it is used for:

- Rotation function (Storoni *et al.*, 2004)
- Gyre rotation refinement (McCoy *et al.*, 2018)
- Translation function (McCoy *et al.*, 2005)
- Symmetry packing filter (McCoy, 2017)
- Rigid body refinement (McCoy, 2017)
- Gimble rigid-body refinement (McCoy *et al.*, 2018)
- LLG-based pruning (Oeffner *et al.*, 2018)
- Normal mode analysis (McCoy *et al.*, 2013)

The program allows specifying parameters for the MR search. The *Phaser* executable runs in different modes, which can be called either through a Python interface or as shell scripts (which is the way used in ARCIMBOLDO). The different modes can be parameterised through keywords. A full list of keywords can be found at http://www.phaser.cimr.cam.ac.uk/index.php/Keywords. Some examples include the resolution, the sampling size, the RMSD of the models, etc.

Part of the initial research described in this thesis relied on *Phaser* version 2.5, and at the time of the end of this work, the current *Phaser* version 2.8 has been used. The latest *Phaser* version has always been used in order to profit from its development. Nevertheless, original results have not been recalculated with the latest version. Therefore the particular version used is appropriately indicated along the text.

2.2.2 SHELXE

http://shelx.uni-goettingen.de/

Versions from 2013 to 2018 from the Shelx distribution server

SHELXE (Sheldrick, 2010) is required, to provide density modification based on the sphere of influence algorithm (Sheldrick, 2002) and for phase extension and main chain autotracing. The program allows to specify appropriate values or to accept defaults for a series of parameters including the number of cycles of density modification, resolution cut-off for start phases or data and solvent content autotracing (Usón & Sheldrick, 2018). A full list of the parameters is available at http://shelx.uni-goettingen.de/shelxe_keywords.php, or by launching SHELXE without any argument.

2.2.3 PHSTAT

PHSTAT is a FORTRAN prototype written by George Sheldrick that performs clustering of phase sets using a cyclical procedure. It takes a set of phase files in .phs format (*h*, *k*, *l,* F, FOM, PHI, sigF) as input and sets one of them as a reference. Then, it calculates the E- or F-weighted mean phase error (MPE) for each phase set, taking into account either the discrete or estimated origin shifts. Keeping the shifts with the lowest MPE, weights for each phase set are adjusted to minimise the MPE to the combined set, until convergence. Customizable parameters are selecting amplitudes or normalized amplitudes, the number of cycles (default 3), the reference file for clustering (default highest CC), the resolution limit for the phase sets (default 2Å) and the tolerance in degrees for MPD between the sets of phases to be clustered.

2.2.4 Molecular graphics

Model and maps were examined with Coot (Emsley *et al.*, 2010). Figures were prepared with PyMOL (Schrodinger, 2015).

2.3 Programming resources

The source code for all the distributed ARCIMBOLDO programs is organised in a modular structure, as described in Massimo Sammito's thesis dissertation (Sammito, 2015). Modules for input/output, grid connections, crystallographic symmetry and all required features are available. All of them are written in Python 2.7. A version compatible with the new standard, Python 3, is underway. ARCIMBOLDO has dependencies on scientific python libraries that are not included in the standard Python library, and for that reason we have chosen an Anaconda Python distribution (https://www.anaconda.com) as software environment. Anaconda comes with many data science and statistics packages, and, more importantly, it has a tool, conda, that serves as a virtual environment manager for Windows, Linux, and MacOS. It allows to test different versions/features of the libraries and automatically detects the interdependencies between packages.

In the work here described, ARCIMBOLDO used the following scientific libraries:
BioPython: to read/write/modify pdb files

matplotlib: for plotting
numpy: various numerical functions and use of its arrays
python-igraph: provides the graph data structure in which we express secondary and tertiary structure relations
scikit-learn: for clustering
scipy: for clustering and to compute correlations

2.4 Distribution of the software

At the time of this work, the ARCIMBOLDO programs are distributed either through the CCP4 (Winn *et al.*, 2011) suite or directly from our website. In our website, the current distribution is made available in form of frozen binaries generated with Pyinstaller, although this may change to distribution through the PyPI (Python Package Index) project. The ARCIMBOLDO frozen binaries are deployed for Linux and Macintosh current OS (Mavericks to High Sierra), and they are generated with Pyinstaller 3.3 and Python 2.7. The software is under the BSD 3 clause license.

2.5 Computing resources

Most structure solution and tests were run on a local HTCondor version 8.4.5. (Tannenbaum *et al.*, 2001) grid made up of 160 nodes totalling 225 GFlops. Submitter machines were 8 core workstations with 24GB RAM running Debian or Ubuntu Linux.

For the work described in this thesis, the remote grid and the supercomputing mode of the software were occasionally used. In particular, we accessed two remote grids, one at ALBA synchrotron (Barcelona) and another at the CIMR (Cambridge Institute for Medical Research, Cambridge).

Lastly, some of the tests described in section 5.7 were performed on a MacBook Pro (Retina, 15-inch, Mid 2015), with four 2.5 GHz Intel Core i7 and a RAM of 16 GB 1600 MHz DDR3.

2.6 Test data

Apart from the previously unknown structures described in chapter 5, a few other test structures from the PDB (Berman *et al.*, 2000, Bernstein *et al.*, 1977) have been used as test cases for initial development of the software, described in chapters 3 and 4 and are presented in the following paragraphs:

Test case PRDI

PRDII (PDB ID 3GWH) is a transcriptional antiterminator of the BglG family from Bacillus subtilis, which was solved *ab initio* with ARCIMBOLDO (Rodríguez *et al.*, 2009). The crystals belonged to space group $P2_1$, with unit-cell parameters a=37.39, b=65.75, c=38.19Å, β=109.58º. The ASU contains two copies of the monomer with 111 residues and 40% solvent, although the solvent content was deliberately increased to 45% in SHELXE runs. The data resolution is 1.95Å, and it is available as amplitudes. The structure comprises ten α-helices ranging from 11 to 20 residues.

Test case EIF5

Crystals of the C-terminal end (residues 232–431) of eukaryotic translation initiation factor 5 (EIF5) were obtained in space group $P2_12_12_1$, with unit-cell parameters a=32.23, b=71.08, c 80.64Å. The ASU contains one monomer of 185 residues and 42% solvent content, which was set to 45% in SHELXE. Data to 1.67Å resolution were available as amplitudes. The structure (PDB ID 2IU1) was originally solved by experimental phasing (Bieniossek *et al.*, 2006) and contains ten α-helices.

Test case MltE

MltE (PDB ID 2Y8P) is a bacterial outer membrane-anchored endolytic peptidoglycan lytic transglycosylase (Artola-Recolons *et al.*, 2011). Diffraction data to 2.0Å resolution as intensities were available. The crystals belonged to space group $C222_1$, with unit-cell parameters a=123.32, b=183.93, c=35.29Å. They contained two copies of the 194 amino-acid MltE monomer in the ASU and 45% solvent.

Test structure Xylose isomerase

The xylose isomerase from *Streptomyces rubiginosus* is a TIM-barrel protein for which in-house data were available to 1.54Å resolution as intensities. The space group is $I222$ and the unit-cell parameters are a=92.89, b=98.46, c=102.68Å. The ASU contains a monomer of 388 residues along with 50% solvent. A structure of the same crystal form with data to 0.99Å resolution is deposited under the PDB ID 1MNZ (Carrell *et al.*, 1994) and contains 15 α-helices ranging from 5 to 27 residues.

Test structure Brd4

The structure of the P-TEFb-activating protein Brd4 from *Mus musculus* (Vollmuth *et al.*, 2009) has been deposited in the PDB as entry 3JVL and data are available to 1.2Å resolution as amplitudes. The space group is $P2_12_12$ and the unit-cell parameters are a=52.06, b=73.05, c=32.30Å. The ASU contains a 120 residues monomer with 44% solvent.

Test structure ALR-MIA40

The 125 amino acid structure of the human FAD-linked augmenter of liver regeneration ALR (Banci *et al.*, 2011) is composed of a bundle of roughly parallel α-helices. Data to 1.9Å, in space group $C222_1$, and as intensities were available (PDB ID 3O55).

Test structure acylhydrolase

1YZF is a lipase/acylhydrolase from *Enterococcus faecalis* (unpublished, PSI initiative). The structure shows a central β-sheet flanked by α-helices. Data to 1.9 Å as amplitudes are available from the PDB, in space group $P3_221$, with unit cell parameters a=b= 45.92 and c=148.03Å. There is one monomer totaling 195 residues in the ASU, corresponding to a low solvent content of 36%.

Test structure Tom71

3FP2 is the crystal structure of Tom71 in complex with Hsp82 C-terminal fragment (Li et al., 2009). Data to 2.0 Å available from the PDB is in the form of amplitudes and belong to space group $P2_12_12_1$, with unit cell parameters a=47.86, b=116.29 and c=150.74Å. There is one monomer of TOM71 of 537 residues plus a 12 residues fragment of the chaperone, totaling 549 residues in the ASU, and corresponding to a solvent content of 63%. The structure is mainly helical.

Novel structure LTG

LTG (Lee *et al.*, 2018) is a soluble lytic transglycosylase from *Pseudomonas aeruginosa* (PDB ID 5OHU). Diffraction data collected at the ALBA synchrotron to 2.1Å resolution and as intensities were available. The crystals belonged to space group $P6_3$, with unit-cell parameters a=b=163.98Å, c=56.71Å. The ASU contains a monomer of 613 residues of the mainly helical structure, along with 61% solvent.

Novel structure PPAD

PPAD (Goulas *et al.*, 2015) is a peptidylarginine deiminase from *Porphyromonas gingivalis*. 20 diffraction data sets from different crystals were available, ranging from 2.97 to 1.5Å resolution. 16 of these, with unit cells of similar dimensions and rendering an average Rint of 0.37 and Rσ of 0.02, were combined. Data was used as intensities. The crystals belonged to space group $P2_12_12_1$ and contained one copy of the 432 residues monomer in the ASU, corresponding to a solvent content of 40%, which was set to 50% in SHELXE. The structure features short helices and twisted β-sheets along with a high proportion of coil.

Novel structure Hhed2

Hhed2 is a halohydrin dehalogenase from a gammaproteobacterium (Koopmeiners *et al.*, 2017, Schallmey *et al.*, 2014). Diffraction data collected at the ALBA synchrotron to 1.6Å resolution were available as intensities. The crystals belonged to space group $P2_12_12_1$, with unit-cell parameters a=78.02Å, b=94.86Å, c=140.27Å. The asymmetric unit contains four copies of a monomer, totalling 922 residues, along with 50% solvent content.

Pool with tests in 43 space groups

The remaining 43 structures constitute a pool of cases which were aimed to represent a wide spectrum of possible space groups for testing the generality of the method in different conditions.

The pool comprises cases with resolutions ranging between 1.35 and 2.3Å and sizes between 86 and 522 residues distributed in the ASU. Each structure belongs to a different space group, and up to 13 crystal classes from the 7 crystals systems are represented. 29 contain amplitudes and 14 intensities. Table 1 characterizes the test set.

Table 1 Test set for performance evaluation of SHREDDER and ALIXE

The test set covers a range of folds, resolutions, and size, but covering different symmetries was the main aim.

ID PDB	Solvent content	Crystal class	Space group	Resolution (Å)	Data are	Residue count
4DB5	59.53	23	$F23$	1.5	Intensities	125
3CYO	29.35	23	$P2_13$	2.1	Intensities	86
2G2D	45.95	432	$I4_132$	2	Amplitudes	193
5HGN	53.56	6	$P6$	1.9	Intensities	231
3OU2	47.97	6	$P6_1$	1.5	Amplitudes	218
3KWR	53.43	6	$P6_2$	1.4	Amplitudes	195
5G4Z	65.10	6	$P6_4$	1.9	Amplitudes	179
3MN2	42.39	6	$P6_5$	1.8	Intensities	216
5IX3	54.98	622	$P622$	1.8	Amplitudes	168
3HP4	61.44	622	$P6_122$	1.4	Amplitudes	185
2QCK	52.26	622	$P6_222$	1.9	Amplitudes	167
5O7G	63.10	622	$P6_322$	1.9	Intensities	345
2QCV	65.12	622	$P6_422$	1.9	Amplitudes	332
3T1S	37.91	622	$P6_522$	1.7	Amplitudes	136
2V71	56.50	2	$C2$	2.2	Amplitudes	378
4CSV	43.42	2	$I2$	2.0	Intensities	275
4J2F	55.50	222	$C222$	1.9	Amplitudes	223
5VOG	53.69	222	$F222$	1.5	Amplitudes	183
1V6T	47.24	222	$I222$	1.7	Amplitudes	255
3GH6	64.07	222	$I2_12_12_1$	1.7	Amplitudes	210
1S6Y	41.66	222	$P222$	2.3	Amplitudes	450
1SS4	58.52	222	$P222_1$	1.8	Amplitudes	306
2ODL	48.07	4	$I4$	1.9	Amplitudes	373
1NNH	50.91	4	$I4_1$	1.7	Intensities	294
3F4W	48.65	4	$P4$	1.6	Intensities	422
5M3Y	60.94	4	$P4_1$	2.3	Intensities	458
2QG3	39.01	4	$P4_2$	1.9	Amplitudes	416
4MH4	56.05	4	$P4_3$	1.9	Intensities	294
4ROT	57.96	422	$I422$	1.8	Amplitudes	268
2YG5	60.11	422	$P4_122$	1.9	Amplitudes	453
1UQ4	48	422	$P4_12_12$	1.9	Amplitudes	263
5W2G	45.09	422	$P4_22_12$	1.8	Amplitudes	257
5NA1	60.98	422	$P4_332$	2.3	Intensities	408
4PYI	37.97	1	$P1$	1.3	Intensities	221
2AIF	27.60	3	$H3$	1.9	Amplitudes	135
1O5J	39.45	3	$P3$	1.9	Amplitudes	113
3VPE	47.19	3	$P3_1$	1.6	Amplitudes	262
4CZL	51.47	3	$P3_2$	1.6	Intensities	348

3MYI	57.10	3	*R*3	2.2	Amplitudes	172
5H7E	63.94	32	*H*32	1.6	Intensities	182
5UDN	63.39	312	$P3_{1}12$	1.9	Amplitudes	522
2QX2	52.90	312	$P3_{2}12$	1.9	Amplitudes	344
2HYT	51.22	321	*P*321	1.6	Amplitudes	197

RESULTS

3 COMBINING PHASES FROM PARTIAL SOLUTIONS: ALIXE

3.1 Introduction

Fragment-based MR is a possible route to follow when a whole accurate template model is not available. The effect on the success of maximum likelihood phasing depending on the interplay among accuracy, completeness and available experimental data is now well understood, from single atoms to ribosomes, due to the work from the *Phaser* developers in (Oeffner *et al.*, 2018, McCoy *et al.*, 2017). The signal for an MR search can now be estimated before the calculation as the eLLG (the LLG expected for a correctly placed model). This value will depend on the quality of the model, its size, and the resolution of the diffraction data. Therefore, even with small fragments, if the model is very accurate and the resolution is high enough, correct solutions will be among those selected.

In general, in standard MR, when the LLGi score for a first fragment search is above 60, one can be confident that a correct solution has been found. This value comes from a study with thousands of MR trials, shown in Figure 6. In particular, a LLGi 10 times the number of degrees of freedom of the symmetry class is sufficient to be confident of success (so, 60 for non-polar space groups, 50 for polar, 30 for *P*1). However, for smaller values, both correct and incorrect solutions can be found. These solutions are not distinguishable until the signal increases, for example, after fixing a component and searching for the next one. Therefore, when phasing with small fragments, discriminating these partial, correct solutions, from incorrect ones is quite challenging, especially at early stages. One approach is to pursue many solution hypotheses in parallel until the point when they can be distinguished. In fact, even for inconclusive values, because there is a sigmoidal relationship between the chances of a solution being correct and the LLGi (Figure 6), solution lists with low LLGi values are likely to contain the correct solution (Oeffner *et al.*, 2018). But discriminating it from the incorrect ones is often not possible relying only on its FOMs. Combining information from different partial solutions is then an effective way both to increase the SNR, enhancing identification of correct solutions, as well as a way to complete the starting substructure to be expanded by density modification and autotracing.

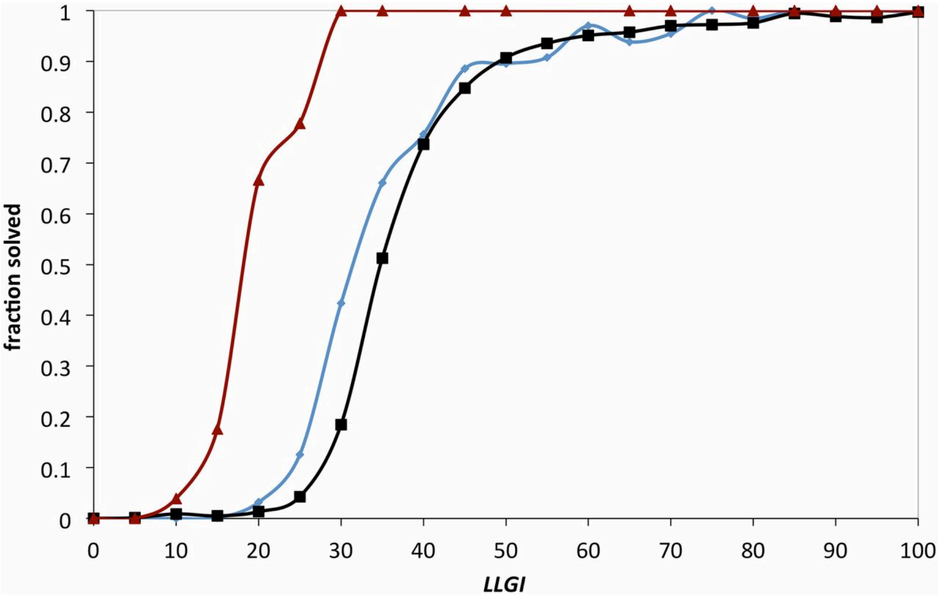


Figure 6 The LLGi score and the confidence in MR solution

Confidence in MR solution as a function of the final LLGi score. The final refined LLGi score provides a clear diagnostic for success in MR. The three curves show how the success rate for placing the first copy by MR varies with LLGi in 3 different space-group symmetry situations: *P*1 (only 3 rotational degrees of freedom; red; total of 263 MR trials), polar (3 rotational and 2 translational degrees of freedom, with an arbitrary origin along one axis; blue; 4,738 MR trials), and nonpolar (3 rotational and 3 translational degrees of freedom; black; 16,740 MR trials). Reproduced with permission from (McCoy *et al.*, 2017)

The comparison and merging of solutions can be undertaken both in real and reciprocal space. Either way, referring the solutions to a common symmetry-allowed origin is required.

In real space, one of the most successful examples is, in the context of MLMR in *Phaser*, the use of *amalgamation* (Bunkoczi et *al.*, 2013), which allows to combine partial solutions by merging their coordinates, checking their packing and rescoring, while considering the possible differences in origin imposed by the symmetry of the space group. In this case, for non-polar space groups, all origins are tried and roto translated solutions are merged, avoiding to perform successive searches or starting them from a more complete hypothesis. However, for polar space groups, only the rotation step is saved, and a translation function is performed. Another example of real space merging can be found in the structure solution pipeline FRAP (Shrestha & Zhang, 2015, Wang *et al.*, 2016). This method derives fragments from *ab initio* modelling and after clustering, tries to assemble a larger starting model, by using real space superpositions to a reference fragment for non-polar space groups and *Phaser*'s fast cross-translation Patterson function for polar ones. The use of these strategies is not limited to the phasing of proteins, with examples of successful iterative searches for phasing of small RNAs (Robertson & Scott, 2007, Scott, 2012).

Phased translations functions (Colman *et al.*, 1976, Read & Schierbeek, 1988, Bentley & Houdusse, 1992), both in their reciprocal or real space formulations, have been used for structure solution of challenging cases and provide a way to exploit partial information. The prior knowledge might come from experimental maps, from other models, or from copies of models in the ASU (Strokopytov *et al.*, 2005).

In reciprocal space merging of partial solutions, some of the first attempts published were those applied by Lunin and colleagues for low-resolution cases (Lunin *et al.*, 1995, Lunin *et al.*, 1990). They established how an average phase set derived from multiple partial solutions may be more precise than the individual phase sets, due to two aspects: the cancellation of errors and the weighting down of structure factors whose phases are not collinear between different models. However, in their attempts, phase sets to evaluate were generated randomly, whereas in the context of MR, particularly with small general fragments, both correct and incorrect phase sets might be correlated because of their genesis. In a more recent study with other collaborators (Buehler *et al.*, 2009), cluster analysis of MR solutions was explored. In it, solutions from differently parameterised rotation functions were grouped in real space, and the translations produced by their representatives, clustered in reciprocal space. The results suggested that clustering would produce a smoothened average between two solutions with partially correct and incorrect information, and how their average could be better than their single contribution. This kind of 'smoothened' average is difficult to get in real space, where the geometry of the fragments has to be taken into account. Plus, sets of phases with their amplitudes counterparts can be density modified, therefore benefitting from the improvement that such modification can bring upon the phase set.

In this work, reciprocal space clustering will be explored. The three ARCIMBOLDO programs (LITE, BORGES and SHREDDER) use very accurate yet incomplete models, placing them and providing starting phases that might be successful in the SHELXE expansion step. The type and size of these models makes them an ideal target for investigating the use of phase combination strategies.

In ARCIMBOLDO_LITE, where unspecific general fragments such as ideal α-helices are used, it is possible that in early stages, alternative correct locations are found. Phase combination can provide a way to reduce the need for subsequent searches after few fragments have been found by retrieving their coherent relations from the beginning. For this reason, both ideal and real solutions from ARCIMBOLDO_LITE were chosen to characterise the MPDs between overlapping and independent correct and incorrect fragments. This work is described from section 3.3 to section 3.4.

In ARCIMBOLDO_BORGES, instead of a single model, a library of models is used to represent a given geometry. These libraries contain several instances of variations of a given small local fold, as for example three antiparallel β-strands or two parallel α-helices. Such models, even if they constitute a tertiary structure fold, are general and unspecific, so they can fit and be located in multiple ways, including an overlapping manner in a given crystal structure. Therefore, reconstituting their overlap in reciprocal space might be a good strategy to complete the starting hypothesis while adjusting for geometrical deviations. This work is described in section 3.7.

In ARCIMBOLDO_SHREDDER, models are derived from an initial distant homology model. Since all of them are generated from the same starting template and will certainly have an overlap between them, they can also produce overlapping solutions. However, the adjustment through internal refinement of the models during the MR procedure has the potential to change their geometry in different ways, therefore making reconstitution in real space more complex. Instead, in reciprocal space, it would be possible not only to merge them but to weigh more the contributions of the most constant regions. This work is described in section 3.5.

3.2 Phase comparison measures

In order to compare phase sets, a measure of similarity is required. Multiple formulations are available, both local and global, but as the aim of the project is to compare whole sets, global measures will be described and used.

3.2.1 Mean phase differences

Phase errors can be expressed as the difference between two phases (Lunin & Woolfson, 1993):

$$\Delta\boldsymbol{\varphi}(\boldsymbol{h}) = \boldsymbol{\varphi_1} - \boldsymbol{\varphi_2} \qquad \textbf{(4)}$$

Phase differences (equation 4) vary between 0 and 180º. An average of all the phases will provide the mean phase differences (MPD). If instead of a simple arithmetic mean, weights are used to modify the contribution of each reflection, a weighted mean phase difference will be computed (wMPD, equation 5).

$$\textbf{wMPD} = \frac{\sum_{i=1}^{n} \boldsymbol{w_i}\, \Delta\boldsymbol{\varphi_i}}{\sum_{i=1}^{n} \boldsymbol{w_i}} \qquad \textbf{(5)}$$

If two given phase sets are not related at all, their wMPD will be close to 90º. If instead, they are almost completely equivalent, their wMPD will tend to 0º.

For the rest of the work presented in this thesis, the term mean phase error (MPE) will be used to refer to the error of the phases under study with respect to the true phases, and the term mean phase difference (MPD) for the differences between phase sets in general.

3.2.2 Map correlation coefficients

Correlation coefficients measure the direction and strength of a linear relationship between magnitudes. These magnitudes can be scalars, but also vectors. They can therefore be used to estimate similarity between phase sets.

The map correlation (Lunin & Lunina, 1996) measures the correlation between two electron density maps. If the two maps are coming from the same set of observed magnitudes, but with different phases, it can be computed by using the structure factor amplitudes and the difference in their phases, as shown in equation 6:

$$\boldsymbol{mapCC}[\boldsymbol{\varphi_1}(\boldsymbol{h}), \boldsymbol{\varphi_2}(\boldsymbol{h})] = \frac{\sum_{\boldsymbol{h}\in\boldsymbol{S}}[\boldsymbol{F^{obs}}(\boldsymbol{h})]^2\, \boldsymbol{cos}[\boldsymbol{\varphi_1}(\boldsymbol{h}) - \boldsymbol{\varphi_2}(\boldsymbol{h})]}{\sum_{\boldsymbol{h}\in\boldsymbol{S}}[\boldsymbol{F^{obs}}(\boldsymbol{h})]^2} \qquad \textbf{(6)}$$

Map correlation is one for identical phase sets and zero for uncorrelated phases. The contributions of each reflection to the mapCC can also be weighted, as shown in equation 7.

$$\boldsymbol{wmapCC}[\boldsymbol{\varphi_1}(\boldsymbol{h}), \boldsymbol{\varphi_2}(\boldsymbol{h})] = \frac{\sum_{\boldsymbol{h}\in\boldsymbol{S}} \boldsymbol{w}(\boldsymbol{h}) \left|\boldsymbol{F^{obs}}(\boldsymbol{h})\right|^2 \boldsymbol{cos}[\boldsymbol{\varphi_1}(\boldsymbol{h}) - \boldsymbol{\varphi_2}(\boldsymbol{h})]}{\sum_{\boldsymbol{h}\in\boldsymbol{S}} \boldsymbol{w}(\boldsymbol{h}) \ \ [\boldsymbol{F^{obs}}(\boldsymbol{h})]^2} \qquad \textbf{(7)}$$

3.2.3 Modifications to the computation of MPD and MapCC

When phase sets are not random but only partially correct, the comparison between them can be improved by considering variations of the measures described above. In this work, the aim is to compare and merge sets that are related but that will have differences between them (because then complementary information will be obtained).

A first, obvious modification, is to cut the resolution to a lower value, so that detail is lost and a most general comparison can be performed. Another option is to weigh the contribution of each reflection according to the value of the intensities, taking or not their sigmas into account, giving more weight to the strongest reflections. On the other side, *normalised structure factors (E-values)* can be used. These can be obtained by dividing the square amplitudes by the mean value for reflections at the same resolution bin. In this case, using E-values will indirectly give more weight to high-resolution data, as all resolution shells are brought to one scale. Finally, weights based on the FOMs of the phases can be used. These FOMs measure the agreement between the computed and the observed structure factors.

3.2.4 Crystallographic symmetry considerations

Combining phase sets correctly requires relating them to the same crystallographic origin since MR solutions do not have an absolute reference. This generates a space group dependency in the way relative origins are constrained. The optimal, space group allowed shift that relates two solutions must be identified in order to compute the true difference between them. Non-polar space groups have a finite number of possible origin shifts. As a consequence, even if the correct origin shift could not be identified, for non-polar space groups it is possible to undertake the exhaustive calculation imposed by matching fragment pairs trying all possible origin shifts. As shown in Figure 7, more than 60% of the structures deposited in the PDB as of May 2018 belong to non-polar space groups, the most frequent space group for proteins, $P2_12_12_1$, being also non-polar. For polar space groups, the number of possible origins is unlimited as particular directions are unconstrained by symmetry. This means that either a very computing demanding sampling must be done in order to find the relative shifts or an approximation must be computed that allows estimating an optimal shift for such cases.

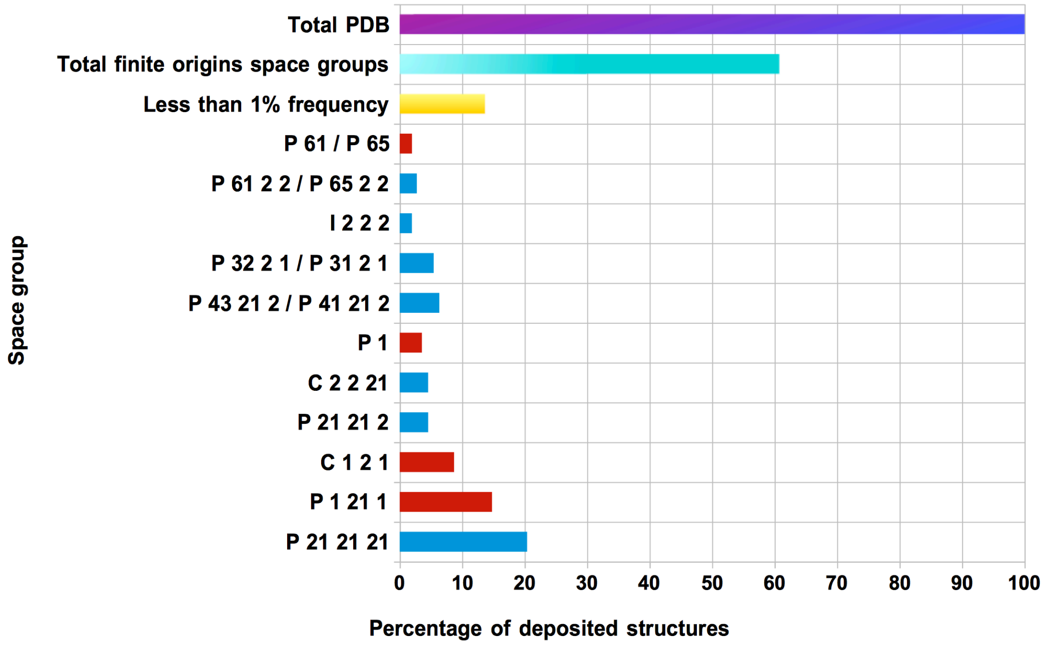


Figure 7 Space-group distribution in the PDB

Space-group distribution within the approximately 125916 crystallographic entries deposited in the PDB as of May 2018. Non-polar space groups are shown in blue and polar space groups in red. Non-polar space groups make up half of the total. $P2_12_12_1$ is the most common space group, occurring in almost one-quarter of the deposited entries.

3.3 Proof of principle on using phase combination from partial solutions

In order to develop a clustering procedure for phase sets derived from partial solutions, a study about the MPD as a metric for the phase comparison and merging was required. This was undertaken to characterise the MPDs between partial maps in both ideal, favourable case scenarios, and more realistic ones in which the phase combination might be tried.

The best scenario for the combination of partial solutions is one in which we are sure that our fragments are correct. For that purpose, tests with ideal secondary structure fragments, cut from the final structures, were performed (Millán, Sammito, Garcia-Ferrer*, et al.*, 2015). Main-chain α-helices are the most successful search fragments for *ab initio* phasing with ARCIMBOLDO. They are very rigid and constant, and when considering short helices, they are very similar across different structures, typically with a Cα r.m.s.d below 0.2Å. In order to explore the combination of perfect, yet partial, solutions, all helical fragments were extracted from three test structures (3GWH, 2IU1, 1MNZ). 2IU1 belongs to the most frequent among all Sohncke space groups (non-polar $P2_12_12_1$), 3GWH adopts the most common polar space group for proteins, $P2_1$. 1MNZ, a TIM barrel in non-polar space group *I*222, does not have an all-helical composition as the other two cases but contains a substantial fraction of β-strands.

In the tests, phases were calculated from each of the helices in the structure, after truncating residues to alanine and setting the B-factors to a common value, to emulate the search fragments used by ARCIMBOLDO_LITE. Phase sets were subject to a variable number of density modification cycles with SHELXE: 0, 5, 10, 15, 20, and 30 cycles. All remaining parameters were set to constant values, adopting the program defaults but for the solvent content, particular to each structure. Subsequently, all possible phase combinations derived from two, three and four fragments were calculated, determining the origin shifts leading to minimal phase differences. The evaluation of phase similarity and the phase combination tests were done using the PHSTAT subroutine from SHELXE (Sheldrick, 2002), described in section 2.2.3. PHSTAT performs clustering of phase sets by a cyclical procedure. For nonpolar space groups, it applies all of the allowed origin shifts to the phases and calculates the E- or F-weighted MPD for each case. This allows to examine the difference between the lowest and second lowest MPD obtained in order to assess whether a large difference would indicate a more reliable discrimination of the origin shift. For polar space groups, the symmetry-allowed discrete origin shifts are tested and an initial origin shift is estimated in the polar direction on a data subset and then refined against all reflections. Keeping the shifts with the lowest MPD, weights for each phase set are adjusted to minimize the MPD to the combined set until convergence. As all fragments are extracted from the final structures, a relative origin shift signals a failure in the clustering process.

The phase sets resulting from these combinations, were subject to further density modification cycles (0, 5, 10, 15, 20, and 30). Evidently, at this stage the relative origin shift has been fixed, but density modification might still evidence *a posteriori* discrimination of correct versus incorrect shifts.

3.3.1 An all helical structure in a non-polar space group, 2IU1

The structure of the carboxy-terminal domain of human translation initiation factor eIF5 (Bieniossek *et al.*, 2006) is displayed in Figure 8a. It contains 10 helices, whose extension, average B values and wMPEs relative to the deposited structure are represented in Figure 8b. Two of the shortest helices present a higher average B-factor than the rest of the helices in the structure.

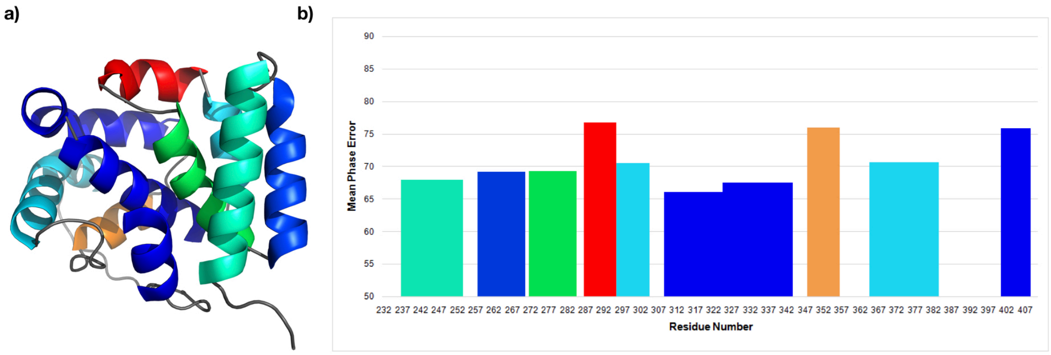


Figure 8 Characterization of the ten helices in the structure 2iu1

(a) Cartoon plot, with a rainbow color gradient representing the main-chain average B factor of the helices (red for highest and blue for lowest B). (b) MPE of the phase set obtained from each helix versus residue count, color coded according to the B values as in (a). The results are comparable except for two of the smallest helices.

For all possible binary combinations of helical fragments, adopting either fragment in the pair as reference, the relative F-wMPDs considering the 8 origin shifts allowed in the space group were computed. Values range from 81.3 to 89.9º. A CC was calculated for each file resulting from fusing both fragments after applying the selected origin shift. The fragment selected as reference may influence the outcome but in general results are consistent within each pair of fragments, and are thus displayed for a sparse set of all possible combinations. The effect of density modification on the discrimination between the two lowest MPD for each pair of fragments is illustrated in Figure 9. Negative bars indicate those cases where the correct origin shift would be missed, as it does not correspond to the minimum MPD.

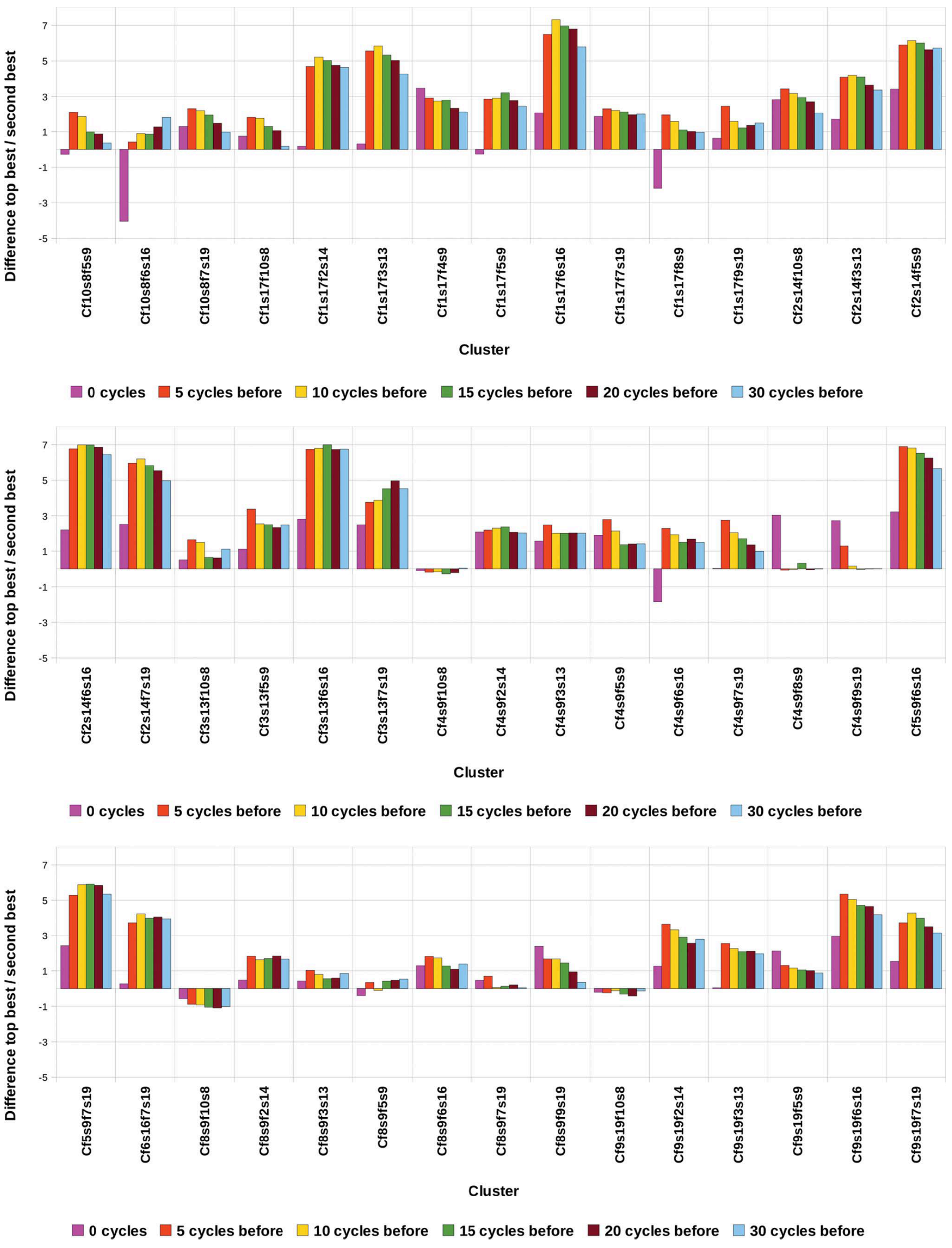


Figure 9 MPD difference between clusters of perfect helices from 2iu1

Fragments are labelled by their order in Figure 8 and the number of amino acids. The bars show the difference in degrees between the MPD corresponding to the correct origin shift and the MPD for the best-scoring wrong origin shift. Negative bars represent cases where a wrong origin shift yields the minimal MPD and would have been selected instead of the correct one. Density modification assists in the selection of the correct origin shift in six cases and generally improves the discrimination of the correct origin choice (pink bars versus the rest).

The correct origin shift would be unequivocally chosen in 39 of the 45 pairs, and even in 42 if the fragment in the pair characterized by a higher CC against the native data, usually the largest, was trusted as reference in case of discrepancy. This would seem the natural choice and solutions from fragment search could be sorted according to this FOM. The use of density modification enhanced discrimination, 4 cases did actually require it in order to avoid selecting a wrong origin shift, but the number of density modification cycles did not determine the outcome. The three cases where the correct origin shift was missed involve as common fragment the smallest helix in the structure (8 residues), in combination with two other small helices and a distorted one. Still, there are some of the clusters that even if correctly matched, present absolute MPD differences among different origin choices, below one degree, indicating that the correct shift could have been accidentally chosen but would hardly be trusted if the structure was unknown. This occurs in other matches involving the smallest helices. Also those cases where discrimination is clearer tend to contain common fragments, such as the long sixth helix (16 residues), whose scattering contribution is enough to generate a non-random map yielding the lowest MPE against the final structure.

As density modification helped to reveal the common origin, the effect of the number of cycles on the resulting MPE versus the final structure was assessed (Figure 10).

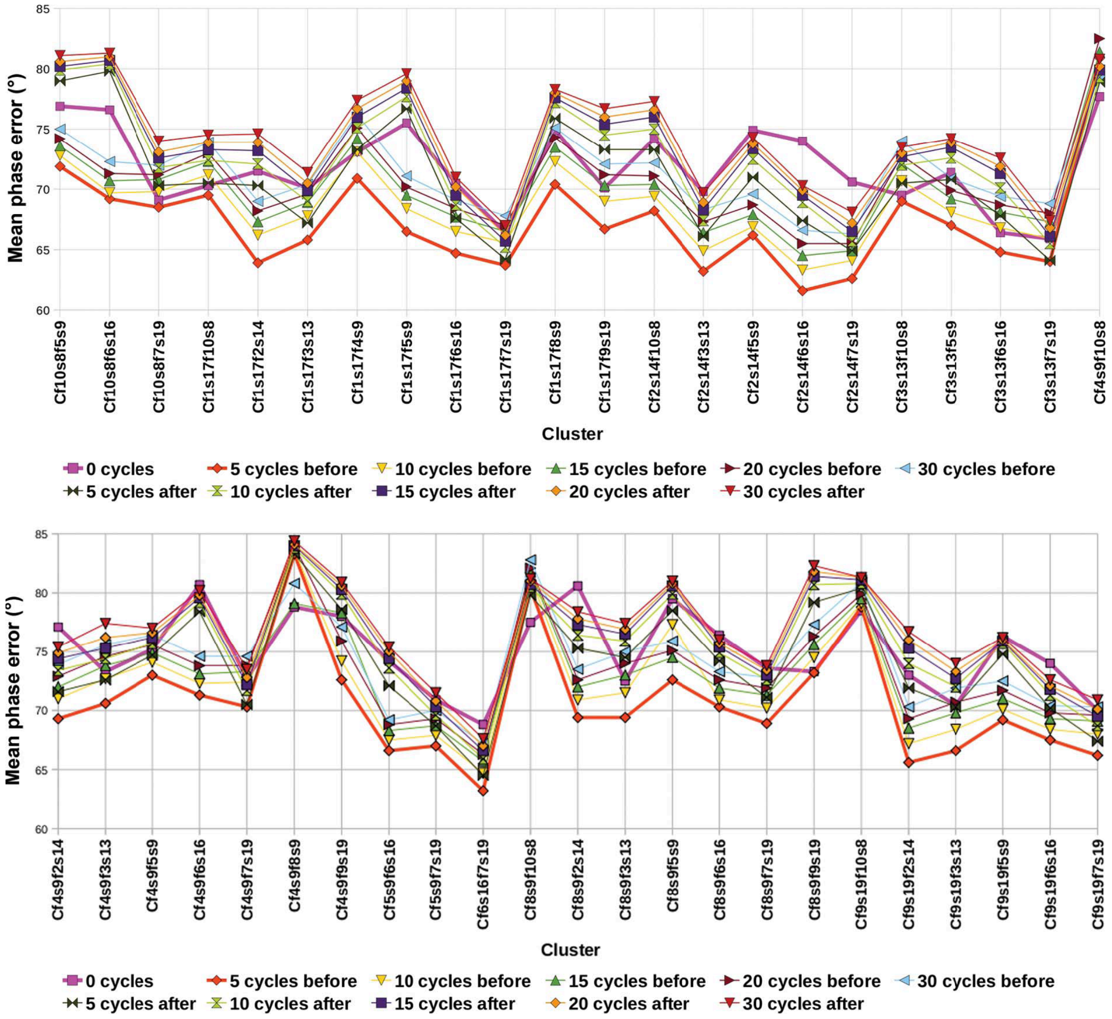


Figure 10 Effect of density modification on the 2iu1 binary clusters

Effect of density modification applied at different stages on the MPE of clusters combining perfect helices extracted from 2IU1 with respect to the true phases as calculated from the deposited structure. The orange line, representing five cycles of density modification, always leads to the best phases with the lowest MPE.

The general trend, even when the phase information derived from fragment pairs is not yet enough to solve the structure, is that application of 5 cycles of density modification prior to phase combination renders the lowest MPE (43 of 45 cases). Exceptions correspond to a very small MPD difference value and involve helices barely 10 residues long. Alternatively, using the E-weighted, rather than F-weighted MPD to cluster these perfect fragments led to the correct origin shift being identified in all cases.

In the case of ternary combinations, two relative origin shifts need to be determined. Clustering all possible combinations, considering each of the three fragments in a set as reference in turn, generates 360 sets. Density modification enhances correct origin discrimination from 254 to 304 sets, representing 60 and 80% of the cases, respectively. Failure occurs only for clusters gathering at least one of the three smallest helices, of 9 residues. Figure 8b shows these fragments render the phase sets with highest MPE. In practice, such small fragments are of limited use within ARCIMBOLDO, unless very high resolution data is provided. Clustering using as reference the best fragments in each triplet increases success rate. Sorting fragments according to their CC and taking as reference the top CC for a triplet would appear advisable, as fragments characterized by better FOMs (LLG or CC) are more likely to be correct. Consistent clustering of a triplet when alternative references are used and discrimination between lowest and second lowest MPD obtained for all allowed origin shifts can be used to evaluate confidence in origin choice. Figure 11 plots graphical results for the combinations of phases derived from three fragments subject or not to density modification. Helix 8, which was found to fail in most clustering tests, yields a negative CC (-0.14%) and high MPE and is omitted in the figure. In all remaining cases, applying density modification prior to clustering enhances origin shift recognition.

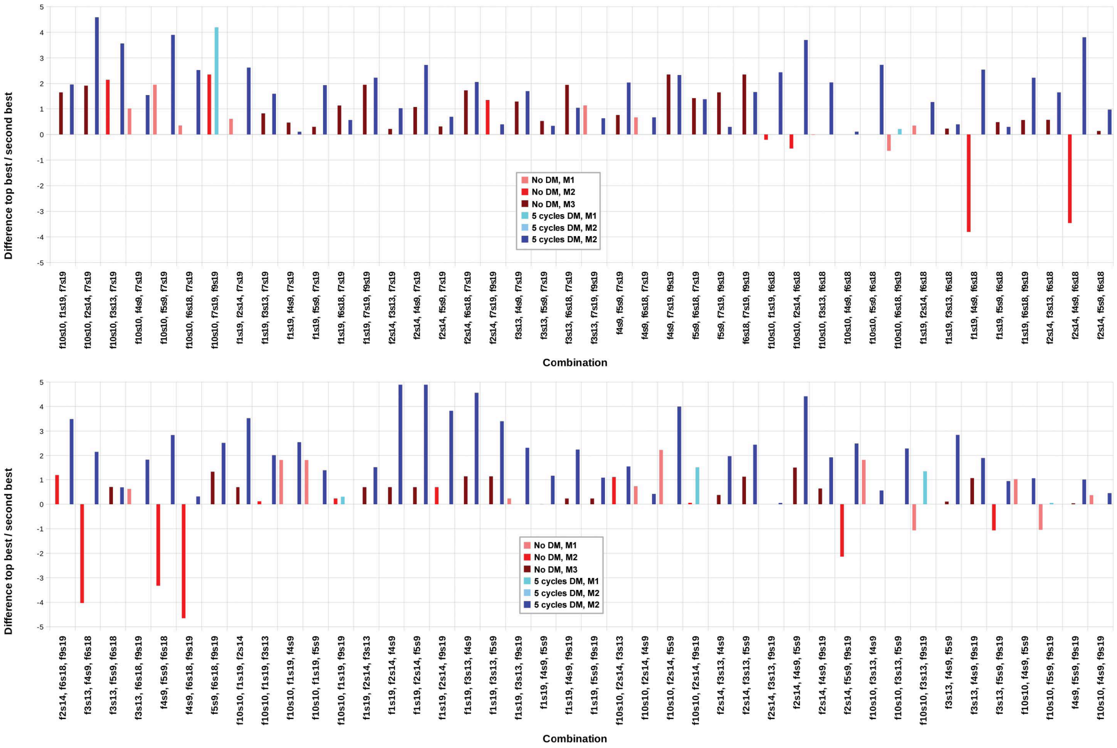


Figure 11 MPD difference between ternary clusters from 2IU1

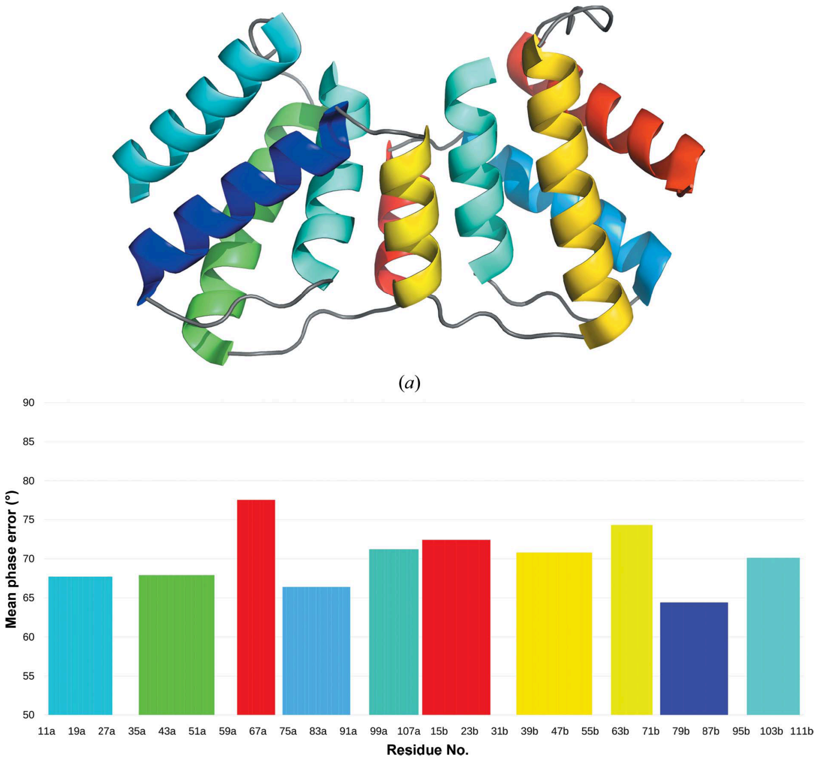
MPD difference between clusters combining three perfect helices from 2iu1 for the two origin shifts corresponding to the lowest and second-lowest MPD. Fragments are labelled in the order indicated in Figure 8 and by the number of amino acids. The bars quantify the difference in degrees between the MPD corresponding to the correct origin shift and the MPD for the best-scoring wrong origin shift; shades of red represent no density modification and shades of blue represent five cycles of density modification; the darker the colour, the higher the consistency among different reference choices. Negative bars represent cases where a wrong origin shift yields the minimal MPD and would have been selected instead of the correct one. With density modification (blue bars) the correct origin is always identified.

In the case of quaternary combinations, origin recognition succeeds every time and for all conditions MPEs decrease. As more correct fragments are being added, phases improve and determining the correct origin becomes easier; as expected the structure can be reconstructed from so many correct fragments. The general trend of optimal results when applying 5 cycles of density modification prior to clustering is maintained for helices of a minimum length of 10. Possibly the early discrimination of protein and solvent regions brought about by density modification enhances correct clustering.

The experiments with the 1MNZ structure produced comparable results to this case, and therefore, are not reproduced in this thesis, but are presented in detail in the reference work (Millán, Sammito, Garcia-Ferrer, *et al.*, 2015).

3.3.2 The transcription antiterminator PRDII (3GWH) in a polar space group

The structure is composed of 220 residues, diffracts to 1.95 Å and belongs to the most common polar space group in the PDB, $P2_1$. 10 helical fragments ranging from 11 to 20 residues were extracted from the deposited structure. A cartoon characterizing the helical fragments and a graph showing their characteristics and MPE for the phases they originate are displayed in Figure 12.



(*a*)

(*b*)

Figure 12 Characterization of the ten helices in the structure 3gwh

(a) Cartoon representation, with a rainbow colour gradient representing the main-chain average B factor of the helices (red for highest and blue for lowest B). (b) MPE of the phase set obtained from each helix versus residue count, colour-coded according to the B values as in (a). The monomers in the asymmetric unit present different B factors; for equivalent helices, the lower the B-factor average the better the MPE.

The same procedure described for the previous case was followed for this structure, except that along the direction of the monoclinic symmetry axis, the possible origin shift is not constrained. This precludes the differences among possible origins as a criterion as it would involve additional intensive calculations. When calculating the

CC for a structure composed of the fused fragments, their positions were allowed to refine locally. The correct origin is selected in over one-quarter of the combinations (26 out of 90 possible combinations). The model selected as reference influences the outcome, so that the resulting shift may differ within a pair. Sorting after MPD and selecting the 20 pairs where origin shift was equivalent regardless of the reference chosen allowed to identify 10 reliable clusters. Correct matches are furthermore characterized by relatively lower MPD and higher CC, although the difference is not clear-cut. The fact that origin shift is unconstrained in one of the directions was expected to hinder discrimination and even with perfect fragments, the correct combinations are often missed. All preliminary tests with this structure using placed model fragments in real searches failed to recognize the correct origin shift relating correct solutions. In view of these results, triclinic symmetry was not further investigated.

3.3.3 Density modification enhances origin shift recognition

From the studies with perfect fragments extracted from test cases, it became evident that applying a few cycles of density modification prior to clustering enhances both recognition of the origin shift and improvement of the clustered phases to be expanded in autotracing. As the partial solutions constitute approximations to the complete structure, the improvement brought about by solvent flattening possibly enhances the approximation to the correct phases and aids recognition. In the case of nonpolar space groups, evaluating the F-weighted MPD among fragments for all possible origin shifts allows the discrimination between the lowest and second-lowest MPD to be relied on. Fragment combinations characterized by larger differences can be trusted to be correct for fragments that are large enough to be located in a search.

3.4 Partial solutions from ARCIMBOLDO_LITE

Moving on from perfect fragments extracted from final structures, the next logical step was to work on partial solutions generated in an ARCIMBOLDO_LITE run which uses ideal polyalanine helices as search models. A model helix can fit a structure at different, partially overlapping positions but a real case will yield incorrect as well as correct locations so that search solutions can be wrong or right, related or independent. In the following test cases, partial solutions from an ARCIMBOLDO_LITE run with default parameters using as search fragments model helices of 14 alanine residues were clustered in reciprocal space. The conclusions derived from the tests on perfect fragments were assumed, that is 5 cycles density modification were applied to single fragments prior to phase combination and discrimination among lowest / second lowest MPD was assessed for non-polar space groups. The aim was to reach a solution after just one round of fragment placement, in cases where the structure cannot be solved expanding from a single fragment. All tests described in this section were performed using *Phaser* version 2.6.

3.4.1 The all-helical structure in non-polar space group, 2IU1

The run yielded 128 different fragment locations originating from 4 distinct rotation clusters. Three of them are correct. Clustering phase sets produced after 5 cycles density modification within 60º MPD, reduces the pool to 113, identifying positions related by an elongation of the helix regardless of their correctness. The extent of the elongation is typically of 1 to 3 residues, and the MPD indicates the value. The fused pdb files yielding highest CC within each cluster were selected as references to identify solutions originating from the remaining rotation clusters that would blend within 87º MPD. In this case, this produces just 4 clusters that can be sent to SHELXE expansion. One of them corresponds to the combination of correct solutions and 10 cycles of density modification and autotracing render a solution, characterized by an MPE of 47.8º and CC of 47.70 %.

3.4.2 3JVL from a standard ARCIMBOLDO_LITE

This 120 amino acid structure, with data resolution up to 1.2Å, was used to test the effect of limiting the resolution of the data used in solution clustering. A default limit of 2Å has been generally used, in order to assess the suitability of the method at least up to this resolution limit. On the other hand including the higher resolution might highlight differences and thus be detrimental. Initially, resolution limits of 2.0, 1.8, 1.5 and 1.2Å were compared. The ARCIMBOLDO run yielded 98 different fragment locations originating from 4 distinct rotation clusters. Two of the solutions, characterized by wMPEs computed at full resolution of 73.7 and 75.4º, were clearly correct, and a third one corresponding to one of the previous helices backtraced (MPE 83.7º). Two non-random solutions, with partial overlap to correct fragments, were characterized by MPE around 86º. Clustering phase sets produced after 5 cycles density modification within 60º MPD, reduced the pool to 90, identifying positions related by an elongation and one corresponding to a slight turn of a helix. These elongated solutions are consistent but wrong. Indeed, given the small number of correct solutions and the fact that they are not independent, consistency is not an indication of correctness, which is only revealed at the expansion stage in case of a successful solution. The two helical fragments placed at the same position in reversed directions are not clustered as their mutual MPD is around 68º (depending on the resolution) but this value is far from those produced by independent helices. After combination of matching solutions into single coordinate files, all fragments were sorted by CC within each rotation cluster and sequentially used as a reference to identify solutions originating from the remaining rotation clusters that would blend within 87º MPD. Evaluating all possible 90 clusters through SHELXE expansion, the ones corresponding to the combination of both correct solutions, the one placed backwards and a few other solutions including the non-random ones, after 10 cycles of density modification and autotracing succeed in solving the structure (MPE of 19º and CC of 46%). These solutions are joined whenever one of the non-random fragments is used as a reference, although the spurious fragments also included may vary. Nevertheless, correct solutions are distinguished since their MPD differences for best/second best origin shift are markedly higher. The same results as to fragment selection and numerical differences for MPDs are observed at each of the 4 resolution limits tested, thus the default of 2.0Å is adopted. Systematically testing the effect of the resolution limit from 2.0 to 10 in 0.1Å steps led to correct origin shift and gradual lowering of the MPD from 85 to 77º, preserving correct clustering until 5.4Å. Below this resolution, the correct shift is always missed. The disadvantage is that as the resolution is lowered, an increasing number of wrong solutions is drawn into the cluster, preventing successful expansion. In contrast, at 2.0Å, the reverse helix stands out through its MPD discrimination.

3.4.3 3O55 from a standard ARCIMBOLDO_LITE

Data to 1.9Å from this 125 residues structure was used in an ARCIMBOLDO_LITE run. The run yielded 79 different fragment locations originating from 2 distinct rotation clusters matching true rotations. After translation search, two positions were correct. The second being an elongation of the other, they belonged to the same rotation cluster. Clustering phase sets produced after 5 cycles density modification within 60º MPD, reduced the pool to 57 clusters, identifying positions related by elongation or slight translation. In this case, the fused solution from the 60º clustering, characterized by a CC among the top 5, may already be developed into a solution after 27 cycles density modification and autotracing (MPE of 57º and CC of 30%), whereas none of the single fragments succeeds. As these are the only two correctly placed fragments, further clustering with a high MPD threshold does not yield any additional solutions.

3.5 Partial solutions from fragments from distant homologs: MltE

MltE (Artola-Recolons *et al.*, 2011) (2Y8P) is a bacterial endolytic peptidoglycan lytic transglycosylase, from *E. coli*. Crystals belong to space group C222$_1$ and contain two copies of the mainly helical, 194 amino acids MltE monomer in the ASU. This structure was first solved with 2 Å data, using the SHREDDER sequential approach (Sammito *et al.*, 2014) with the structure of Slt70 (PDB-ID 1QTE) as the starting template. Previous MR attempts had failed using this nearest homolog to extract a search fragment, which displayed an r.m.s.d of 3.1Å over 160 Cα atoms. Instead, smaller partial fragments were required for successful placement. ARCIMBOLDO succeeded using, as an alternative to single helices, a set of partially overlapping, small models generated by systematically omitting parts of the loop trimmed 1QTE structure.

Originally, placement of two copies and trimming the solution amino acid by amino acid scoring against the CC (Sheldrick & Gould, 1995) was needed to solve the structure. Alternatively, combining information from partial structures after searching for one monomer should be a suitable strategy, cutting down the computing effort associated to searching for a second copy to complete each of the putative solutions yielded for each model. For this study, 90 polyalanine models were created by omitting blocks of 50 residues from the initial 140 in windows of one residue. From these 90 models, 5929 solutions were generated with *Phaser* (version 2.5) but most of them were incorrect, which is unsurprising given the high r.m.s.d of the models to the target structure. Phases generated from these models are random except for four sets, depicted in Figure 13a, characterized by wMPEs of 70.5, 71.1, 71.8 and 73.2º. They did not correspond to top solutions within their runs, as they occupied positions 25/25, 35/50, 17/95, 20/35 in *Phaser*'s LLG rank of solutions. This highlights how the central problem within the ARCIMBOLDO scope is often to extract extremely weak signals from the dominating noise. Combining these 4 sets succeeds in solving the structure while no isolated solution does, as displayed in Figure 13b, showing one of the solutions and the map obtained with the cluster. Clustering was accomplished in a two-step procedure. First, similar solutions from overlapping models placed on top of each other are identified and clustered. Then, complementary clusters corresponding to very different models or placed on different monomers are built. In this case where partial solutions with errors are combined, emphasising the high-resolution data is adverse and use of EwMPD rather than FwMPD leads to higher values by a small margin (up to 2º) and differences between top and second best become smaller. This has no practical consequences when identifying similar solutions but compromises detection of complementary fragments.

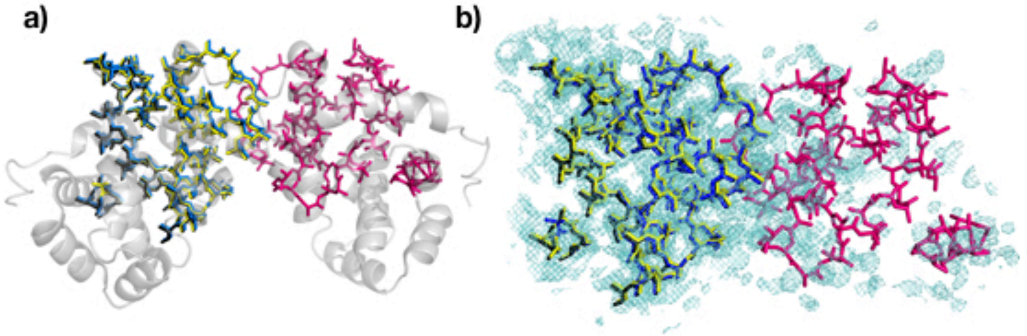


Figure 13 Structure of MltE

(a) Cartoon representation displaying on the same origin the four partial fragments combined in blue, black, yellow and pink, whereas the final structure is drawn as a grey cartoon. (b) Electron-density map generated from the phases of the three overlapping models (blue, black and yellow), showing some features of the missing monomer in pink.

3.5.1 Identifying similar solutions from different models

The procedure was started by ranking solutions according to the CC yielded by the fragment against the experimental data and generating phase sets from each coordinate file, subjecting them to 5 cycles density modification. Then, for the top 10% solutions, phases were clustered against all other files within 60° MPD. 550 clusters were obtained, each grouping 3 to 6 solutions. In practice, the 60° threshold is generous, as in general, clustered solutions are characterized by MPDs below 40°. They involve mainly contiguous models, differing in a few residues. At the end of this round, 3 of the correct solutions are joined in one of the 550 clusters formed, while the non-overlapping one remains single.

3.5.2 Combining complementary information

The clustered phase sets generated were used as seed for a further clustering round within 87° MPD in order to combine sets containing complementary information, like the 4 models represented in Figure 13a, covering parts of different monomers in the structure. One of the clusters obtained at this step, characterized by a difference between correct and second best origin shifts of 3.1º contained the combination of a previous cluster of 3 of the correct solutions described and the fourth one. Using the clustered phases as input, 3 cycles of SHELXE density modification and autotracing were able to discriminate a solution characterized by a trace of 165 residues with a

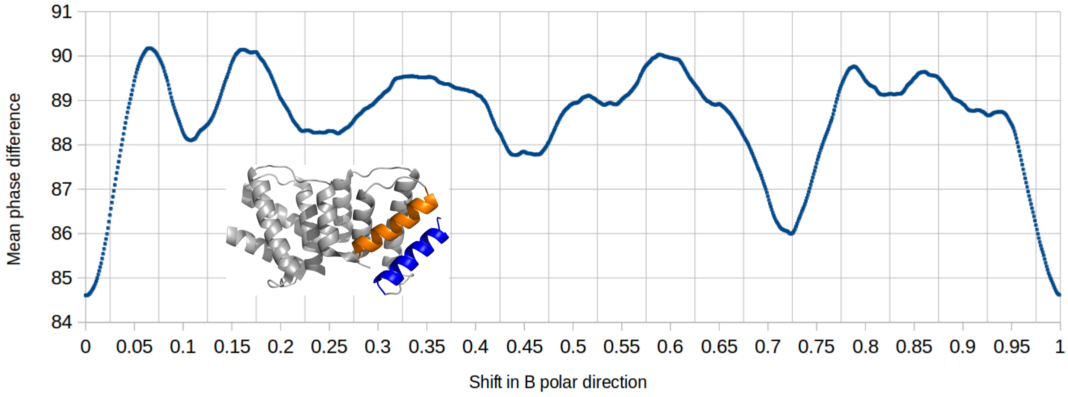
CC of 24.27%. More cycles can be applied to improve this solution up to a main-chain trace of 304 residues characterized by a CC above 45 after 9 iterations.

3.6 MPDs between independent and overlapping fragments

The studies performed with perfect and model helices, as well as with overlapping fragments, showed that pairs of partial independent correct fragments can render MPD barely better than random, around 87º. In the case of overlapping fragments, MPDs can range from 60º to 20º or less, depending on the degree of overlapping and the scattering fraction represented by the model.

As discussed previously, the approximation for estimating the origin shift in the unconstrained directions in polar space groups did not look particularly promising for independent fragments. In fact, all tests with independent helices either as perfect fragments or real solutions in an ARCIMBOLDO_LITE run were often missing the right origin shift. However, all of these represented very partial models that were placed in completely different part of the structures. The results with 2y8p indicated that partially overlapping fragments enhance the correct assessment of the origin shift. Therefore, a small experiment was performed to visualise the landscape for all the possible shifts in the unconstrained direction, for different pairs of fragment solutions.

In Figure 14 top, the MPDs between two perfect independent helices from the structure 3GWH, in space group *P*2$_1$, are shown. MPDs were computed for all possible shifts between (0,0,0) to (0,1,0), that is, for the non-constrained direction, a sampling of possible translations in steps of 0.0001 in fractional coordinates. In Figure 14 bottom, the same procedure was followed for 1JOV structure, in space group *C*2. In this case, perfect fragments were cut out from the final structure, but this time two sets of three antiparallel β-strands were chosen, in which two of them were shared by both fragments. This means that approximately 2/3 of the model is the same. As it can be seen from the figure, the landscape in the case of the independent fragments has a lot of minima that are relatively close to that of the correct shift (0,0,0) or its equivalent (0,1,0). The risk of choosing one of those local minima instead of the true one is high. In contrast, the value for the correct shift is much lower in the case of the overlapping fragments. It decreases very steeply for shifts close to the correct one, increasing the possibility of succeeding even if using an approximation.



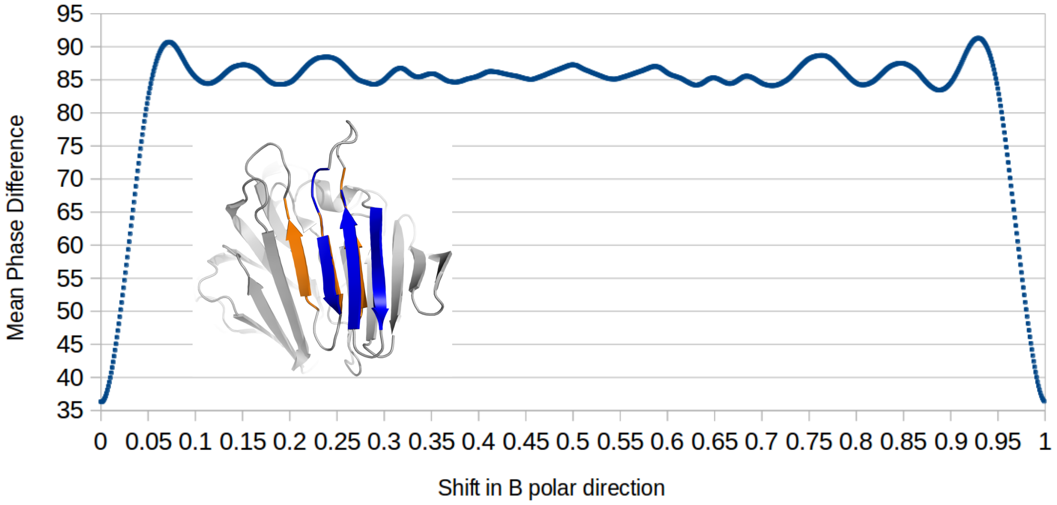


Figure 14 The landscape of MPDs along polar axes

> A sampling in steps of 0.0001 Å along the polar axis (b) was used, and the MPDs were computed at each 0,b,0 origin. The top figure shows the results of two perfect helical fragments from a structure in space group P2₁ and the bottom one, two sets of 3-stranded β-sheets that have an overlapping of two β-strands between them and that come from a structure in space group *C*2.

These results and the successful phasing of the MltE structure with a two-step clustering provided the base for an automated procedure implemented in our software for phase combination from partial solutions. If more than one monomer is expected in the asymmetric unit, the two steps strategy can be used. Otherwise, only a first step targeting overlapping fragments is followed.

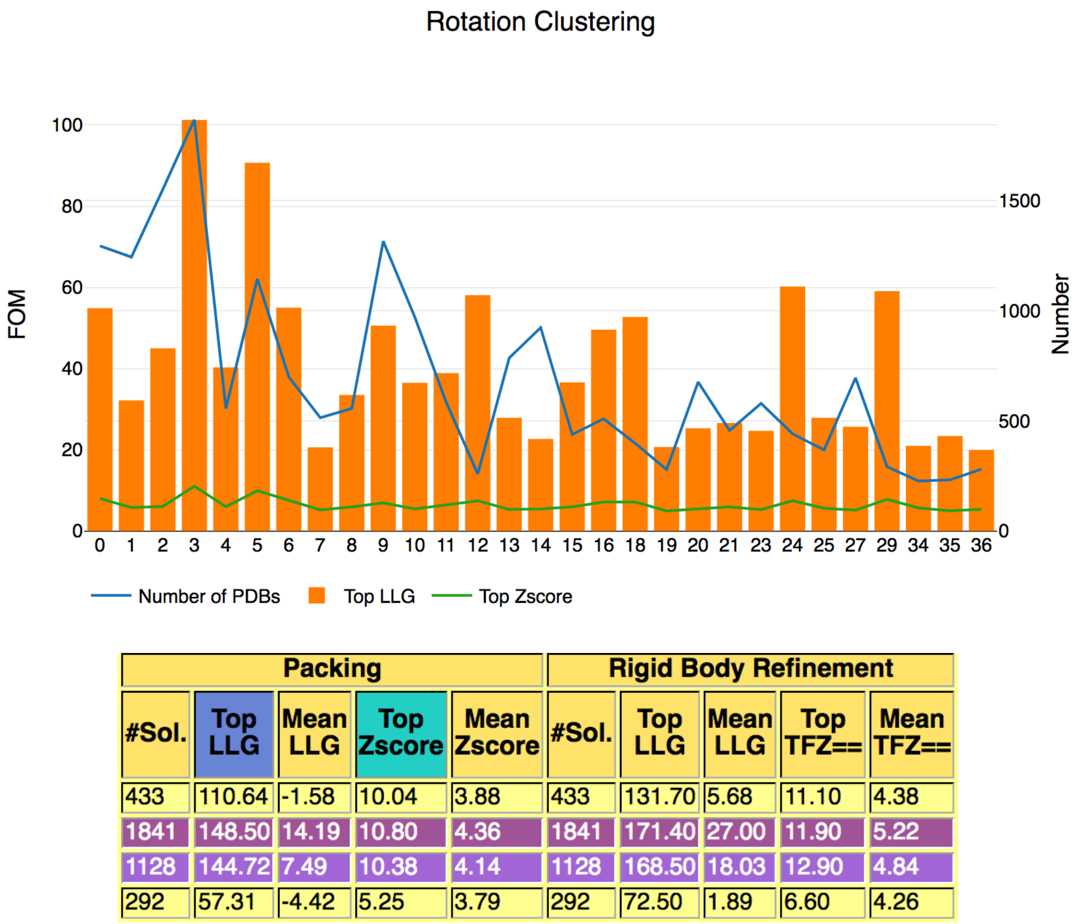
3.7 Partial solutions from ARCIMBOLDO_BORGES libraries

The combination of solutions from large libraries produced with ALEPH (Sammito *et al., in preparation*) and evaluated in ARCIMBOLDO_BORGES (Sammito *et al.*, 2013) has also been explored.

3.7.1 A structure made of immunoglobulin-like folds

The structure solution of a structure containing immunoglobulin-like folds, with 363 residues in the ASU was achieved with ARCIMBOLDO_BORGES using a library of β-sandwich folds, and combining the solutions with ALIXE. Data in *P*2₁ space group and resolution up to 1.6Å was available.

The library contains 3069 models, extracted from a subset of the PDB database containing only immunoglobulin-like folds. Models contain six β-strands, forming two antiparallel sheets that face against each other. Figure 15 shows the FOMs at different stages.



Packing					Rigid Body Refinement				
#Sol.	Top LLG	Mean LLG	Top Zscore	Mean Zscore	#Sol.	Top LLG	Mean LLG	Top TFZ==	Mean TFZ==
433	110.64	-1.58	10.04	3.88	433	131.70	5.68	11.10	4.38
1841	148.50	14.19	10.80	4.36	1841	171.40	27.00	11.90	5.22
1128	144.72	7.49	10.38	4.14	1128	168.50	18.03	12.90	4.84
292	57.31	-4.42	5.25	3.79	292	72.50	1.89	6.60	4.26

Figure 15 Output from an ARCIMBOLDO_BORGES run

Upper graph shows all rotation clusters, along with the number of pdb models represented in each of them and the top LLG and ZSCORE. In the table below, only the four rotation clusters that were selected for further processing are shown.

Out of the 7388 solutions that pack and pass to *Phaser*’s rigid body refinement and SHELXE initial CC computation, 1819 have wMPE with respect to the final solution of less than 80º, and 40 of less than 70º. These correct solutions are found in all four rotation clusters. Rotation clusters 5 and 3 were characterized by the highest LLGs after rigid body refinement, with a top LLG of 171.40 and 168.50. In rotation clusters 24 and 29, there are also correct solutions characterized by high FOMs, but their number is smaller.

A first round of clusterization in each of the rotation clusters was performed. Not all solutions are tried at that stage, so only the 100 top solutions are considered. In rotation clusters 5 and 3, characterized by the larger figures of merit, all top 100 solutions merge together in single clusters, presenting wMPEs to final structure of 65.5º and 64.7º respectively. Rotation clusters 24 and 29 also produce phase clusters but of smaller dimensions, between 2 to 6 phase sets, and characterised by slightly worse wMPE (68 to 73º). When these phase sets are use in a second round of clustering, this time with a much larger tolerance for the MPD (87º), they merge together in a single cluster. The values of the MPD at this stage show that rotation clusters 3 and 5 match the same monomer, as well as two phase clusters coming from rotation cluster 24, although in this case their overlap is partial. In the case of the phase clusters coming from rotation cluster 29 and a couple more from rotation cluster 24, they match different monomers. The phase set produced by the combination of these seven phase clusters of the first round develops to a full solution with a trace characterized by 36.70% CC.

3.8 Lessons learned from the combination of phases from fragments

The results from both the ideal and the real fragments have allowed to conclude a few general aspects about the combination of phase information between partial solutions. First of all, that a few cycles of density modification enhance origin discrimination for all cases. Secondly, that in principle, we can distinguish the MPDs that relate overlapping fragments (always below 60º) and non-overlapping (very high, marginal, as high as 87º). It also became obvious that overlapping fragments provide a really good scenario for phase combination as, if correct, they will enhance coherence between solutions. In the particular case of polar space groups, finding the appropriate shift when the differences are very high has proven to be quite challenging. It is possibly only worth if alternative strategies have previously failed, as the risk of losing correct partial solutions due to their combination under a wrong shift must be taken into account.

3.9 Alixe implementation

ALIXE is available both as a command line standalone program and integrated within the ARCIMBOLDO software.

As a standalone program, it can be used by providing a list of phase sets to evaluate, as well as cell and space group data and parameterisation.

ARCIMBOLDO_BORGES evaluates both the BORGES libraries and the spherical ARCIMBOLDO_SHREDDER compact models. After all *Phaser* steps have been performed, the result is a large number of solutions that corresponds to locations of different models in different rotation clusters. The phase sets are generated during the step that computes the initial CC (and only 5 cycles of density modification are applied) and are kept in their corresponding folders. If the ASU is expected to contain only one monomer, a single step of phase combination will be performed, and it will be tested in every rotation cluster. However, if more copies are expected, two phase clustering steps will be performed. In the first step phase combination is performed within rotation clusters and then their resulting combined phase sets are used for yet a second round of combination using a higher tolerance (87º) on all the available clusters from the first round.

In the coiled-coil mode of ARCIMBOLDO_LITE (Caballero *et al.*, 2018), where multiple correlated correct and incorrect solutions are often found, ALIXE is used to reduce the number of redundant hypotheses and to provide a better starting map for autotracing. In the case of the standard ARCIMBOLDO_LITE, where a sequential search of fragments builds up a solution, ALIXE is not used by default.

For all these possible scenarios, the common denominator is that at some point, we have a list of solutions, characterised by their FOMs. This list will be used to perform a referential clustering, in which, successively, the top solution will be used as reference, then merged with whichever phase sets are below a threshold of MPD, and then, removed from the list. Below, the core algorithm that performs the referential clustering is described.

Pre-processing step:

Input files required are the phase sets (in the SHELX format, .phs or .phi, that contains [h,k,l, F, Phase, FOM, sigF], as well as either a SHELX .ins file or a .pdb file with a CRYST card in order to retrieve symmetry information.

In order to perform the referential clustering on a list of phase sets from one structure, the program requires the following parameters to be set:

- The number of cycles of phase combination to perform (by default, 3 cycles)
- The tolerance in degrees (º) for the MPD between the phase sets that will be combined
- The FOM to weigh the contribution of the phases (F or E-values)
- The resolution limit to determine the origin shift (by default is 2.0 Å)
- The phs or list of phs files that will be used as references

The first steps involve the preparation of the general data structures and variables, and proceed as follows:

1. Required symmetry information is retrieved from a space group dictionary
2. Unit cell dimensions and angles are used to precompute the volume of the unit cell as well as other metric coefficients that will be used for computations between reciprocal and real space.
3. The reference phs file is read and used to compute the normalised structure factors and the epsilon factors. A resolution cut is performed if set and reflections are sorted. Also, the structure factors (F) and their errors (sigF) from the reference are saved independently, as all solutions share the same set.
4. The rest of phs files are read and saved into arrays. The same steps that were performed to the reference array are applied to all the arrays.

The subsequent steps are described in Algorithm 1 using pseudocode.

Algorithm 1: Referential clustering of phase sets

```
Input:
      list of phase sets
      symmetry operations and precomputed coefficients
Output:
      Hashtable with information about the clustering.
      merging was successful: Phi file with the merged phase set
      merging was NOT successful: Phi file with the resolution cut

n_cycles = 3
list_phasesets = list with paths of input phs files
list_references ⊆ list_phasesets
for reference in list_references:
    for cycle in n_cycles:
        for phaseset in list_phasesets:
            list_wmpd, list_mapcc = shift_cmp(phaseset,reference)
            MPD = sorted des list_wmpd by MPD
        asc sort list_phasesets by MPD (phaseset ∊ list_phasesets)
        reference = combine(sorted_list_phasesets,threshold_MPD)

if reference has been combined
      return merged_reference_as_phi_file, hash_table_results
else:
      return original_reference_as_phi_file with the resolution cut
```

Additional functions:

shift_cmp

```
Note:
      If space group is non-polar, origin shifts are read
      If it is polar an approximation is computed
Input:
   phaseset1(reference), phaseset2, symmetry_hash
Output:
   list_wmpd, list_mapcc

list_possible_origins = retrieve_origins from stored symmetry_hash
list_wmpd = empty list
list_mapcc = empty list

for origin in list_possible_origins:
    for reflection in list_reflections:
        apply origin shift to phaseset2
        Δphase between phaseset2 and phaseset1
        sum(phase_difference ∗ weight)
        sum(weight)
        sum(amplitudes$^2$ ∗ cos(Δphase))
        sum(amplitudes$^2$)
    wMPD = sum(Δphase ∗ weight) / sum(weight)
    mapcc= sum(amplitudes$^2$∗cos(Δphase)/sum(amplitudes$^2$)
    list_wmpd.append((wmpd,origin))
    list_mapcc.append((mapcc,origin))
    return list_wmpd, list_mapcc
```

Combine

```
Input:
     sorted_list_phasesets, threshold_MPD
Output:
     merged_phaseset

merged_phaseset = empty array
for phaseset in sorted_list_phasesets:
  t = sqrt(1/position of phaseset in sorted_list_phasesets)
  for refl in list_reflections:
      weight_combi[refl]= min(1,sqrt((FOM[refl])2+(phase[refl])2)*t)
      phase[refl] = weight_combi[refl]*evalue[refln]
      merged_phaseset[refl] = [phase[refl],FOM[refl]]
return merged_phaseset
```

Then MPDs are computed with respect to the reference. The computation of the MPD involves asserting or inferring which is the origin shift between the two phase sets being compared:

3.9.1 MPD and shift computation for P1 space group

Though the FORTRAN prototype for phase combination, PHSTAT, included an algorithm to compute an estimated origin shift for triclinic cases, it proved unfeasible for its use with partial maps, even with overlapping fragments. This result, taken together with the difficulties found with polar space groups and independent fragments, has made us decide to leave out of this work the implementation of ALIXE for triclinic cases.

3.9.2 MPD and shift computation for polar space groups

the allowed discrete origin shifts are tested and an initial origin shift is estimated in the polar direction using the layer of index 1 which is later refined against all reflections (Lunin & Lunina, 1996). This approximation is calculated to take advantage from the fact that, for a phase to be equivalent, whichever shift has been applied must be correct according to the symmetry of the space group. An equivalent phase, φ_m, must hold the following:

$$\boldsymbol{\varphi_m = \varphi - 2\pi t_m = \varphi - 2\pi(ht_1 + kt_2 + lt_3)} \qquad \textbf{(8)}$$

For the non-polar directions, the origins will be restricted, and if all possibilities are tested, the only unknown is the index that affects the polar direction.

$$\boldsymbol{\varphi_s = \varphi - 2\pi h t_1 \; 2\pi} \qquad \textbf{(9)}$$

Therefore, if we only use the indexes of the polar direction with a value of 1 or -1, we can get an initial approximation to the translation t_1.

$$\boldsymbol{\frac{(\varphi_s - \varphi)}{2\pi} = t_1} \qquad \textbf{(10)}$$

The polar direction will either be B ($P2$, $P2_1$, $C2$, $I2$) for monoclinic space groups, or C, for tetragonal ($P4$, $P4_1$, $P4_2$, $P4_3$, $I4$, $I4_1$), trigonal ($P3$, $P3_1$, $P3_2$) and hexagonal ($P6$, $P6_1$, $P6_2$, $P6_3$, $P6_4$, $P6_5$, $R3{:}H$). Finally, three equivalent polar directions are found for the non-standard setting $R3$ space group.

3.9.3 MPD and shift computation for non-polar space groups

In this case, all possible origin shifts for the space group will be tested in the function `shift_comp`.

3.9.4 Symmetry information retrieval

With the objective of facilitating symmetry handling within ALIXE, a Python dictionary is defined, in which the keys are the space group numbers, and the value is another dictionary, which contains all symbol representations, the symmetry operations, the point and laue group, a list of possible origins, a boolean indicating whether the space group is polar and the symmetry cards required for SHELXE in order to perform density modification and autotracing starting from a phase set.

4 PHASING THROUGH GENERATION OF COMPACT FRAGMENTS FROM A DISTANT HOMOLOG: ARCIMBOLDO_SHREDDER

4.1 Introduction

MR has become the default choice for structure solution when a homologous model is available (Berman *et al.*, 2013). The increasing availability of experimental structures and the introduction of ML based targets (Read, 2001) has pushed the boundaries of how much a template might deviate structurally and still be able to succeed in MR.

Numerous studies have proven that sequence identity does correlate with structural similarity, but in order to infer homology accurately, there is a minimum amount of sequence identity that is required. Already back in the 60s to 80s, when as few as 213 structures were available, the first analyses proving this relationship were performed (Chothia & Lesk, 1986, Wang, 1985, Kabsch & Sander, 1983b, Doesburg & Beurskens, 1983, Doolittle, 1981, Zuckerkandl & Pauling, 1965). Nowadays, with a PDB that has grown as much as 600 times what it was in those days (it holds more than 140000 entries), more recent studies (Rost, 1999, Krissinel, 2007, Krissinel & Henrick, 2004) have shown that many of the conclusions from this early work still hold, while some of them have been revisited. It is well established that over 35% sequence identity, there is a strong correlation between sequence identity and structural similarity. But for values between 20%-35%, the so-called twilight zone, this correlation starts to be lost, and both true and false positives are found. For sequence identities below 20% (the midnight zone), and although there are some examples of homologous protein pairs, the likelihood of structural homologs is negligibly small. The length of the alignment (Sander & Schneider, 1991) plays also a role, that is taken into account via the e-value of the alignment. Lastly, proteins might undergo conformational changes when bound to ligands or in different crystal forms, thus hindering solution by MR even in the presence of significant sequence identity.

MR relies on these structure/sequence relationships because it requires a model from which to compute structure factors (including their phases) and compare them to that of the unknown data. But in MLMR, an estimation of the errors in the model is also required, because it is a parameter in the likelihood functions. The estimation of the model error is given by an expected r.m.s.d between model and target structure. A better estimate of the r.m.s.d. for use in MLMR has been described in recent work (Oeffner *et al.*, 2013). The new equation includes a size factor, which accounts for the fact that homologous large structures have long-range structural perturbations. It also includes the possibility to refine this parameter (VRMS).

When dealing with unsuccessful models in MR, there are various paths that can be tried. Model improvement can be guided by the use of information about sequence conservation. For example, using paired or multiple sequence alignments between the homologs and the target sequence (Schwarzenbacher *et al.*, 2004) to trim non conserved regions or sidechains, or removing the coil regions from the model, as they tend to be less conserved than the secondary structure elements (Mizuguchi & Blundell, 2000, Sitbon & Pietrokovski, 2007). Consistency analysis of the conservation of residues between models is automatically performed in tools such as Sculptor (Bunkoczi & Read, 2011), or pipelines such as MrBUMP (Keegan *et al.*, 2018). MolRep (Vagin & Teplyakov, 1997) has a number of model-preparation and Patterson function techniques that allow specific adjustment of the model modification parameters, and that can make use of sequence, surface accessibility and experimental data (Lebedev *et al.*, 2008). The combination of multiple models, that is, the generation of ensembles (Leahy *et al.*, 1992, Pieper *et al.*, 1998) is also a successful approach, that allows to capture the degrees of confidence in various regions of the model, by reinforcing the common areas and down-weighting the more variable ones. Preparation of ensembles is implemented in Ensembler (Bunkoczi *et al.*, 2013), and both Molrep and *Phaser* can use them in MR. Modifying the models can also be achieved by normal mode analysis (NMA), which allows to calculate vibrational modes and anticipate protein flexibility (McCoy *et al.*, 2013, Suhre & Sanejouand, 2004) or by modelling within protocols devised for this purpose in Rosetta (DiMaio *et al.*, 2011), Quark (Xu & Zhang, 2012) or iTasser (Wang *et al.*, 2017, Zhang, 2008). Fragmenting and reassembling search models has also been explored (Shrestha & Zhang, 2015).

The assumption behind ARCIMBOLDO_SHREDDER is that even if globally a model might deviate too much to succeed as a whole in standard MR, it is possible to find smaller folds within it that are locally very similar, and that these local folds can be improved relying on the experimental data.

In the first implementation of ARCIMBOLDO_SHREDDER (Sammito *et al.*, 2014), the aim was to improve a starting model by trimming regions that did not agree with the experimental data. This template trimming relied on a rotation function based scoring, the SHRED-LLG. The whole template was initially used to find the maxima of the rotation function. The list of peaks in the rotation function was clustered within a given tolerance. For each of these clusters, the template was systematically shredded (omitting from the polypeptide chain continuous stretches in a range of sizes) and fragments were scored against each unique solution of the rotation function. Then, results were combined into a score per residue and the template was trimmed accordingly. The sequential shredding and its derived model trimming can improve models where the high average deviation to the target is due to dissimilar or flexible regions reducing the signal from a core of low r.m.s.d from the target structure.

In 2018, a new implementation of the shredder algorithm was published (Millán *et al.*, 2018). It has been extended to use fragments defining an approximately spherical volume in order to extract compact structural units from a distant homolog. To increase the radius of convergence of this approach, additional degrees of freedom are given to the models, which are decomposed in rigid body groups and subject to refinement against the intensity based likelihood rotation function target (RF) (Read

& McCoy, 2016) and again after they have been placed in the unit cell. This refinement is accomplished in *Phaser* with the gyre and gimble modes (McCoy *et al.*, 2018).

4.2 Methods for model improvement

In ARCIMBOLDO_SHREDDER, a series of strategies for model refinement in *Phaser* can be used, and what follows is a short description for each of them:

Gyre (McCoy *et al.*, 2018):

Gyre is a maximum likelihood replacement for Patterson Correlation (PC) refinement (Brünger, 1990). Both aim to refine rigid bodies defined within a model for which the orientation has been found. This can effectively improve the model and increases the convergence to a correct rotation (DeLano & Brunger, 1995, Grosse-Kunstleve & Adams, 2001). In fact, we have been previously using successfully PC refinement for structure solution in ARCIMBOLDO_BORGES with libraries of pairs of helices (Sammito *et al.*, 2013). While the target function in PC refinement is the CC on the structure-factor intensities, in gyre it is the rotation maximum likelihood function (Storoni et al., 2004). The centre of mass of the model is used to define the angular perturbations and the translations in orthogonal directions in space. Refinement can be tethered to the initial orientation, and the estimation of model accuracy can be refined.

Gimble (McCoy *et al.*, 2018):

Gimble is a maximum likelihood rigid-body refinement strategy, that similarly to gyre, uses a model divided into rigid groups to perform refinement of their relative orientation and position, but this time against the translation-function/refinement maximum likelihood function (McCoy *et al.*, 2005).

LLG-guided pruning (Oeffner *et al.*, 2018):

This method allows to perform pruning of residues when the phase error is high after the model is already placed. However, it does not directly refine atomic occupancies, as that would increase the risk of overfitting. Instead, it decides what number of residues *n* needs to be removed to reach a target eLLG. Then, occupancies are refined in windows of *n* residues for each offset of the window along the protein chain. The results are combined and an occupancy-refined structure with per-residue occupancies in the range between 0.01 and 1 is obtained. The occupancy-refined structure is then converted to a pruned structure, by the application of an occupancy threshold above which the refined occupancies are set to 1 and below which they are set to 0. The optimal threshold is selected by testing thresholds and calculating the LLGi for the model pruned at each value, choosing the threshold generating the highest LLGi.

4.3 Proof of principle on using compact fragments for phasing

The solution of the structure of PPAD (described in section 5.2) was achieved by manual generation of compact models starting from different homologous templates. This result, together with the success in using ALIXE with overlapping fragments (as described in section 3.5), prompted the development of an automatic mode for ARCIMBOLDO_SHREDDER in which to exploit the use of libraries of fragments derived from a homolog. The first approach to the generation of the models was somewhat simple, computing spheres of a given radius centred in a residue, and defining a model including all residues enclosed within the radius of the sphere. Still, this implementation was successful in the solution of another previously unknown structure, LTG (described in section 5.3). After this proof of principle was established, a more sophisticated generation of the models was developed, which is described in the following sections, discussing the current implementation of the program and its application to different cases.

4.4 The selection of a starting template

Finding an appropriate template for MR always involves, in the case of a known sequence, the search of putative homologs against databases. For that purpose, BLAST and PSI-BLAST (Altschul *et al.*, 1997), SSEARCH (Pearson, 1991, Smith & Waterman, 1981), FASTA (Pearson & Lipman, 1988) and HMMER3 (Johnson *et al.*, 2010) are commonly used algorithms to perform the required alignments and searches. HHPRED (Zimmermann *et al.*, 2017, Soding *et al.*, 2005), is devised particularly for distant homolog detection. It is based on pairwise comparison of profile Hidden Markov Models (HMMs). These HMMs are generated from multiple alignments and also include information on predicted secondary structure. Once the HHMs are generated, they can be compared with a database of precalculated HMMs (Soding, 2005). In the case of the PDB database, the secondary structure information is provided by DSSP (Kabsch & Sander, 1983a) from the 3D structure.

Since ARCIMBOLDO_SHREDDER is aimed at structure solution when obvious homologs are not available or the models are expected to present high deviations, in all the cases described in this work (both test and previously unknown structures), HHPRED has been used to search for structurally similar models. HHPRED produces a list of putative hits, sorted by a score that includes the secondary structure matching between target and query. It also outputs the multiple alignment, that can be used to trim the models if required.

4.5 Spherical mode implementation

The program accepts a configuration file, with extension `.bor`, which contains the parameterization of the run. Most parameters have appropriate defaults, and the only mandatory input is the data description, a template model and the shredding mode. The generation and evaluation of sequentially shredded models is mostly unchanged from the algorithm described in 2014 (Sammito *et al.*, 2014). From here on, the spherical mode is described. Figure 16 summarizes the flow of ARCIMBOLDO_SHREDDER's spherical mode.

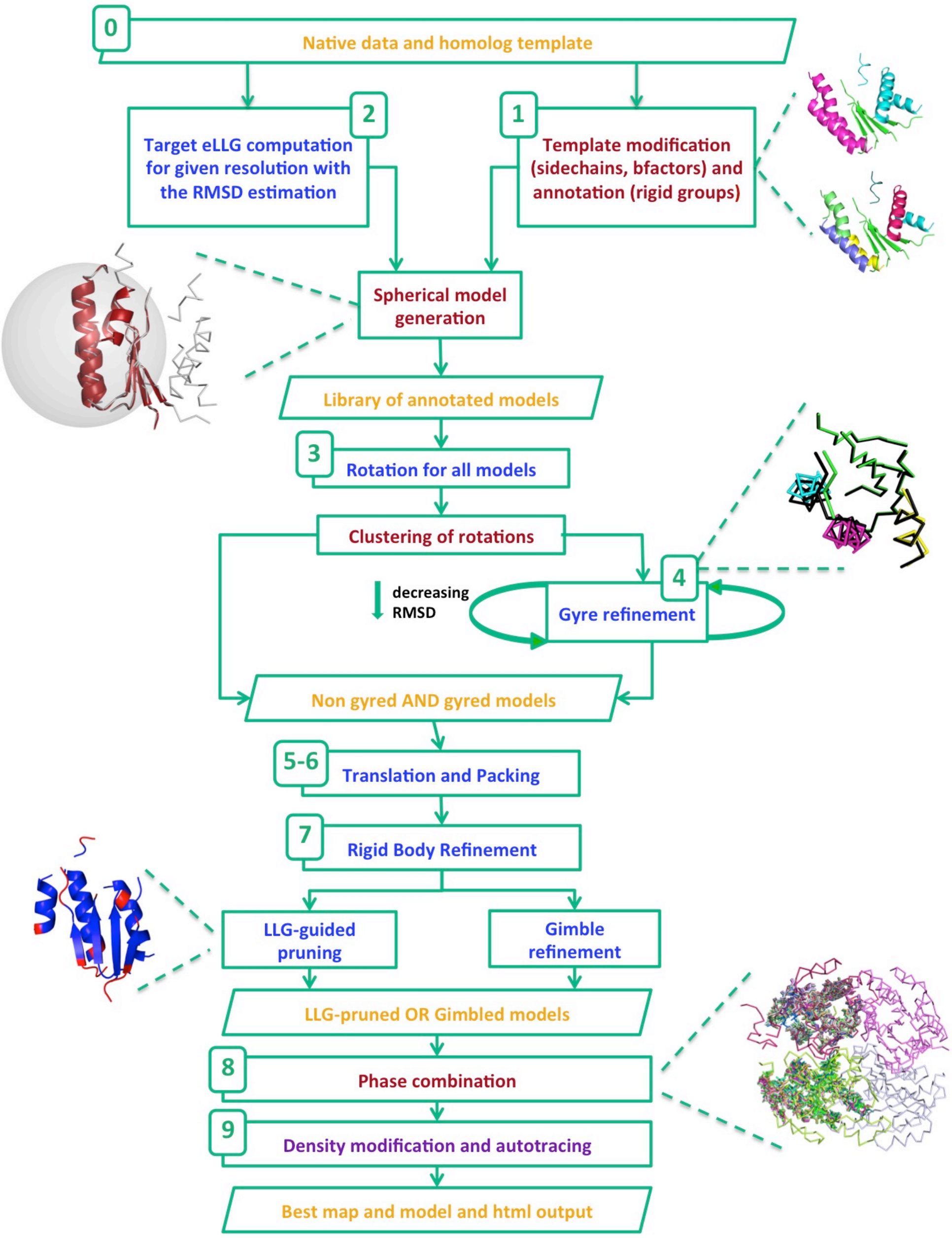

0
Native data and homolog template
2
Target eLLG computation for given resolution with the RMSD estimation
1
Template modification (sidechains, bfactors) and annotation (rigid groups)
Spherical model generation
Library of annotated models
3
Rotation for all models
Clustering of rotations
4
decreasing RMSD
Gyre refinement
Non gyred AND gyred models
5-6
Translation and Packing
7
Rigid Body Refinement
LLG-guided pruning
Gimble refinement
LLG-pruned OR Gimbled models
8
Phase combination
9
Density modification and autotracing
Best map and model and html output

Figure 16 ARCIMBOLDO_SHREDDER spheres workflow

The numbers reference the steps described in 4.5. Orange color refers to input/output, blue to *Phaser* steps, red to ARCIMBOLDO steps and purple to SHELXE steps.

4.5.1 Initial checks

The first task performed by the program is the validation of the instruction file, which must contain all mandatory parameters and may override defaults. Non-existent or misspelt instructions will be ignored and physically impossible values, such as a negative value for the molecular weight, or a model size larger than the given template will cause the program to terminate. Further checks to ensure the run is viable comprise validation of paths to files and folders, format correctness of the input files, retrieval of hardware information, and compatibility of *Phaser* and SHELXE versions. Resolution of the data is also checked and if below 2.5 Å, the run will be terminated.

4.5.2 Partitioning and annotation of the template

The template model is pre-processed, analysed and annotated in terms of fragments that will be treated as rigid groups in gyre and gimble refinement. Default pre-processing trims side chains and sets a common B-value for all atoms. The user can override either default to preserve this information in the template model. Secondary structure elements present in the model are identified relying on the distribution, distances and angles between characteristic vectors defined from the centroids of α carbons to the centroids of carbonyl oxygen atoms from all tripeptides. Relations among characteristic vectors allow characterizing tertiary structure as well (Sammito *et al.*, 2013). Unless otherwise selected, coil regions in the template are trimmed. The first level of annotation divides the secondary structure elements into a few groups defined by distance and preservation of folds such as the association of strands into a sheet. A second level further separates individual helices. This partition scheme is established on the template using an algorithm based in community clustering (Pons & Latapy, 2005, Clauset *et al.*, 2004, G. & T., 2006, Rosvall *et al.*, 2010). Community clustering is the general name for algorithms aiming to find community structure in networks. A network is said to have communities if groups of nodes are more densely connected within the groups than with the rest of the network. The community clustering algorithm integrated in ARCIMBOLDO_SHREDDER is part of ALEPH (Sammito et al, *in preparation*) and it has been modified to weigh favourably homogenous clusters, as well as to consider differently the interactions between β strands packing together. The tertiary structure constraints derived from the community clustering are adopted for each of the partial models derived, using chain identifiers are to mark rigid groups. By default, they are set and modified by the program in the course of the ARCIMBOLDO_SHREDDER run as the fragment is progressively decomposed into more rigid bodies. Alternatively, the user may input a template already annotated with chain identifiers and set the program to preserve them. The secondary structure in the template model will also be used to automatically configure the SHELXE parameters that target autotracing of α-helices or β-strands.

4.5.3 Generation of the models

After the template has been annotated for partition, a library of equal sized models is generated. The eLLG (McCoy *et al.*, 2017) provides an estimate of the model size required to identify correct solutions. *Phaser's* MR_ELLG mode is thus used to estimate the number of polyalanine residues needed in order to reach a target eLLG for the available data, assuming an RMSD value. An assumed RMSD value is a key parameter in the likelihood calculations, determining the relative weights assigned to low- and high-resolution data (here, RMSD is used to describe the parameter value, to distinguish it from the actual deviation from the final structure, which is denoted r.m.s.d.). Models are generated to fit the calculated size. eLLG defaults used within SHREDDER are somewhat on the lower limits compared to a general MR case since as long as non-random solutions are generated, a combination of partial solutions and the subsequent density modification and autotracing will discriminate the correct solutions. The computation of the eLLG is performed even if the user sets the model size, and the program will issue a warning if the chosen parameterization appears unfavourable. The sequential shredding mode previously described is still available (Sammito *et al.*, 2014). In this mode, fragments of different sizes are systematically omitted from the template to identify simultaneously all the most incorrect regions. Conversely, the spherical mode provides a way to cut models in a spatial way, retrieving compact fragments that are structurally close rather than contiguous in sequence. This is performed by traversing the sequence and using each residue in turn as the centre of a sphere containing the number of residues estimated from the eLLG. For each model, residues are selected by their distance to the central amino acid, subject to the constraints of preserving secondary structure continuity and avoiding unconnected stretches of less than 4 residues for strands or 7 for helices.

The subsequent steps are described in Algorithm 2 using pseudocode.

Algorithm 2: Shredder spheres model generation

```
Input:
      CA-CA triangular distance matrix
Output:
      Model files

list_residues = extracted from CA-CA distance matrix
list_dist = extracted from CA-CA distance matrix

for current_res in list_residues:
    model = empty_pdb
    sorted_list_residues = asc sort list_dist by dist to current_res

    if len(model) ≅ target_size:
        model.add (current_res)

    for candidate_res in sorted_list_residues:
       if candidate_res is not used:
        residues_in_between = range(current_res,candidate_res)
        total_res_add = candidate_res + residues_in_between

    if candidate_res[‘ss_type’] == current_res[‘ss_type’]:
            model.add(total_res_add)
            flag total_res_add as used

    if len(model) ≅ target_size:
        write(model)
    else:
        elongation(model, list_dist, target_size)
```

Additional functions:

```
elongation

Input:
      model, current_res, list_dist, target_size
Output:
      model
```

```
extremities = residues that break continuity
for extremity in list_residues:
    list_dist_to_ext.append(list_dist[extremity])
sort_to_add = sort asc list_dist_to_ext by dist to current_res
for res in sort_to_add:
    if len(model)==target_size:
        break
    else:
        model.add(res)

if len(model) ≅ target_size:
    write(model)
    return True
else:
    return False
```

After all the models have been generated using as central residue every residue from the template, a step of filtering is performed, in which the redundant models are eliminated.

All models are gathered in a library. In subsequent steps, the library is used and evaluated with an algorithm like the one previously described for ARCIMBOLDO_BORGES (Millán, Sammito & Usón, 2015) although parameterization and default options are specifically devised for ARCIMBOLDO_SHREDDER, as will be described in the next paragraphs. In contrast, the models derived from homologs in the sequential mode undergo an ARCIMBOLDO_LITE like subsequent treatment. ARCIMBOLDO_BORGES was originally designed to evaluate libraries of superimposed local folds of the same size (Sammito *et al.*, 2013), such as three β-strands forming an antiparallel β-sheet, extracted from the PDB (Berman *et al.*, 2000). Common size ensures that FOMs are comparable and given the superposition of the initial models, equivalent rotations bring models to the same position.

4.5.4 Evaluation against the likelihood rotation function target

An independent rotation search is performed on each of the models in the library. The resulting rotation angles are clustered within a given threshold (15º by default) taking symmetry into account and all models producing rotations in the same cluster are gathered. A model usually populates more than one cluster, either because the ASU contains more than one copy of the structure or because small fragments may fit different parts of a structure or because wrong solutions are obtained along with correct ones. In either case, it is convenient to isolate these different situations, so from that point on, every step is performed independently on each rotation cluster. Also, subsequent default filters are used independently unless a given cluster is aborted so that diversity is preserved while keeping the number of solutions within manageable limits.

4.5.5 Gyre refinement

Models can be subject to a step of gyre refinement (McCoy *et al.*, 2018) against the intensity likelihood rotation target (Read & McCoy, 2016) starting at their highest scoring rotation solution for the given cluster. Atoms with different chain identifiers within an ensemble will be treated as independent rigid groups refining their rotation and relative translation. In gyre refinement, an initial RMSD parameter is chosen as a trade-off between convergence radius and sensitivity to coordinate accuracy, iterating refinement and decreasing the RMSD parameter estimation sequentially. The goal is to improve and select among the many possible models those with an r.m.s.d. versus the real below 0.6Å, and thus susceptible of being expanded to the full solution in the density modification and autotracing step. The chain definition also changes between cycles in order to increase the number of fragments and thus the degrees of freedom for model refinement, as predefined in the template partitioning step (Figure 16, step 1).

4.5.6 Translation search

Both rotated and gyred models in each cluster are subjected to a translation search. The RMSD value of the last cycle of gyre refinement will be used for the translation search and all the subsequent steps until VRMS refinement, for both gyred and non-gyred models.

4.5.7 Packing test

Translated solutions are filtered with *Phaser*'s packing function. In ARCIMBOLDO_SHREDDER, as models tend to be larger and are expected to be less accurate, the default for the packing test allows 10% clashes instead of the very stringent default in the other ARCIMBOLDO modes, which accepts no clashes.

4.5.8 Refinement

Phaser's rigid body refinement is performed on all solutions accepted in the packing test. If refinement of the VRMS (Oeffner *et al.*, 2013) has been set, it will be performed at this stage. Optionally, the original template may be superimposed on each placed fragment and trimming and refinement of the model is revisited. Whether on the small, placed fragments or on the whole template, 2 different ways to optimize are available. A gimble (McCoy *et al.*, 2018) refinement step, subdividing the placed model into the same rigid groups differentiated in gyre, can be subsequently applied. Alternatively, *Phaser*'s likelihood-based pruning can be used to eliminate from the refined model those residues whose removal leads to an increase in the LLG (Oeffner *et al.*, 2018). The RMSD set at the pruning step will determine the trade-off between completeness and accuracy in the resulting model.

4.5.9 Phase combination

Solutions from both the original and refined models are passed on to SHELXE, computing the initial CC and 5 cycles of density modification. This leads to some discrimination between protein and solvent regions. This is possibly the cause why even for phase sets with wMPEs too large to be improved, determination of the relative origin shift is enhanced. The phase sets produced are sorted according to their FOMs (CC, LLG at refinement, TFZ-Score). At this point, consistent phase sets can be combined in order to complete partial solutions and increase their information content. This is performed within SHREDDER by an integrated version of ALIXE (Millán, Sammito, Garcia-Ferrer*, et al.*, 2015), using a two-step procedure. First, for each rotation cluster, partially overlapping solutions are identified within 60º MPD to the clustered phases. Subsequently, if the ASU is expected to contain more than one monomer, a second round combines phase sets gathered in the first step from different rotation clusters, allowing a higher tolerance (87º).

4.5.10 Density modification and autotracing

The single or combined phase sets are used to calculate starting maps for iterative density modification and autotracing with SHELXE. If phase combination is disabled or the combined phases do not yield a solution, the procedure is performed on selected individual solutions. The FOMs used for selection depend on the previous steps: CC after having performed a correlation CC-guided trimming (-o) in SHELXE, or LLG otherwise. In either case, solutions characterized by top CC, LLG and ZSCORE will be included in the selected set.

4.5.11 Best solution traceback and output of FOMs

Throughout the run, an HTML output that is generated at the beginning keeps being updated with the FOMs corresponding to each of the steps. While SHELXE's density modification and autotracing are being performed, the trace with the highest CC is highlighted at every cycle in the HTML. Values above 30% typically indicate a solved structure at a resolution better than 2.5Å (Usón & Sheldrick, 2018). When the program finishes, the HTML output file describes the best solution found and its FOMs, together with links to its map and coordinate files.

4.6 An illustration on the use of ARCIMBOLDO_SHREDDER with test structures

This section describes a detailed analysis with the final version of the program for the cases of PPAD and LTG, which were originally solved with a prototype (as described in sections 5.2 and 5.3) and prompted the development of the spherical ARCIMBOLDO_SHREDDER. In addition, the all-helical repeat protein (PDB ID 3FP2) and a mixed α/β protein (PDB ID 1YZF) have been selected to test and illustrate parameterization for ARCIMBOLDO_SHREDDER.

4.6.1 LTG

LTG is a highly helical structure (86%) with a low coil fraction. Despite sharing the overall fold, the search template presents an r.m.s.d. versus the true structure of 4.6Å but helical fragments should be particularly suited for rigid body refinement, even though the original solution, described in section 5.3 was obtained with phase combination of partial solutions, before gyre and gimble refinement were implemented. In addition, many solutions are produced and the effect of parameterization should be more patent than in borderline cases when solutions are spurious. In particular, eLLG-derived model size, VRMS refinement, and LLG-guided pruning as an alternative to gyre and gimble refinement were probed. In all tests summarized in Table 2, template annotation and therefore model disassembling were predefined as displayed in Figure 17.

Table 2 Summary of the tests performed with the LTG structure

Parameterization and results for all tests are summarized.

	No gyre (reference)	Default	LLG-guided pruning	VRMS refinement	Variation in starting RMSD parameter and model size (runs 5,6,7,8)			
RMSD (Å)	1.0	1.0, 1.2	1.0, 1.2	1.0, 1.2	2.0	2.0	3.0	3.0
Model size (no. res)	128	128	128	128	127	180	127	180
Cycles of *gyre* refinement	0	2	2	2	1	1	1	1
Unique models	417	417	417	417	408	436	408	436
eLLG	28.4	28.4	28.4	28.4	1.7	3.4	0.17	0.34
Correct solutions	205	295	450	296	135	136	5	23
Total solutions	1228	2162	3201	2132	1852	1756	852	1012
Correct ratio	0.17	0.14	0.14	0.14	0.07	0.07	0.006	0.02
Best wMPE (^{o})	66.3	61.8	61.7	63.1	66.6	67.8	76.9	72.6
Top CC (%)	33.88	34.76	31.84	32.19	30.79	31.39	10.92	32.24

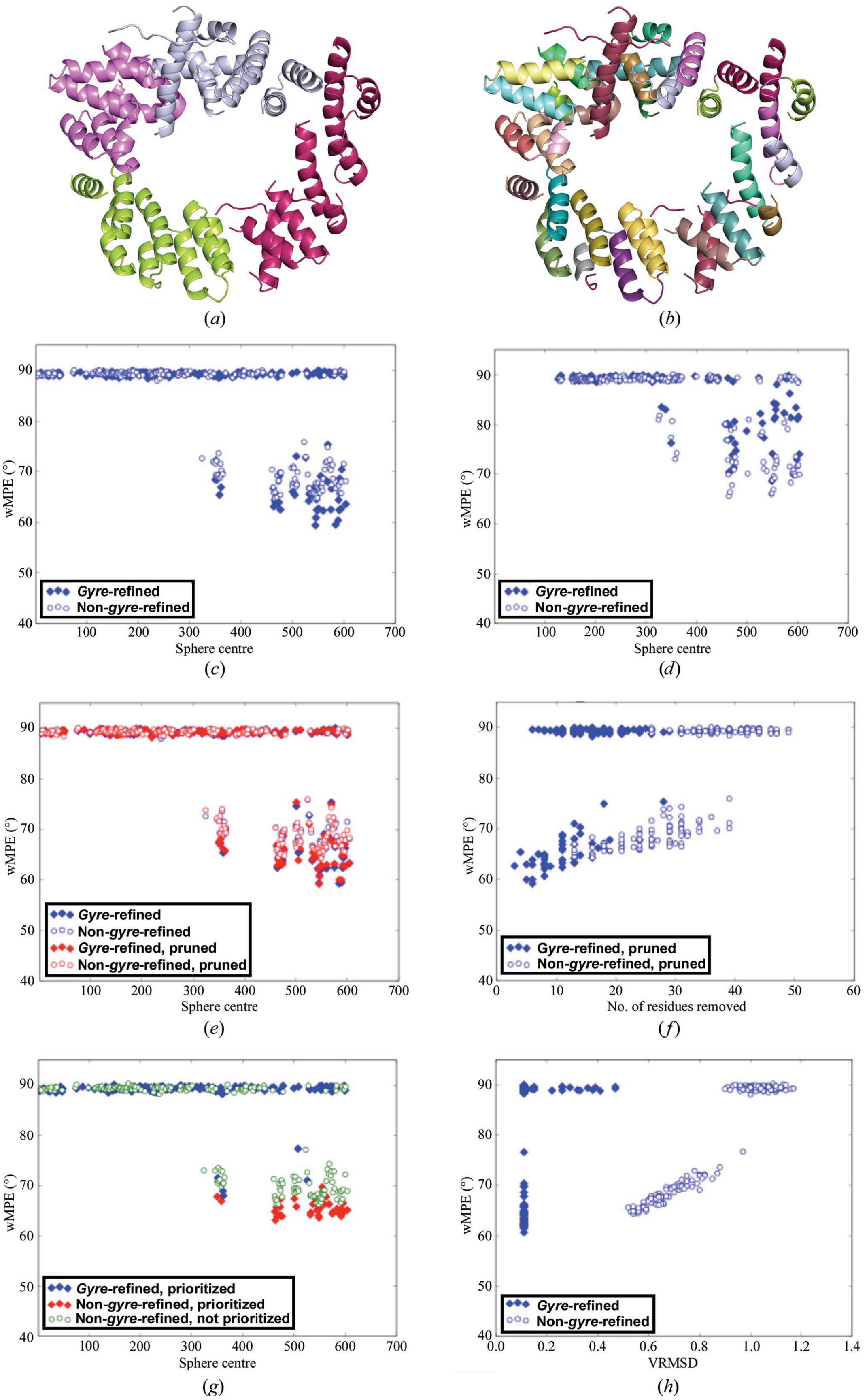


Figure 17 Tests on LTG structure

Each scatter plot corresponds to a correct rotation cluster. In (c), (d), (e) and (g) the horizontal axis represents the number of the central residue of the model. (a) First-level annotation groups. (b) Second-level groups of helices. (c) wMPE versus model centre for solutions in gyre and gimble refinement run 2. (d) wMPE for solutions in the run with one cycle of gyre refinement at 2.0Å RMSD (run 5). (e) wMPE for all solutions in the run with LLG-based pruning (run 3). (f) wMPE against the number of residues trimmed from each solution after LLG-based pruning in run 3. (g) wMPE versus model centre for solutions in the VRMS refinement run (run 4). A red colour marks solutions that have been prioritized for expansion. (h) VRMS against wMPE for all solutions.

If gyre/gimble were performed, a first cycle differentiated 4 groups in the template whereas a second cycle would treat each helix as an independent rigid group. Models of 128 or 180 residues were used, corresponding to eLLGs below 30, depending on the RMSD estimation.

4.6.1.1 Base run without gyre or gimble refinement

As a reference, the SHREDDER parameterization that best corresponds to the original solution was chosen. The main difference is that in this test all 417 models generated shared a common size, corresponding to an eLLG of 28 at 1.0Å RMSD. The selected size would be expected to yield solutions reaching around the inflection point of the LLG sigmoidal curve (McCoy *et al.*, 2017). Nevertheless, correct solutions of the rotation function become parts of two close clusters, populated by more than half of the models, which eventually produce clear discriminated solutions with an LLG well over 60, twice as high as in clusters that fail to lead to a solution. 17% of the substructures are non-random and the best one develops within SHELXE to a main chain trace of 466 residues and a map of 66º wMPE.

4.6.1.2 Gyre and gimble refinement

The same models were subjected to an initial rotation search and gyre refinement at an assumed RMSD of 1.2Å, distinguishing two rigid groups of the total four present in the template (Figure 17a), followed by 1.0 Å refinement of the rotations and relative translations of each helix in the model (Figure 17b). In this case, the initial rotation solutions are divided in the same two close clusters seen previously to contain correct solutions. After both refined and original models were placed, those passing the packing filter were gimble-refined, subject to the same decomposition as the last gyre step.

The solution leading to the best polypeptide trace, with a CC of 34.76%, had been processed by gyre and gimble. The final wMPE, 62º, decreases versus the original run.

The graph in Figure 17c, displaying all solutions from the main correct rotation cluster shows how in general, gyred models improve the wMPE versus non-gyred ones. For correct solutions in this run, r.m.s.d. between the placed fragments and the LTG structure range from 0.3 to 0.45.

4.6.1.3 Likelihood-based pruning on gyred and non-gyred solutions

As an alternative to the gimble refinement in the previous run, this run was set to trim incorrect residues using the likelihood-based pruning in *Phaser* (Oeffner et al., 2018). This refinement is performed for a window size producing a significant change in the eLLG. A threshold in the refined occupancy values for residue trimming is derived by probing different values and choosing the one for which the trimmed model shows the highest LLG. In the present case, model improvement through LLG-pruning prior to density modification and autotracing, solves the structure as well.

Graphs of the solutions for the main correct rotation cluster identifying them as gyred or non-gyred and pruned or not, reveal how the best phases correspond to solutions gyred and trimmed, and how the LLG-based pruning improves the wMPE (Figure 17e). Suitably, pruning removes fewer residues from the more correct gyred versus non-gyred solutions as seen in Figure 17f. Pruning of only a few residues can be a good indication of quality, especially for non-gyred solutions. Even if phasing is achieved in either case, starting the density modification step from models containing fewer errors may be beneficial. Some geometrical differences between search model and target, such as backbone torsions, cannot be improved by rigid body refinement.

4.6.1.4 VRMS refinement on gyred and non-gyred solutions

As models improve upon gyre and gimble refinement, r.m.s.d. to the target structure is expected to decrease. This is partially accounted for by decreasing the RMSD value in successive steps but VRMS refinement in *Phaser* should provide a better estimate of the final r.m.s.d. (Oeffner *et al.*, 2013), leading to a clearer identification of solutions to be selected for SHELXE expansion.

Figure 17g and h show graphs of the solutions in the major correct rotation cluster. Noticeable from the plot in Figure 17h is that the lowest VRMS correspond to the best wMPE. VRMS reaches values ranging from 0.11 (for gyred solutions) to 0.56 (for non-gyred solutions) in correct solutions. The values for the final r.m.s.d. obtained after gyre and gimble refinement for such correct solutions have indeed improved and range from 0.33 to 0.45Å.

In both the VRMS refined run and the non-refined (run 2) all selected solutions have been gyred. Some solutions represented by the red diamonds (gyred and prioritized) achieve lower starting wMPE in the case of the VRMS refined run (Figure 17g).

4.6.1.5 Runs with 1 cycle gyre refinement and large starting RMSD parameter (Runs 5, 6, 7 and 8)

Finally, four runs were computed with large initial RMSD to probe whether this could lead to an increase in the radius of convergence in model refinement. A single gyre step, with a few large groups, was undertaken. In run 5, the RMSD was set to 2.0 Å, even though for the same set of models this implies a substantial drop in the eLLG, which becomes 1.7. As in previous runs, close to correct rotations eventually leading to a solution are found in two different clusters but this time, a non-gyred solution is the best before expansion and phases are poorer (wMPE of 66.6°). As can be seen in Figure 17d, non-gyred models predominate. The gyre and non-gyred versions of the model are geometrically very similar as only a few large groups have been refined.

In run 6, with the same RMSD of 2.0 Å but larger models of 180 alanines, the eLLG increases to a still very modest 3.4. Nevertheless, the number and percentage of correctly placed fragments does not improve compared to the last run, neither do the phases of the placed fragments, corresponding to a wMPE of 67.8º for the best solution, which comes from an original model.

Runs 7 and 8 probe the same 127 and 180 alanine models setting the initial RMSD to 3.0 Å and confirm the trend. The smaller models in run 7, altogether fail to produce a correctly phased final structure. Neither refined nor original fragments are placed accurately enough for extension to succeed. Start phases for the few non-random solutions are worse by 10º (wMPE 76.9º) than in previous runs. Again, there is no improvement of refined versus non-refined models. The larger models in run 8 lead to an increase in the number of correctly placed fragments and the start phases they produce improve enough (72.6º) to provide one full solution. In this context, performing the initial refinement of few fragments at high RMSD does not appear to aid convergence, as non-gyred models are closer to true solutions. Accordingly, the program's RMSD default is chosen as 1.2Å.

In conclusion, for this highly helical model with diffraction data to 2Å resolution, gyre and gimble refinement of individual helices improves models, provided that the RMSD parameter is set to sufficiently low values around 1Å. Solutions can be identified by VRMS refinement, while LLG-guided pruning can be also used to trim incorrect fragments and enhances solution.

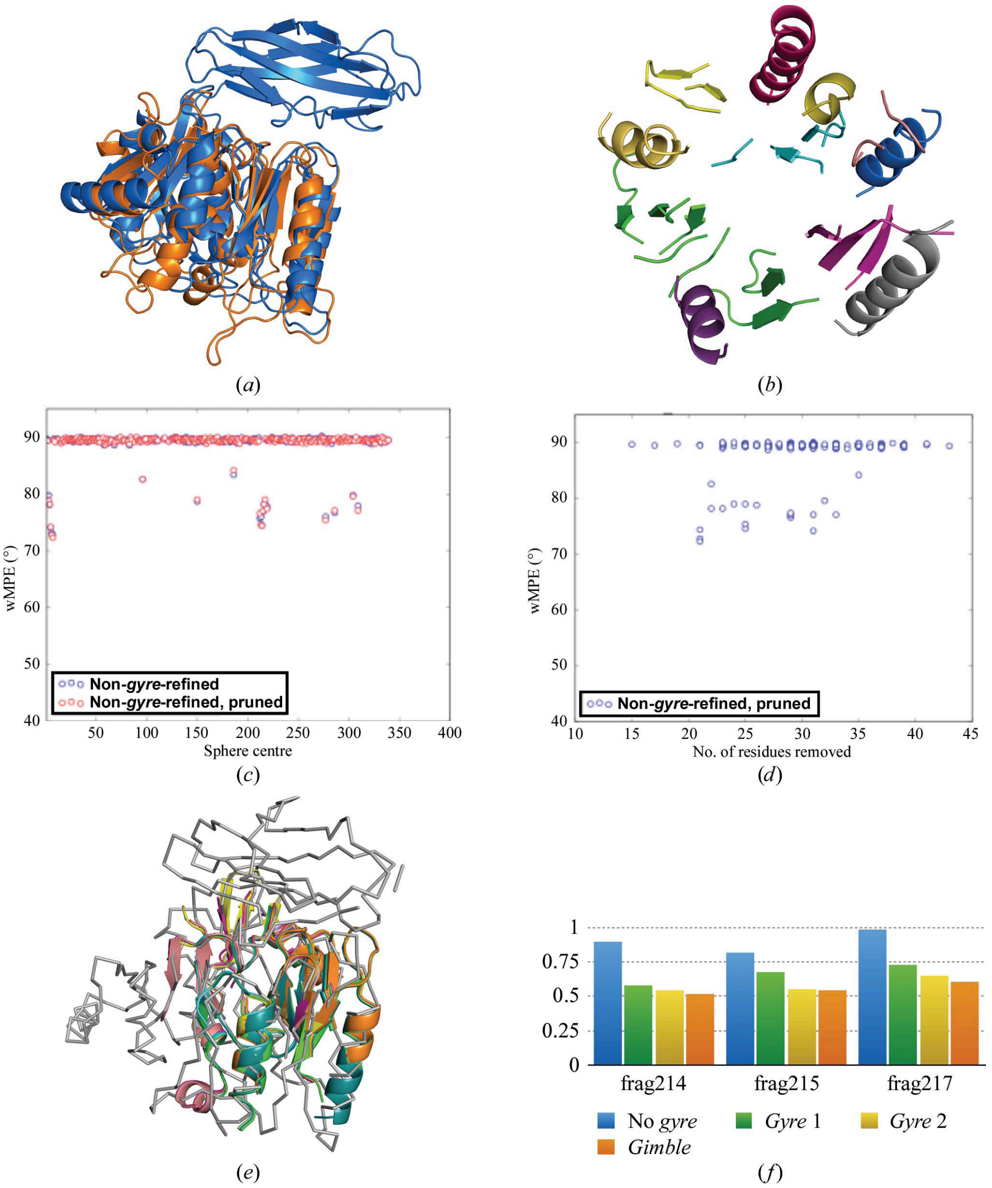
4.6.2 PPAD

The final structure of PPAD, superimposed on the template used to solve it, is displayed on Figure 18a. 1ZBR (Northeast Structural Genomics Consortium, unpublished work) shares 19% sequence identity with PPAD and the r.m.s.d. over a core of 231 Cα is 1.6 Å. The original solution of this structure (described in section 3.3) involved the combination of two partial solutions from overlapping models derived from 1ZBR. These models contained 108 and 127 residues respectively and had been obtained preserving coil regions in the starting template. Trimming the coil parts eliminates half of the model, and the resulting fragments fail to produce a solution. The PDB annotates this structure as containing 28% α-helices and 28% β-strands based on DSSP (Kabsch & Sander, 1983a). Our automated choice of secondary structure annotation for SHREDDER templates is slightly more conservative leading to a noticeably low secondary structure content in the case of this template with 25% α-helices and 33% β-strands, leaving 41% for coil and turns. Considering the large coil fraction in this structure, and the fact that previous successful solution had been accomplished with models preserving it, maintaining coil residues in model generation in ARCIMBOLDO_SHREDDER is a choice and may be appropriate in some cases. It must be considered too that the comparatively low fraction of residues in defined secondary structure elements leads to very fragmented models, dispersed over a large volume when coil residues are removed. Setting the RMSD to 0.8Å requires polyalanine models of 101 residues to reach an eLLG of 60. Three runs were compared under such conditions: two of them maintaining the coil regions in the template and one trimming it. In the first two, as models are continuous, local folds are not disassembled and thus not given additional degrees of freedom through gyre or gimble refinement. In the second run, model improvement was attempted within *Phaser* by LLG-guided pruning of residues in the placed model prior to input into SHELXE. In the third run, the "spherical" search models were generated from the coil-trimmed template and groups of secondary structure elements (Figure 18b) were refined using gyre and gimble methods. Results of all three runs are summarized in Table 3.

Table 3 Summary of the tests performed with PPAD

Parameterization and results of all three tests with PPAD are summarized.

	Maintain coil	Maintain coil, prune	Remove coil
RMSD (Å)	0.8	0.8	0.8
Models size (no res)	101	101	101
Unique models	335	335	160
eLLG	60	60	60
Correct solutions	32	48	6
Total solutions	1652	2478	1504
Correct ratio	0.019	0.019	0.0039
Best wMPE (º)	72.7	72.1	67.7
TopCC (%)	30.69	31.43	31.05



(*a*) (*b*)

(*c*) (*d*)

(*e*) (*f*)

Figure 18 Tests on PPAD

In runs 1 and 2 coil residues were kept, and run 2 included LLG-guided pruning. In run 3 coil was removed and the models were subjected to gyre and gimble refinement. (a) Superposition between the 1ZBR template (orange) and the final structure (blue). The r.m.s.d. is 1.57Å for a core of 231 Cα atoms. (b) First level of annotation for the decomposition used in run 3. (c) wMPE of solutions versus the model centre in run 2. (d) Number of residues removed by the LLG-guided pruning against wMPE in run 2. (e) The coloured cartoon shows solving fragments from run 2 that clustered together and the grey ribbon shows the final structure. (f) r.m.s.d. to the final structure for each of the three correct fragments in run 3. Values at different refinement stages are calculated over a common core.

The first run yields numerous partial solutions within one of the rotation clusters. This is clearly discriminated from all other clusters by its LLG of 64 versus less than 30. One of the placed models, whose phases correspond to a minimum wMPE of 72º, expands to a full solution identifiable by a mainchain trace encompassing 331 residues and characterized by a CC above 30%. The second run is identical to the first, but for the last pruning step modifying the models and their selection for density modification and autotracing. The starting phases are marginally better in some cases (Figure 18c and d) and lead to a comparable trace.

Among all placed models in runs 1 and 2 with non-random phases only one could be expanded into a full solution. It does not correspond to the top scoring solution, so the use of the phase combination with ALIXE was tested to increase the convergence of the method. The solution identified by the top TFZ (7.02) gives rise to a cluster of 14 phase sets gathering solutions with mean phase differences below 60º. Its expansion yielded a trace of 342 residues in 11 chains, characterized by a CC of 37%. All models contributing to this phase cluster are depicted in Figure 18e.

No decisive difference is seen by using pruning in terms of number of solutions or FOMs, but in borderline cases even a slight improvement may help. In general, many residues are being removed (Figure 18d), and in this case there is no clear correlation between correct/incorrect solutions and number of residues removed, even though solutions with the lowest wMPE are among those less trimmed.

A third run with less compact models from which coil residues were trimmed, subject to gyre and gimble refinement, gave rise to fewer but more accurate solutions than the previous runs. Three partial solutions with initial wMPEs of 67.7º, 68.7º and 70.8º, correspond to gyred and gimbled models. As seen in Figure 18f, the r.m.s.d. to the final structure improves in each gyre and gimble cycle. One of these solutions develops into a full solution characterized by a CC of 31.05%.

4.6.3 1YZF

The $P3_221$ crystal form of the lipase/acylhydrolase from *E. faecalis* at 1.9Å contains a monomer with 195 residues in the asymmetric unit and 36% solvent. It has a sequence identity of 21% to the homologous esterase EstA from *Pseudoalteromonas sp. 643A*, deposited as 3HP4 (Brzuszkiewicz *et al.*, 2009), and an r.m.s.d. of 2.4 Å over 121 atoms (Figure 19a).

This case exemplifies a borderline solution due to the large deviation to the search model while, despite the low solvent content, the structure can easily be solved with the same protocols described but using closer homologs such as 4RSH (1.15Å r.m.s.d. over 116 Cα atoms). Secondary structure annotation of the 185 residues in the 3hp4 template, assigned 88 to α-helices and 45 to β-strands. Polyalanine models of 83 residues were generated, corresponding to an eLLG of 60 for an expected RMSD of 0.8Å. Rotation search and the first cycle of gyre refinement (annotation shown in Figure 19b were performed with RMSD at 1.2 Å, while from the second gyre cycle on (annotation shown in Figure 19c), the rest of the steps were performed at a setting of 0.8 Å. Only one model produced non-random solutions. These belonged to rotation cluster 0, one of the four clusters selected by default but containing neither the top LLG scoring solution, nor the highest number of models. Among the six correct solutions, the one undergoing gyre refinement as well as LLG-pruning had the lowest wMPE and better FOMs. This solution occupies position 51 in the list of 60 substructures prioritized for expansion. Compared to the 74º wMPE the unrefined fragment yielded, both the gyre and gimble or the gyre and LLG-pruning combinations improve it to 67º. Given the low solvent content, expansion is difficult and a large number of cycles with the latest SHELXE version, featuring constrained autotracing (Usón & Sheldrick, 2018), are needed to lower the wMPE to 54º and produce an identifiable solution.

An attempt was made to design an improved protocol, which would make solution pathway for this test case more robust. We implemented the possibility to revisit refinement and/or trimming of the original model. The full, annotated template is superimposed on the solutions that have survived the packing test, whether gyred or non-gyred. These full models are then rigid-body refined and also subject to either gimble or LLG-guided pruning. In this case, starting from a correctly placed model with high deviations failed to improve on the initial wMPE, which remained above 72º in spite of the increase in scattering mass, as refinement or trimming did not eliminate the errors sufficiently. Nevertheless, this feature is described as it can be used in the program (using the keyword `mend_after_translation`) and may prove useful in other cases.

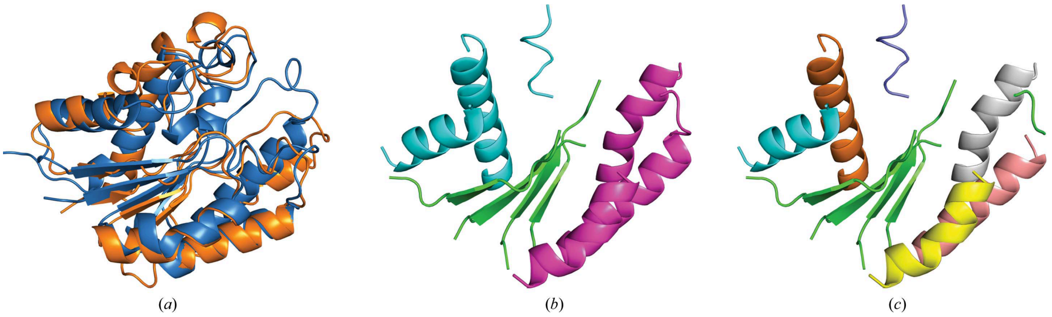


Figure 19 Tests on 1YZF

(a) Final structure (blue) versus the template used in ARCIMBOLDO_SHREDDER (orange). The r.m.s.d. between the structures computed with super in PyMOL is 2.4 Å over a core of 121 Cα atoms. (b) Community clustering groups. (c) Sheet and independent helices grouping.

4.6.4 3FP2

Tom71 is a tetratricopeptide repeat (TPR) containing protein made up of 537 residues comprising 27 helices with 6 to 22 residues each. TPR domains usually consist of tandem arrays of two antiparallel α-helices that generate a right-handed helical structure. Diffraction data from the PDB entry 3FP2 (Li *et al.*, 2009) extend to a resolution of 1.98 Å. The homolog tested was the superhelical TPR domain of the O-linked GlcNac transferase 1W3B (Jinek *et al.*, 2004), sharing 19% sequence identity with the target structure. Accordingly, the expected RMS (eVRMS) is 1.61Å but given the plasticity of the fold, both structures can only be partially superimposed. The search model contains 45 helices of 7 to 14 residues arranged in a fold that locally resembles through the TPR domains the target structure while presenting large overall differences. The superposition displayed in Figure 20 a matches 208 residues with an r.m.s.d. of 5.0Å.

Figure 20b and c show the template annotation for the first cycle of gyre refinement and subsequent refinement steps, respectively. Models with different sizes, comprising three to seven helices each, were tested as well as a range of starting RMSD values, from 0.8 up to 2.0Å. The only run that was successful in producing correct solutions was the one using the smallest models and the lowest estimated RMSD. In this run, the starting rotation search and first cycle of gyre refinement were performed at 0.8Å RMSD with models of 36 residues corresponding to an eLLG target of 25. Two more cycles of gyre refinement were run decreasing the RMSD down to 0.4 Å, which was the value adopted for all remaining steps. Three non-random solutions are found among the prioritized ones, all of them matching models that correspond to arrangements of three helices. The two solutions in the main rotation cluster zero (initial wMPE of 73.4º and 74.5º). Both of them develop to a full solution after density modification and auto tracing with SHELXE and can be identified by the main chain traces with CC of 44 and 46% respectively. A third solution is found in a different rotation cluster (wMPE 76.6º). It was not sent to expansion as ARCIMBOLDO_BORGES stops evaluating clusters once the structure is solved. The successful models are remarkably small, barely 5% of the mainchain atoms but their starting r.m.s.d. to the target structure is already close to 0.5Å, as seen in Figure 20d.

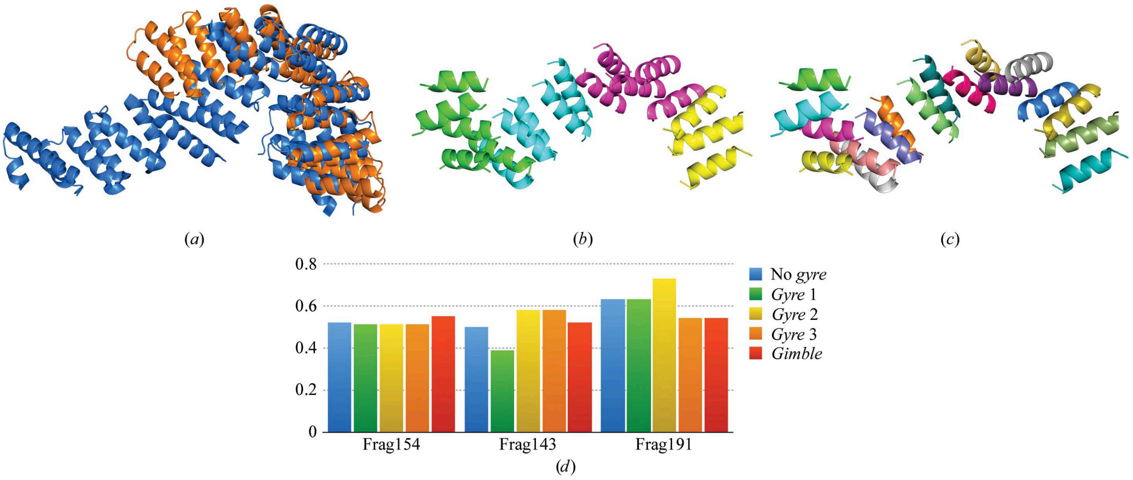


Figure 20 Tests on 3FP2

a) Final structure (blue) vs 1W3B template used in ARCIMBOLDO_SHREDDER (orange). RMSD between both structures is 4.95 Å over a core of 208 Cα atoms b) First level of annotation for refinement c) Second level of annotation for refinement d) RMSD of each of the three correct fragments to the final structure and over a common core at different refinement stages.

4.7 Lessons learned from testing ARCIMBOLDO_SHREDDER

ARCIMBOLDO_SHREDDER in spherical mode has been used to solve new and test structures. The program features have been illustrated, as well as the possibilities of its parameterization. The convergence of the method can be improved by using *gyre* and *gimble* refinement, and VRMS refinement can increase the chances of recognizing correct solutions. LLG-guided pruning can be used in cases where is preferable to keep the coil in model generation. A protocol to revisit model refinement after translation and use the original template superimposed to partial solutions is also available. Phase combination of consistent solutions can be exploited to merge solutions corresponding to either overlapping fragments or different monomers of the structure.

Current defaults are based on the tests described, but as the tests have shown, often variations are required since many features are available and the starting template will have different characteristics.

5 PRACTICAL IMPACT OF THE METHODS DEVELOPED IN THE SOLUTION OF STRUCTURES OF BIOLOGICAL INTEREST

5.1 Introduction

Methods development requires the availability of test cases in which to apply successfully or not the methodologies, learning in the process what might work best in the general case. However, very often, new features are developed within already existing algorithms or new algorithms are implemented altogether because there is an exciting, previously unknown structure that might require them. This has undoubtedly been my experience during the Ph.D., and thanks to the collaborations we have established with different groups, I have developed new features in my projects and solved challenging cases. These structure solution processes have been also useful to illustrate the use of the programs, therefore contributing too to the final aim of my work, which is to produce software that other structural biologists can use to solve their structures. In the following chapter, the solution of a few structures of biological interest will be described.

5.2 PPAD

PPAD is a peptidylarginine deiminase from *Porphyromonas gingivalis* (Goulas *et al.*, 2015). Peptidylarginine deiminases (PADs) catalyse the citrullination reaction, a post-translational modification of higher organisms that selectively deiminates arginines in proteins and peptides. It occurs in physiological processes but also pathologies such as multiple sclerosis, Alzheimer's disease or rheumatoid arthritis (RA). In fact, increased levels of citrullinated proteins are found in several if not all inflammatory diseases (Gudmann *et al.*, 2015). Until recently, peptidylarginine deiminases (PADs), had only been found in vertebrates but not in lower organisms. Uniquely among microbes, *P. gingivalis* secretes a PAD, termed PPAD (*Porphyromonas* peptidylarginine deiminase), which is genetically unrelated to eukaryotic PADs. Periodontal disease (PD) is among the most prevalent infectious diseases of mankind (Seymour *et al.*, 2007). Its major causal agent is *Porphyromonas gingivalis*. PPAD protects *P. gingivalis* during acidic cleansing in the mouth through ammonia generated during host and endogenous protein citrullination. The studies suggest that the link between RA and P. gingivalis-induced PD may result from PPAD-mediated citrullination (Maresz *et al.*, 2013).

Our collaborators from the group of Xavier Gomis-Rüth collected over 20 datasets from different crystals at macromolecular beamlines ID23-1 (Nurizzo *et al.*, 2006) in the ESRF and XALOC (Juanhuix *et al.*, 2014) in ALBA synchrotrons. Crystals belong to space group $P2_12_12_1$ and contain one copy of the 432 amino acids monomer in the asymmetric unit, corresponding to a solvent content of 40%. Table 4 shows the statistics for the merged dataset used for phasing.

Table 4 PPAD data statistics

	PPAD
Space group	$P2_12_12_1$
Unit cell parameters	
a (Å)	58.631
b (Å)	60.357
c (Å)	113.884
α (º) = β (º) = γ (º)	90
Resolution (Å)	53.33 – 1.50
I/σ(I)	31.62 (5.51)
Completeness (%)	99.1 (95.6)

A protein-protein BLAST (Altschul *et al.*, 1997) search with the PPAD sequence showed that the closest sequence available correspond to chain A in the PDB entry 3HVM (Jones *et al.*, 2010), with an E-value of 2e-05 and 23% sequence identity. However, a search in HHPRED (Soding *et al.*, 2005) broadened the choice of templates to a set of 6 structures characterized by 100% probability. Apart from 3HVM, the following structures were found to have a homologous relationship: 1ZBR, 1XKN, 2JER, 3H7C, and 2EVO. The common fold to all these structures is a pentein, β/α propeller, composed of five α-β-β-α-β units arranged around a pseudo fivefold axis.

MR did not succeed with any of these six models as search fragments. Instead, a variety of fragments were generated from all these templates. The structures were decomposed into the five pseudo-repeats, and models cut in a number of ways: with and without the helices, with and without side-chains, further trimmed from loops and partially overlapping. The resulting models and libraries were used as search fragments in ARCIMBOLDO runs with various parameterizations, systematically varying resolution for *Phaser* rotation and translation (McCoy *et al.*, 2005) searches as well as the r.m.s.d estimation. One of the models cut out from the 1ZBR template, composed of the poly-Ala-trimmed fifth and first repeats stood out for producing a unique rotation cluster and a low number of solutions with higher LLG than any other trial, for a rotation resolution cut-off of 2.1Å and translation cut-off of 1.7Å. Still, its expansion would not yield a solution. The top LLG solution for this model was used as reference to cluster phases from all other 350 solutions produced by the pool of five models. From the 350 phase sets derived from these models after five cycles density modification, one solution produced by a model derived from the fourth and fifth repeats was matching within a tolerance of 60º. The merged solutions succeeded in rendering a trace of 368 residues with a CC of 40.14%.

After the structure of the wild-type PPAD was solved by our group, it was used by our collaborators as a model for MR for a series of mutants and complexes. The analysis of the structure and function of PPAD in these different functional states allowed them to propose the structural relation between PPAD and the AgDIs (agmatine deiminases). Moreover, they demonstrated for several substrates that this enzyme has developed a unique function among citrullinating enzymes, which is the deimination of peptides with a C-terminal arginine. This activity helps the pathogenic bacteria to modify endogenous proteins by citrullination thus generating epitopes that are not recognised by the host immune system. This might aggravate inflammation by initiation of autoimmune reactions, contributing to both RA and other inflammatory diseases.

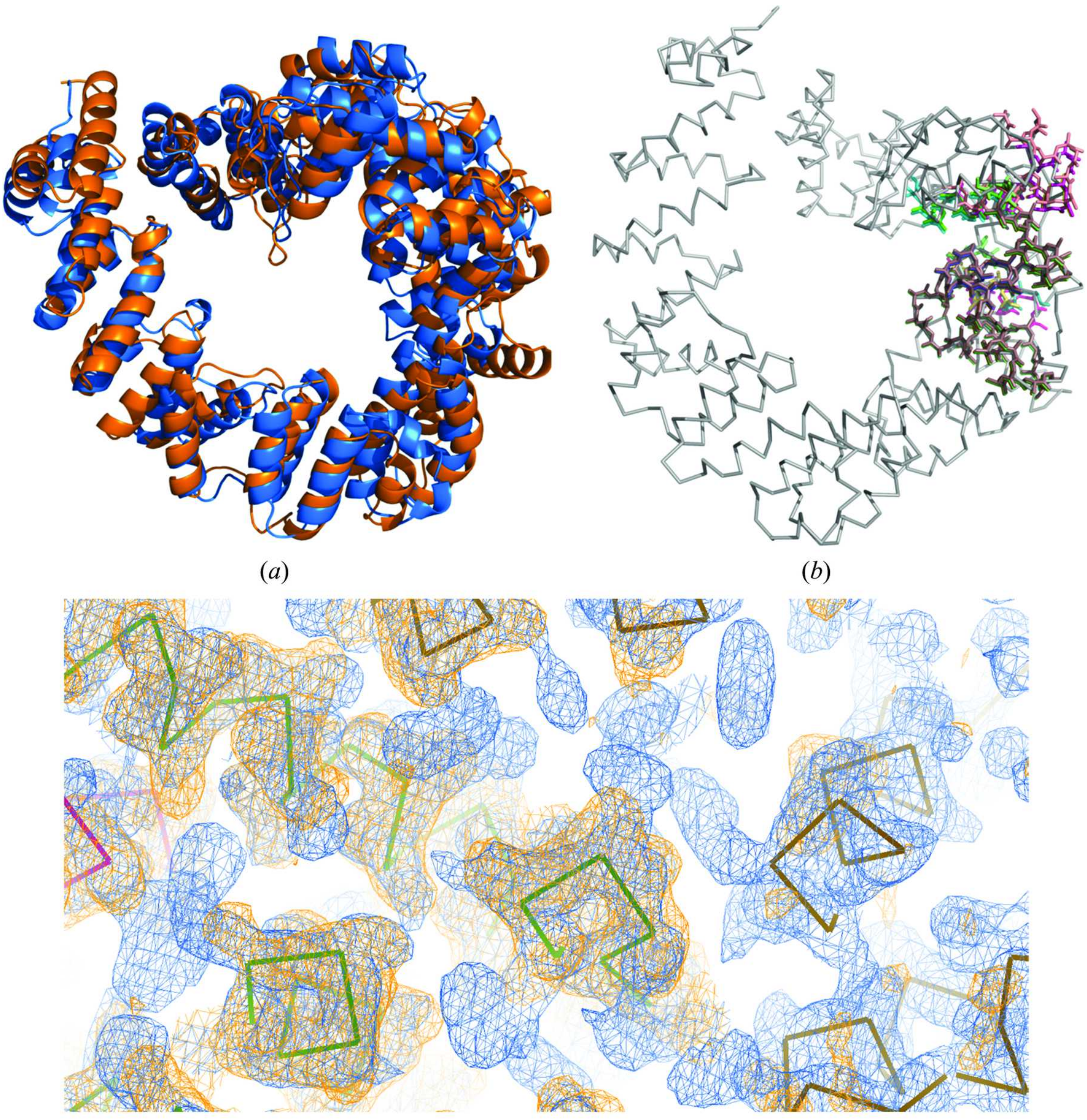
5.3 LTG

LTG is a soluble lytic transglycosylase from the bacterial pathogen *Pseudomonas aeruginosa* (Lee et al., 2018). This pathogen is fought by using β-Lactam antibiotics, that prevent bacterial cell wall from crosslinking, which leads to the accumulation of long non-crosslinked strands of peptidoglycan. The lytic transglycosylase Slt is the enzyme that P. aeruginosa uses in an attempt to repair this aberrantly formed peptidoglycan. Native structure is deposited under the PDB ID 5OHU, with a monomer of 613 residues in the ASU. Datasets were collected on the ALBA beamline XALOC (Juanhuix *et al.*, 2014), and their statistics are found in Table 5.

Table 5 LTG data statistics

	LTG
Space group	$P6_3$
Unit cell parameters	
a (Å)	163.98
b (Å)	163.98
c (Å)	56.71
α (º) , β (º) , γ (º)	90, 90, 120
Resolution (Å)	47.3 – 2.2 (2.28 – 2.20)
I/σ(I)	14.8 (1.35)
Completeness (%)	99.98 (100.0)

A homology search for the target sequence using HHPRED (Söding *et al.*, 2005) provided a list of possible templates for MR. The best scoring model was another soluble lytic transglycosylase, SLT70 from *Escherichia coli* (PDB ID 1QSA), with 31% sequence identity. The estimated VRMS for this degree of conservation is 1.5 Å but on account of its flexibility, the r.m.s.d. of the final structure with respect to the 1QSA model is 4.6 Å, as computed with the PyMOL super algorithm on a core of 582 residues. Figure 21 shows the superposition of final structure and template (a), fragments used in the solution (b) and a detail of the electron density maps before and after expansion (c).



(*a*)

(*b*)

(*c*)

Figure 21 Original solution of LTG

(a) Final structure (blue) versus the template used in ARCIMBOLDO_ SHREDDER (orange). The r.m.s.d. between the structures is 4.6 Å over a core of 582 C atoms. (b) Colored sticks show the solving fragments that clustered together and the black ribbon shows the final structure. (c) A detail of the SHELXE Fo FOM electron-density maps with the C trace. Orange, initial map from phase combination; blue, final map after density modification and autotracing; both are contoured at 1 σ.

The structure was originally solved with ARCIMBOLDO_SHREDDER with the first implementation of the spherical mode, described herein. The full pdb of 1QSA was used as the initial template, preserving the coil regions and the original B-factors, but trimming the side chains to alanine. Spheres of 20Å radius centred on each amino acid of the template were defined, without further modification, to extract 619 models. Those models ranged in size from 42 to 177 residues, making FOMs not directly comparable across fragments. It should be stressed that all models are naturally superimposed on the template they derive from and correspond to different parts of a common fold. Therefore, they can be input as a library into ARCIMBOLDO_BORGES. Similar rotations would map fragments to consistent regions of the target structure if the original fold was maintained. Moreover, partially overlapping solutions, if produced, should be found within one rotation cluster and their maps could be combined to improve the starting phases. In this case, one of the rotation clusters stood out through solutions with TFZ-scores above 8. These solutions were used as references to cluster phases. One of the combined phase sets developed into a full solution, with a CC of 48.08% and 563 residues traced in 7 chains. All 12 models so grouped were targeting the same region of the query structure, corresponding to residues 478 to 592 in the template.

5.4 HHED2

Hhed2 is a 230 amino acids long halohydrin-dehalogenase (Schallmey *et al.*, 2014) from *Gamma proteobacterium*. Data to a resolution of 1.6Å were available in space group $P2_12_12_1$ with four monomers in the asymmetric unit totalling 920 residues. An homology search for the target sequence using HHPRED provided a list of possible templates for MR, sharing a typical Rossmann fold, characterised by a series of alternating β-strand and α-helical segments with the β-strands arranged into a parallel β-sheet.

Three homologs were selected, two from the same family of dehalogenases; HhedB (PDB ID 4ZD6) with a sequence identity of 47%, and HheA (PDB ID 4Z9F) with a sequence identity of 30% (Watanabe *et al.*, 2015) and one from the same superfamily of short-chain dehydrogenase reductases (SDR); EbN1 (Büsing *et al.*, 2015), with 26% sequence identity (PDB ID 4URF).

All three templates lead to a successful solution. The two dehalogenases show r.m.s.d. to the target structure over a core of 185 Cα atoms of 0.7 (4Z9F) and 1.12Å (4ZD6), respectively, for the SDR 4URF, the r.m.s.d. over a core of 149 Cα atoms is 1.3Å. Templates were trimmed, removing short α-helices of less than 7 residues, β-strands of less than 4 residues and coil regions. The annotation for the first gyre cycle leaves the central β-sheet present in the fold as a single, indivisible group. The second level of annotations separated the helices as independent groups. In all cases, the rotation search and first cycle of gyre refinement were performed at 0.8Å RMSD. The second cycle of gyre refinement and subsequent *Phaser* steps were performed at 0.5Å RMSD. The size of the search fragments was set in order to achieve a target eLLG of 60 at the last RMSD set in the run (0.5Å).

The template derived from 4ZD6 is so close to the target structure that solution is trivial. Fragments derived from this model are correctly placed corresponding to all four monomers in the asymmetric unit, although approximate alignment of non-crystallographic and crystallographic symmetry axes leads to three, rather than four rotation clusters. All best scoring fragments have been improved by gyre and gimble. Consistent solutions were combined using as a reference the best scoring solution, characterized by a TFZ score of 12.6. Two consecutive combination steps setting MPD thresholds of 60 and 87° identify remaining correct solutions placed on the same and different monomers, respectively. This phase set, when submitted to

SHELXE for density modification and autotracing, solves the structure and reaches a CC of 37.99%, with 859 residues traced in 13 chains.

5.5 SYCP1-αC-end

SYCP1 (Dunce *et al.*, 2018) is a protein that forms part of the synaptonemal complex. This is a supramolecular protein assembly that plays a relevant role in the reduction in chromosome number during meiosis. The αC-end part from the protein (676-770) was crystallised in two different crystal forms: *C*2 (diffraction data to 2.15Å, four SYPC1 chains per ASU, PDB ID 6F63) and $I4_122$ (diffraction data to 2.48Å, one SYPC1 chain per ASU, PDB ID 6F64). While structure solution of the orthorhombic crystal form was achieved employing the coiled-coil mode in ARCIMBOLDO_LITE (Caballero *et al.*, 2018), data in *C*2 yielded a solution using the spherical shredding in ARCIMBOLDO_SHREDDER (Millán *et al.*, 2018). For that purpose, the $I4_122$ crystal form was used as a starting template for generating 74 models containing 99 amino acids each. A phase set combining 25 partial solutions expanded into a full solution, recognisable by a CC of 48.2%

5.6 Other previously unknown structures

During the time of the research presented in this PhD thesis, other previously unknown structures have been solved using the methods described. In particular, I have solved the structures of:

- A bacterial enzyme with 1224 residues in the ASU and resolution up to 2.3Å, solved with ARCIMBOLDO_SHREDDER starting from a template with 34% sequence identity.
- A fungal enzyme with 294 residues in the ASU and resolution up to 1.5Å, solved with ARCIMBOLDO_SHREDDER starting from a template with 19% sequence identity and combining the solutions with ALIXE.
- A domain from a plant protein with 692 residues in the ASU and resolution up to 2.5Å, solved with ARCIMBOLDO_SHREDDER starting from a template with 21% sequence identity.
- A complex of antibodies, with 363 residues in the ASU and resolution up to 1.6Å, solved with ARCIMBOLDO_BORGES using a library of β-sandwich folds, and combining the solutions with ALIXE.

Finally, shortly after the publication of the spherical mode, in June 2018, two external users phased their structures using ARCIMBOLDO_SHREDDER. One was a protein of about 180 residues from data up to 1.74Å and with a model with 19% sequence identity, and 2.5Å r.m.s.d. to the final structure. The second case contained two copies of a 260 residues monomer with data up to 1.45Å, and was solved starting from a template with 20% sequence identity and 2.4Å r.m.s.d. to the final structure.

5.7 Assessing the generality of SHREDDER and ALIXE

A set of 43 test cases representing different space groups (described in section 2.6) has been used to test the generality of the algorithms in ALIXE and in ARCIMBOLDO_SHREDDER. Concomitantly different hardware was probed. Some of the tests were performed in our local Condor grid, and some others in a MacBook

Pro with 4 cores (both described in section 2.5). A summary of the cases and the parameterisation used is shown in Table 6.

Table 6 Tests with ARCIMBOLDO_SHREDDER and ALIXE

Summary of the parameterization used in the tests. 23 cases were run in the local grid and 20 on an 4-core machine.

Computing	ID PDB	Template model from	% sequence identity	r.m.s.d to final	Shredder strategy	Result
Local grid	1O5J	3AHP, A	23	1	RMSD 1.0	solved
Local grid	5VOG	5BQP, C	28	1	default	solved
Local grid	2AIF	3VI6	18	1.1	RMSD 0.6	solved
Local grid	5W2G	4FRF, A	28	1.3	default	solved
Local grid	5NA1	5YJW, A	25	1.4	default	expansion issues
Local grid	3F4W	3RJ2	25	1.5	default	expansion issues
Local grid	4PYI	5KVA	24	1.5	default	solved
Local grid	1S6Y	3FEF, B	23	1.7	default	expansion issues
Local grid	4MH4	2ONF, B	18	1.7	default, RMSD 0.8	not solved
Local grid	3VPE	2VW8	20	1.7	default, RMSD 0.8	not solved
Local grid	4DB5	5L19	17	1.8	default	solved
Local grid	1NNH	4J15, A	22	1.8	default	solved
Local grid	2QG3	2DVK	28	2.2	RMSD 0.8	expansion issues
Local grid	3HP4	1YZF	22	2.4	default	solved
Local grid	3MYI	4IGG	31	2.4	default	solved
Local grid	3GH6	2V6K, B	19	2.5	default	solved
Local grid	1V6T	2NLY	14	2.8	default	not solved
Local grid	2YG5	5TTJ, A	20	3	default, RMSD 0.8	not solved
Local grid	5M3Y	1QMN, A	20	3.3	default	not solved
Local grid	4J2F	5U56, C	19	3.3	default	expansion issues
Local grid	2ODL	4RM6	28	3.3	maintain coil	solved
Local grid	4CZL	2V7Y	26	3.7	default	solved
Local grid	3CYO	3K9A	73	4.5	default + coiled coil	solved
Multiprocessing	5H7E	1VGJ	28	1.1	RMSD 0.6 + maintain coil	solved
Multiprocessing	1UQ4	4M1U, A	19	1.1	default	not solved
Multiprocessing	4ROT	5VOL, H	20	1.2	default	solved
Multiprocessing	5IX3	2VI7, C	20	1.3	default	solved
Multiprocessing	3T1S	3KYE, B	15	1.4	default	solved
Multiprocessing	5UDN	4CT3, A	23	1.4	default	not solved
Multiprocessing	2QX2	3IB5	28	1.5	default	solved
Multiprocessing	2QCK	3PFT, B	23	1.6	default	expansion issues

Multiprocessing	4CSV	4CFH, A	26	1.6	maintain coil + RMSD 0.8	expansion issues
Multiprocessing	3KWR	4P78, B	18	1.7	default, RMSD 0.8	not solved
Multiprocessing	2QCV	1TYY, A	23	1.7	default	solved
Multiprocessing	1SS4	5UHJ, A	15	1.8	RMSD 0.6	not solved
Multiprocessing	3MN2	3OOU, A	21	2.1	default, RMSD 0.4, RMSD 0.8	expansion issues
Multiprocessing	2G2D	1NIG	17	2.2	RMSD 0.8 + coiled_coil	solved
Multiprocessing	5O7G	3PFB, A	12	2.3	default	solved
Multiprocessing	3OU2	3LCC	16	2.3	default	solved
Multiprocessing	5HGN	4PH0, A	21	2.3	default	solved
Multiprocessing	2HYT	5MWR, B	15	3.2	default, RMSD 0.8	not solved
Multiprocessing	5G4Z	3UB6, A	21	3.7	default, RMSD 0.4, RMSD 0.6, RMSD 0.8	not solved
Multiprocessing	2V71	2EFR	17	5.8	default	expansion issues

Out of the 43 cases, 23 were solved completely, that is, achieving a complete trace (CC > 30%) and map after the density modification and autotracing step. Nine more cases presented non-random solutions (with wMPE < 80º) that were not progressing in the expansion step.

Regarding parameterisation in the spherical mode of ARCIMBOLDO_SHREDDER, 17 tests have been solved with the current defaults, and the remaining five have required either the decrease of the starting RMSD, the preservation of the coil in the model or the activation of the coiled coil mode. The true r.m.s.d to final structure in the successful solutions ranges between 1 to 4.5Å, and the sequence identity is between 12 to 73%. The 4.5Å r.m.s.d. and 73% sequence identity are from a coiled coil structure in which the monomer, which is an helix, is really similar, and it is the arrangement of the multiple copies in the ASU that varies.

Regarding the use of ALIXE, of the 23 solved, 15 were solved by the combination of phases of multiple correct solutions. Of the remaining cases one was in *P*1 space group, and four only had a single correct solution. Lastly, three produced merged phase sets that, although correct, would not reach CC larger than 30% at the expansions, whereas some of the single solutions would. The successful cases represent symmetries of all crystal systems but for triclinic, and include ten different point groups.

With respect to hardware, no significant differences are found in terms of success, as in the local grid, 12 solved, five did not and five had issues in expansion, whereas in multiprocessing 11 solved, seven did not and three had problems in expansion.

CONCLUSIONS

All the objectives we set at the beginning of this thesis have been achieved, leading to the following conclusions:

The program ALIXE has been designed to explore phase combination and developed to provide a method to solve challenging structures by maximizing the information derived from partial correct solutions. In particular:

- A systematic characterisation of correct, yet partial solutions, has been performed. It established that for the kind of fragments typically used in ARCIMBOLDO, mean phase errors range from 70 to 85º.
- Weighted mean phase differences have been used as a measure for the similarity between phase sets, once referred to the same symmetry origin. Values below 60º characterise overlapping fragments while values barely better than random (as high as 87º) characterise non-overlapping fragments.
- Even for starting phase sets where density modification does not improve the average phases and lead to a solution, it enhances the discrimination of the correct origin shift relating such sets. We proposed this to be related to density modification bringing forth boundaries between solvent and protein regions.
- In all cases examined, the general resolution limit of 2.0Å set for comparison and merging of the phase sets, has been seen to provide appropriate trade-off between highlighting differences at higher resolution and losing detail at lower resolution.
- The procedure for finding the correct shift relating phase sets for non-polar space groups involves testing of all possible shifts and selecting the one yielding the lowest mean phase difference. This has been shown to properly discriminate the right origin shift even for highly dissimilar phase sets.
- For polar space groups, an approximation to get the optimal origin shift is provided, which has been shown to be successful in the case of overlapping fragments, but problematic for phase sets from non-overlapping solutions with large differences, where the correct origin shift is frequently missed. In view of the results with polar space groups, triclinic *P*1 has not been further analysed.
- The clear cut difference between overlapping and non-overlapping cases has been used to develop a two steps strategy for phase combination. A first step of phase combination at a tolerance of 60º is performed within rotation clusters to find overlapping solutions. Then, if multiple monomers are expected in the asymmetric unit, a second step with a higher tolerance (87º) is attempted, aiming to merge solutions from different monomers. This strategy minimizes the incorrect clustering and effectively reduces computing time.
- ALIXE can be used integrated within ARCIMBOLDO for fragment-based MR, or as a standalone tool that can be used independently (for example, to compare solutions).
- Tests with ARCIMBOLDO_LITE showed that combining single solutions from the first round of fragment placement led to a complete

solution in cases where the structure could not be solved expanding from a single fragment.

- Partial ARCIMBOLDO_BORGES overlapping and non-overlapping solutions produced with different models from libraries representing a β-sandwich fold, have been successfully merged with ALIXE and resulted in the solution of a previously unknown structure.
- Reconstitution in reciprocal space with ALIXE of larger hypotheses from partially overlapping fragments in ARCIMBOLDO_SHREDDER has been shown to successfully handle the geometrical deviations and enhance the common regions.

The spherical mode for ARCIMBOLDO_SHREDDER has been designed to address the frequent challenge of using distant homologs as search models in molecular replacement. This software provides a method to derive compact fragments from a template, optimise them by internal refinement and/or trimming and combine consistent solutions. In particular:

- The generation of compact models was initially done by computing a sphere of a given radius centred on each residue in the template, the model comprising all residues falling within the volume of the sphere. This prototype provided a baseline strategy and from that point, the algorithm was improved by considering distances between residues, continuity in secondary structure elements, fold association, and controlling model size through eLLG.
- The annotation of the template in terms of secondary and tertiary structure allows to identify structural information, which will be inherited in the treatment of the small fragments extracted. For example, coiled regions are highly flexible and removal is often required for success. Helices, on the other hand, are rather rigid and refine well as independent bodies. Beta sheets, on the contrary, are best kept packed in small groups.
- The models are further decomposed to give them additional degrees of freedom according to the annotation. Refinement against the rotation (*Phaser*'s *gyre*) or the translation (*Phaser*'s *gymble*) target functions has been seen to bring the models closer with respect to the final structure.
- Alternatively, it has been found effective to trim and refine the models when coil residues have been left (*Phaser*'s LLG-guided pruning, mend after translation).

The algorithms in ALIXE and ARCIMBOLDO_SHREDDER have been extensively tested and subsequently used for the solution of previously unknown structures. The implementations have proven general when tested on 54 test cases representing a range of folds, resolution, and space groups. A default parameterisation has been concluded. The work presented here has provided a guide on how to steer this parameterisation depending on data and model characteristics.

8 previously unknown structures have been solved by the author using the methods described in this work:

- LTG, an helical solenoid of 610 residues, was solved using a template model that shared a similar fold overall, but on account of its flexibility, had an r.m.s.d. to the final structure of 4.6Å, a value far beyond the convergence of MR.

- In PPAD, a structure with a high fraction of coil, the availability of various homologs and the systematic testing of different parameterisations produced a pool of solutions from which only one stood out in discrimination. This solution provided a reference phase set that was used to merge overlapping solutions, which succeeded in phasing.
- Hhed2, a structure with four monomers of 230 residues in the ASU, was solved by exploiting the two steps phase combination strategy to merge solutions from different monomers, and merging the information from a single fragment search.
- SYCP1, a coiled coil domain with four monomers in the ASU, was solved by exploiting the fold information from a solution in a different crystal form with one monomer in the ASU.
- An enzyme with 1224 residues in the ASU was solved with SHREDDER, representing the largest structure solved until now with the ARCIMBOLDO methods.
- A fungal enzyme was solved deriving small fragments yielding an eLLG of 30 from a template with only 19% sequence identity and combining the partial solutions with ALIXE.
- A solution for a domain of an helical plant protein structure containing two monomers of 350 residues was achieved with SHREDDER, which substantially improved partial models with gyre and gimble refinement. Combination in ALIXE and expansion led to a solution.
- A structure with four immunoglobulin-like domains that had not been solved using complete models due to the variability that characterises such folds, was solved with libraries of β-sandwich folds and combining the solutions in reciprocal space.
- Usability of the software has been demonstrated by the fact that two independent users have phased their structures using the spherical mode in ARCIMBOLDO_SHREDDER, shortly after its publication. Both cases used templates with less than 20% sequence identity and r.m.s.d over 2.4Å

Finally, the software derived from this work has been made available via our website, and via CCP4. Both programs have been published and presented in specialised meetings and workshops, thus providing the crystallographic community with new methods to undertake structure solution of challenging cases.

OUTLOOK

A PhD thesis is hardly ever considered as finished by its author. The objectives of the work we set to do have been achieved. Furthermore, they have raised new questions and opened new exciting possibilities to investigate in the future. I would like to propose some of them:

ALIXE:

- MPDs provide a good metric for differences between maps that we have used successfully, but further exploration could be done on the use of correlation coefficients of different type.
- Currently, all phase sets that present a MPD with respect to the reference below the threshold are accepted in a cluster. Use of the variance within a cluster to limit inclusion of phase sets or clustering in successive steps of smaller tolerances could also be explored.
- In coiled coils, the application of ALIXE can help in discriminating solutions and reducing computing effort. Further studies can be done to implement this in an automatic manner for the coiled-coil mode.
- Currently, ALIXE does not support maps originating from different datasets. However, this could be of interest for example to compare solutions obtained with fragments in different datasets, or even between experimental maps.

ARCIMBOLDO_SHREDDER:

- Often, more than one distant homolog is available to use as template, and sequence alignment and secondary structure prediction based scoring do not clearly indicate which model will succeed. Therefore, finding ways to evaluate all different models jointly against the experimental data would be interesting.
- The use of ensembles within ARCIMBOLDO_SHREDDER could also be explored.
- We have explored only a few modifications of the starting template, and in general, the best strategy has been to trim the templates to polyalanine and set all B-factors to a constant value. Other modifications could be conducted and might make a difference for borderline cases. For example:
  - Including the possibility of maintaining particular sidechain classes.
  - Leaving the original B-factors from the distant homolog.
  - Setting the B-factors of the secondary structure elements according to their expected rigidity or the expected errors of their coordinates.
- In terms of model refinement strategies, normal mode analysis (NMA) is available through *Phaser* and implemented within ARCIMBOLDO but has not been used in the context of SHREDDER, and should be explored.
- A strategy parallel to what ARCIMBOLDO_SHREDDER does starting from a template model, could be implemented using maps.

REFERENCES

Abrahams, J. P. & Leslie, A. G. (1996). *Acta Crystallogr D Biol Crystallogr* **52**, 30-42.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res* **25**, 3389-3402.

Artola-Recolons, C., Carrasco-Lopez, C., Llarrull, L. I., Kumarasiri, M., Lastochkin, E., Martinez de Ilarduya, I., Meindl, K., Usón, I., Mobashery, S. & Hermoso, J. A. (2011). *Biochemistry* **50**, 2384-2386.

Astbury, W. T. & Marwick, T. C. (1932). *Nature* **130**, 309-310.

Banci, L., Bertini, I., Calderone, V., Cefaro, C., Ciofi-Baffoni, S., Gallo, A., Kallergi, E., Lionaki, E., Pozidis, C. & Tokatlidis, K. (2011). *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4811-4816.

Bentley, G. A. & Houdusse, A. (1992). *Acta Crystallographica Section A Foundations of Crystallography* **48**, 312-322.

Berman, H. M. (2008). *Acta Crystallogr A* **64**, 88-95.

Berman, H. M., Coimbatore Narayanan, B., Di Costanzo, L., Dutta, S., Ghosh, S., Hudson, B. P., Lawson, C. L., Peisach, E., Prlic, A., Rose, P. W., Shao, C., Yang, H., Young, J. & Zardecki, C. (2013). *FEBS Lett* **587**, 1036-1045.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Research* **28**, 235-242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *European Journal of Biochemistry* **80**, 319-324.

Bieniossek, C., Schutz, P., Bumann, M., Limacher, A., Uson, I. & Baumann, U. (2006). *J Mol Biol* **360**, 457-465.

Bosh, A. S., Elliot, J. L., Kruse, S. E., Baron, R. L., Dunham, E. W. & French, L. M. (1986). *Icarus* **66**, 556-560.

Bragg, W. H. & Bragg, W. L. (1913). *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **88**, 428-438.

Bragg, W. L. (1913). *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **89**, 248-277.

Brünger, A. T. (1990). *Acta Crystallographica Section A Foundations of Crystallography* **46**, 46-57.

Brzuszkiewicz, A., Nowak, E., Dauter, Z., Dauter, M., Cieslinski, H., Dlugolecka, A. & Kur, J. (2009). *Acta Crystallogr Sect F Struct Biol Cryst Commun* **65**, 862-865.

Buehler, A., Urzhumtseva, L., Lunin, V. Y. & Urzhumtsev, A. (2009). *Acta Crystallogr D Biol Crystallogr* **65**, 644-650.

Bunkoczi, G., Echols, N., McCoy, A. J., Oeffner, R. D., Adams, P. D. & Read, R. J. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2276-2286.

Bunkoczi, G. & Read, R. J. (2011). *Acta Crystallogr D Biol Crystallogr* **67**, 303-312.

Bunkoczi, G., Wallner, B. & Read, R. J. (2015). *Structure* **23**, 397-406.

Burla, M. C., Giacovazzo, C. & Polidori, G. (2010). *Journal of Applied Crystallography* **43**, 825-836.

Büsing, I., Höffken, H. W., Breuer, M., Wöhlbrand, L., Hauer, B. & Rabus, R. (2015). *Journal of Molecular Microbiology and Biotechnology* **25**, 327-339.

Caballero, I., Sammito, M., Millán, C., Lebedev, A., Soler, N. & Usón, I. (2018). *Acta Crystallogr D Struct Biol* **74**, 194-204.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *Journal of Applied Crystallography* **41**, 548-553.

Carrell, H. L., Hoier, H. & Glusker, J. P. (1994). *Acta Crystallographica Section D* **50**, 113-123.

Chakraborty, M. & Das, S. (2012). *Procedia Technology* **4**, 830-833.

Chothia, C. & Lesk, A. M. (1986). *The EMBO Journal* **5**, 823-826.

Clauset, A., Newman, M. E. J. & Moore, C. (2004). *Physical Review E* **70**, 066111.

Clothia, C. & Lesk, A. (1986). *EMBO J.*

Colman, Fehlhammer & Bartels (1976). *Crystallographic Computing Techniques*. Copenhagen: Munksgaard.

Cooley, J. W. & Tukey, J. W. (1965). *Mathematics of computation* **19**, 297-301.

Cowtan, K. (1994). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* **31**, 34-38.

Cowtan, K. (2006). *Acta Crystallogr D Biol Crystallogr* **62**, 1002-1011.

Cowtan, K. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 470-478.

Cowtan, K. D. & Main, P. (1993). *Acta Crystallogr D Biol Crystallogr* **49**, 148-157.

Cowtan, K. D. & Zhang, K. Y. (1999). *Prog Biophys Mol Biol* **72**, 245-270.

DeLano, W. L. & Brunger, A. T. (1995). *Acta Crystallogr D Biol Crystallogr* **51**, 740-748.

Diederichs, K. (2017). *Acta Crystallogr D Struct Biol* **73**, 286-293.

DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwai, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature* **473**, 540-543.

Doesburg, H. M. & Beurskens, P. T. (1983). *Acta Crystallographica Section A Foundations of Crystallography* **39**, 368-376.

Doolittle, R. F. (1981). *Science* **214**, 149-159.

Dunce, J. M., Dunne, O. M., Ratcliff, M., Millán, C., Madgwick, S., Usón, I. & Davies, O. R. (2018). *Nature Structural & Molecular Biology*.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 486-501.

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Crystallogr D Biol Crystallogr* **56**, 1137-1147.

Franklin, R. E. & Gosling, R. G. (1953). *Nature* **171**, 740-741.

Friedrich, W., Knipping, P. & Laue, M. (1913). *Annalen der Physik* **346**, 971-988.

Fujinaga, M. & Read, R. J. (1987). *Journal of Applied Crystallography* **20**, 517-521.

G., C. & T., N. (2006). *InterJournal* **Complex Systems**, 1695.

Giacovazzo, C., Monaco, H. L., Artioli, G., Viterbo, D., Milanesio, M., Gilli, G., Gilli, P., Zanotti, G., Ferraris, G. & Catti, M. (2011). *Fundamentals of Crystallography*. OUP Oxford.

Gildea, R. J. & Winter, G. (2018). *Acta Crystallogr D Struct Biol* **74**, 405-410.

Glykos, N. M. & Kokkinidis, M. (2003). *Acta Crystallographica Section D Biological Crystallography* **59**, 709-718.

Goulas, T., Mizgalska, D., Garcia-Ferrer, I., Kantyka, T., Guevara, T., Szmigielski, B., Sroka, A., Millán, C., Usón, I., Veillard, F., Potempa, B., Mydel, P., Sola, M., Potempa, J. & Gomis-Ruth, F. X. (2015). *Sci Rep* **5**, 11969.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **225**, 287-307.

Grosse-Kunstleve, R. W. & Adams, P. D. (2001). *Acta Crystallogr D Biol Crystallogr* **57**, 1390-1396.

Gudmann, N. S., Hansen, N. U., Jensen, A. C., Karsdal, M. A. & Siebuhr, A. S. (2015). *Autoimmunity* **48**, 73-79.

Harker, D. (1936). *The Journal of Chemical Physics* **4**, 381-390.

Harker, D. (1956). *Acta Crystallographica* **9**, 1-9.

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

Heller, I., Mannik, J., Lemay, S. G. & Dekker, C. (2009). *Nano Lett* **9**, 377-382.

Hendrickson, W. A. (1991). *Science* **254**, 51-58.

Hendrickson, W. A. (2014). *Q Rev Biophys* **47**, 49-93.

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature* **290**, 107.

Hodgkin, D. C. (1949). *Adv Sci* **6**, 85-89.

Huber, R. (1965). *Acta Crystallographica* **19**, 353-356.

Hyberts, S. G., Robson, S. A. & Wagner, G. (2013). *J Biomol NMR* **55**, 167-178.

International Tables for Crystallography Volume C: Mathematical, Physical and Chemical Tables, 2004). Kluwer Academic Publishers.

Jia-xing, Y., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Crystallogr D Biol Crystallogr* **61**, 1465-1475.

Jinek, M., Rehwinkel, J., Lazarus, B. D., Izaurralde, E., Hanover, J. A. & Conti, E. (2004). *Nat Struct Mol Biol* **11**, 1001-1007.

Johnson, L. S., Eddy, S. R. & Portugaly, E. (2010). *BMC Bioinformatics* **11**, 431.

Jones, J. E., Causey, C. P., Lovelace, L., Knuckley, B., Flick, H., Lebioda, L. & Thompson, P. R. (2010). *Bioorg Chem* **38**, 62-73.

Juanhuix, J., Gil-Ortiz, F., Cuni, G., Colldelram, C., Nicolas, J., Lidon, J., Boter, E., Ruget, C., Ferrer, S. & Benach, J. (2014). *J Synchrotron Radiat* **21**, 679-689.

Kabsch, W. & Sander, C. (1983a). *Biopolymers* **22**, 2577-2637.

Kabsch, W. & Sander, C. (1983b). *Biopolymers* **22**, 2577-2637.

Kalman, R. E. (1960). *Journal of Basic Engineering* **82**, 35-45.

Karle, J. & Hauptman, H. (1950). *Acta Crystallographica* **3**, 181-187.

Keegan, R. M., McNicholas, S. J., Thomas, J. M. H., Simpkin, A. J., Simkovic, F., Uski, V., Ballard, C. C., Winn, M. D., Wilson, K. S. & Rigden, D. J. (2018). *Acta Crystallogr D Struct Biol* **74**, 167-182.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. & Phillips, D. C. (1958). *Nature* **181**, 662-666.

Kleywegt, G. J. & Read, R. J. (1997). *Structure* **5**, 1557-1569.

Koopmeiners, J., Diederich, C., Solarczek, J., Voß, H., Mayer, J., Blankenfeldt, W. & Schallmey, A. (2017). *ACS Catalysis* **7**, 6877-6886.

Krissinel, E. (2007). *Bioinformatics* **23**, 717-723.

Krissinel, E. & Henrick, K. (2004). *Acta Crystallogr D Biol Crystallogr* **60**, 2256-2268.

Laue, M. (1913). *Annalen der Physik* **346**, 989-1002.

Leahy, D. J., Axel, R. & Hendrickson, W. A. (1992). *Cell* **68**, 1145-1162.

Lebedev, A. A., Vagin, A. A. & Murshudov, G. N. (2008). *Acta Crystallogr D Biol Crystallogr* **64**, 33-39.

Lee, M., Batuecas, M. T., Tomoshige, S., Dominguez-Gil, T., Mahasenan, K. V., Dik, D. A., Hesek, D., Millán, C., Usón, I., Lastochkin, E., Hermoso, J. A. & Mobashery, S. (2018). *Proc Natl Acad Sci U S A* **115**, 4393-4398.

Li, J., Qian, X., Hu, J. & Sha, B. (2009). *J Biol Chem* **284**, 23852-23859.

Lunin, V. Y. & Lunina, N. L. (1996). *Acta Crystallogr A*.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Crystallogr D Biol Crystallogr* **51**, 896-903.

Lunin, V. Y., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Crystallographica Section A Foundations of Crystallography* **46**, 540-544.

Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Crystallogr D Biol Crystallogr* **49**, 530-533.

Maresz, K. J., Hellvard, A., Sroka, A., Adamowicz, K., Bielecka, E., Koziel, J., Gawron, K., Mizgalska, D., Marcinska, K. A., Benedyk, M., Pyrc, K., Quirke, A. M., Jonsson, R., Alzabin, S., Venables, P. J., Nguyen, K. A., Mydel, P. & Potempa, J. (2013). *PLoS Pathog* **9**, e1003627.

McCoy, A. J. (2017). *Methods Mol Biol* **1607**, 421-453.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J Appl Crystallogr* **40**, 658-674.

McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C. & Read, R. J. (2005). *Acta Crystallogr D Biol Crystallogr* **61**, 458-464.

McCoy, A. J., Nicholls, R. A. & Schneider, T. R. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2216-2225.

McCoy, A. J., Oeffner, R. D., Millán, C., Sammito, M., Usón, I. & Read, R. J. (2018). *Acta Crystallogr D Struct Biol* **74**, 279-289.

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Proc Natl Acad Sci U S A* **114**, 3637-3641.

Millán, C., Sammito, M., Garcia-Ferrer, I., Goulas, T., Sheldrick, G. M. & Usón, I. (2015). *Acta Crystallogr D Biol Crystallogr* **71**, 1931-1945.

Millán, C., Sammito, M. & Usón, I. (2015). *IUCrJ* **2**, 95-105.

Millán, C., Sammito, M. D., McCoy, A. J., Nascimento, A. F. Z., Petrillo, G., Oeffner, R. D., Dominguez-Gil, T., Hermoso, J. A., Read, R. J. & Usón, I. (2018). *Acta Crystallogr D Struct Biol* **74**, 290-304.

Miller, R., Gallo, S. M., Khalak, H. G. & Weeks, C. M. (1994). *Journal of Applied Crystallography* **27**, 613-621.

Miller, W. H. (1839). *A treatise on crystallography [microform] / by W.H. Miller*. Cambridge [Eng.] : London: Printed for J. & J.J. Deighton ; J.W. Parker.

Mizuguchi, K. & Blundell, T. L. (2000). *Bioinformatics* **16**, 1111-1119.

Navaza, J. (1994). *Acta Crystallographica Section A Foundations of Crystallography* **50**, 157-163.

Nurizzo, D., Mairs, T., Guijarro, M., Rey, V., Meyer, J., Fajardo, P., Chavanne, J., Biasci, J. C., McSweeney, S. & Mitchell, E. (2006). *J Synchrotron Radiat* **13**, 227-238.

Oeffner, R. D., Afonine, P. V., Millán, C., Sammito, M., Usón, I., Read, R. J. & McCoy, A. J. (2018). *Acta Crystallogr D Struct Biol* **74**, 245-255.

Oeffner, R. D., Bunkoczi, G., McCoy, A. J. & Read, R. J. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2209-2215.

Oszlanyi, G. & Suto, A. (2004). *Acta Crystallogr A* **60**, 134-141.

Patterson, A. L. (1934). *Physical Review* **46**, 372-376.

Patterson, A. L. (1944). *Physical Review* **65**, 195-201.

Pauling, L. & Corey, R. B. (1951). *Proc Natl Acad Sci U S A* **37**, 251-256.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). *Proc Natl Acad Sci U S A* **37**, 205-211.

Pearson, W. R. (1991). *Genomics* **11**, 635-650.

Pearson, W. R. & Lipman, D. J. (1988). *Proc Natl Acad Sci U S A* **85**, 2444-2448.

Perrakis, A., Harkiolaki, M., Wilson, K. S. & Lamzin, V. S. (2001). *Acta Crystallogr D Biol Crystallogr* **57**, 1445-1450.

Pieper, U., Kapadia, G., Mevarech, M. & Herzberg, O. (1998). *Structure* **6**, 75-88.

Pons, P. & Latapy, M. (2005). *Computer and Information Sciences - ISCIS 2005*, edited by p. Yolum, T. Güngör, F. Gürgen & C. Özturan, pp. 284-293. Berlin, Heidelberg: Springer Berlin Heidelberg.

Propper, K., Meindl, K., Sammito, M., Dittrich, B., Sheldrick, G. M., Pohl, E. & Usón, I. (2014). *Acta Crystallogr D Biol Crystallogr* **70**, 1743-1757.

Ravelli, R. B. G. & Garman, E. F. (2006). *Current Opinion in Structural Biology* **16**, 624-629.

Read, R. J. (2001). *Acta Crystallogr D Biol Crystallogr* **57**, 1373-1382.

Read, R. J. & McCoy, A. J. (2016). *Acta Crystallogr D Struct Biol* **72**, 375-387.

Read, R. J. & Schierbeek, A. J. (1988). *Journal of Applied Crystallography* **21**, 490-495.

Robertson, M. P. & Scott, W. G. (2007). *Science* **315**, 1549-1553.

Rodríguez, D. D., Grosse, C., Himmel, S., Gonzalez, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nat Methods* **6**, 651-653.

Rose, J. P., Wang, B. C. & Weiss, M. S. (2015). *IUCrJ* **2**, 431-440.

Rossmann, M. (1972). *The Molecular Replacement Method*. Gordon & Breach.

Rossmann, M. G. (2001). *Acta Crystallogr D Biol Crystallogr* **57**, 1360-1366.

Rossmann, M. G. & Arnold, E. (2001).

Rossmann, M. G. & Arnold, E. (2006). *International Tables for Crystallography*, edited by M. G. Rossmann & E. Arnold, pp. 279-292. Dordrecht: Springer Netherlands.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Crystallographica* **15**, 24-31.

Rost, B. (1997). *Fold Des* **2**, S19-24.

Rost, B. (1999). *Protein Eng* **12**, 85-94.

Rosvall, M., Axelsson, D. & Bergstrom, C. T. (2010). *The European Physical Journal Special Topics* **178**, 13-23.

Sammito, M. (2015).

Sammito, M., Meindl, K., de Ilarduya, I. M., Millán, C., Artola-Recolons, C., Hermoso, J. A. & Usón, I. (2014). *FEBS J* **281**, 4029-4045.

Sammito, M., Millán, C., Rodríguez, D. D., de Ilarduya, I. M., Meindl, K., De Marino, I., Petrillo, G., Buey, R. M., de Pereda, J. M., Zeth, K., Sheldrick, G. M. & Usón, I. (2013). *Nat Meth* **10**, 1099-1101.

Sander, C. & Schneider, R. (1991). *Proteins* **9**, 56-68.

Sayre, D. (1952). *Acta Crystallographica* **5**, 60-65.

Scapin, G. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2266-2275.

Schallmey, M., Koopmeiners, J., Wells, E., Wardenga, R. & Schallmey, A. (2014). *Appl Environ Microbiol* **80**, 7303-7315.

Schoch, G. A., Sammito, M., Millán, C., Usón, I. & Rudolph, M. G. (2015). *IUCrJ* **2**, 177-187.

Schrodinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

Schwarzenbacher, R., Godzik, A., Grzechnik, S. K. & Jaroszewski, L. (2004). *Acta Crystallogr D Biol Crystallogr* **60**, 1229-1236.

Scott, W. G. (2012). *Acta Crystallogr D Biol Crystallogr* **68**, 441-445.

Seymour, G. J., Ford, P. J., Cullinan, M. P., Leishman, S. & Yamazaki, K. (2007). *Clin Microbiol Infect* **13 Suppl 4**, 3-10.

Shannon, C. E. (1949). *Proceedings of the IRE* **37**, 10-21.

Sheldrick, G. (1990). *Acta Crystallographica Section A* **46**, 467-473.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644-650.

Sheldrick, G. M. (2008). *Acta Crystallogr A* **64**, 112-122.

Sheldrick, G. M. (2010). *Acta Crystallogr D Biol Crystallogr* **66**, 479-485.

Sheldrick, G. M. (2015). *Acta Crystallogr A Found Adv* **71**, 3-8.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Crystallographica Section B* **51**, 423-431.

Shrestha, R. & Zhang, K. Y. (2015). *Acta Crystallogr D Biol Crystallogr* **71**, 304-312.

Sitbon, E. & Pietrokovski, S. (2007). *BMC Struct Biol* **7**, 3.

Smith, T. F. & Waterman, M. S. (1981). *J Mol Biol* **147**, 195-197.

Soding, J. (2005). *Bioinformatics* **21**, 951-960.

Soding, J., Biegert, A. & Lupas, A. N. (2005). *Nucleic Acids Res* **33**, W244-248.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Crystallogr D Biol Crystallogr* **60**, 432-438.

Strokopytov, B. V., Fedorov, A., Mahoney, N. M., Kessels, M., Drubin, D. G. & Almo, S. C. (2005). *Acta Crystallogr D Biol Crystallogr* **61**, 285-293.

Suhre, K. & Sanejouand, Y. H. (2004). *Acta Crystallogr D Biol Crystallogr* **60**, 796-799.

Tannenbaum, T., Wright, D., Miller, K. & Livny, M. (2001). *Beowulf Cluster Computing with Linux*, edited by T. Sterling, pp. 307-350: MIT Press.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L. W. (2009). *Acta Crystallogr D Biol Crystallogr* **65**, 582-601.

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. (2008). *Acta Crystallogr D Biol Crystallogr* **64**, 61-69.

Thompson, M. C., Cascio, D. & Yeates, T. O. (2018). *Acta Crystallogr D Struct Biol* **74**, 411-421.

Thorn, A. & Sheldrick, G. M. (2013). *Acta Crystallogr D Biol Crystallogr* **69**, 2251-2256.

Umer, H. & Muhammad Sabieh, A. (2010). *European Journal of Physics* **31**, 453.

Usón, I. & Sheldrick, G. M. (1999). *Curr Opin Struct Biol* **9**, 643-648.

Usón, I. & Sheldrick, G. M. (2018). *Acta Crystallogr D Struct Biol* **74**, 106-116.

Vagin, A. & Teplyakov, A. (1997). *Journal of Applied Crystallography* **30**, 1022-1025.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *Journal of Applied Crystallography* **28**, 347-351.

Vollmuth, F., Blankenfeldt, W. & Geyer, M. (2009). *The Journal of Biological Chemistry* **284**, 36547-36556.

Wang, B.-C. (1985). *Methods in Enzymology*, pp. 90-112: Academic Press.

Wang, Y., Virtanen, J., Xue, Z., Tesmer, J. J. & Zhang, Y. (2016). *Acta Crystallogr D Struct Biol* **72**, 616-628.

Wang, Y., Virtanen, J., Xue, Z. & Zhang, Y. (2017). *Nucleic Acids Res* **45**, W429-W434.

Watanabe, F., Yu, F., Ohtaki, A., Yamanaka, Y., Noguchi, K., Yohda, M. & Odaka, M. (2015). *Proteins* **83**, 2230-2239.

Watson, J. D. & Crick, F. H. (1953). *Nature* **171**, 737-738.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Crystallogr D Biol Crystallogr* **67**, 235-242.

Xu, D. & Zhang, Y. (2012). *Proteins* **80**, 1715-1735.

Yamashita, K., Hirata, K. & Yamamoto, M. (2018). *Acta Crystallogr D Struct Biol* **74**, 441-449.

Zhanabaev, Z. Z., Akhtanov, S. N., Kozhagulov, E. T. & Karibayev, B. A. (2016). *arXiv [physics.data-an]*.

Zhang, K. Y. J. & Main, P. (1990). *Acta Crystallographica Section A Foundations of Crystallography* **46**, 41-46.

Zhang, Y. (2008). *BMC Bioinformatics* **9**, 40.

Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J., Lupas, A. N. & Alva, V. (2017). *J Mol Biol*.

Zuckerkandl, E. & Pauling, L. (1965). *Journal of Theoretical Biology* **8**, 357-366.

CURRICULUM VITAE

Scientific production

Peer-reviewed articles in scientific journals:

J.M. Dunce, O.M. Dunne, M. Ratcliff, C. Millán, S. Madgwick, I.Usón and O. R. Davies. "Structural basis of meiotic chromosome synapsis through SYCP1 self-assembly". *Nature Structural and Molecular Biology* (accepted, *in press*) https://doi.org/10.1038/s41594-018-0078-9

R. D. Oeffner, P. Afonine, C. Millán, M. Sammito, I. Usón, R. J. Read and A. J. McCoy. (2018) "On the application of the expected LLG to decision making in molecular replacement". *Acta Cryst* D74, 245-255. https://doi.org/10.1107/S2059798318004357

C. Millán, M. Sammito, A. J. McCoy, A. F. Z. Nascimento, G. Petrillo, R. D. Oeffner, T. Dominguez-Gil, J. A. Hermoso, R. J. Read and I. Usón. (2018) "Exploiting distant homologs for phasing through the generation of compact fragments, local fold refinement and partial solution combination". *Acta Cryst.* D74, 290-304. https://doi.org/10.1107/S2059798318001365

A. J. McCoy, R. D. Oeffner, C. Millán, M. Sammito, I. Usón and R. J. Read. (2018) "Gyre and Gimble: a maximum likelihood replacement for PC-refinement" *Acta Cryst.* D74, 279-289. https://doi.org/10.1107/S2059798318001353

I. Caballero, M. Sammito, C. Millán, A. Lebedev, N. Soler and I. Usón. (2018) "ARCIMBOLDO on Coiled Coils" *Acta Cryst.* D74, 194-204. https://doi.org/10.1107/S2059798317017582

C. Millán, M. Sammito, I. Garcia-Ferrer, T. Goulas, G. M. Sheldrick and I. Usón. (2015). "Combining phase information in recprocal space for molecular replacement with partial models". *Acta Cryst.* D71, 1931-1945. https://doi.org/10.1107/S1399004715013127

C. Millán, M. Sammito, and I. Usón. (2015) "Macromolecular *ab initio* phasing enforcing secondary and tertiary structure". *IUCrJ*, 2, 95-105. https://doi.org/10.1107/S2052252514024117

M. Sammito, C. Millán, D. Frieske, E. Rodríguez-Freire, R. J. Borges and I. Usón. (2015). "ARCIMBOLDO_LITE: single workstation implementation and use". *Acta Cryst.* D71, 1921-1930. https://doi.org/10.1107/S1399004715010846

T. Goulas, D. Mizgalska, I. Garcia-Ferrer, T. Kantyka, T. Guevara, B. Szmigielski, A. Sroka, C. Millán, I. Usón, F. Veillard, B. Potempa, P. Mydel, M. Solà, J. Potempa and F.X. Gomis-Rüth. (2015) "Structure and mechanism of a bacterial host-protein citrullinating virulence factor, Porphyromonas gingivalis peptidylarginine deiminase". *Sci. Rep.*, Volume 5. https://doi.org/10.1038/srep11969

G. Schoch, M. Sammito, C. Millán, I. Usón and M. Rudolph (2015) "Structure of a thirteen-fold superhelix (almost) determined from first principles" *IUCrJ* 2, 177-187. http://dx.doi.org/10.1107/S2052252515000238

M. Sammito, K. Meindl, I. M. de Ilarduya, C. Millán, C. Artola-Recolons, J. A. Hermoso, and I. Usón. (2014) "Structure solution with ARCIMBOLDO using fragments derived from distant homology models" *FEBS Journal*, 281, 4029–4045. https://doi.org/10.1111/febs.12897

Z. Fourati, B. Roy, C. Millán, P. D. Coureux, S. Kervestin, H. van Tilbeurgh, F. He, I. Usón, A. Jacobson, M. Graille. (2014) "A Highly Conserved Region Essential for NMD in the Upf2 N-Terminal Domain" *Journal of Molecular Biology*, Volume 426, Issue 22, Pages 3689-3702, https://doi.org/10.1016/j.jmb.2014.09.015

M. Sammito, C. Millán, D.D. Rodríguez, I. M. de Ilarduya, K. Meindl, I. De Marino, G. Petrillo, R. M. Buey, J. M. de Pereda, K. Zeth, G. M. Sheldrick and I. Usón. (2013) "Exploiting tertiary structure through local folds for *ab initio* phasing". *Nature Methods,* Volume 10, Pages 1099–1101, https://doi.org/10.1038/nmeth.2644

Book chapter:

I. Usón, C. Millán, M. Sammito, K. Meindl, I. M. de Ilarduya, I. De Marino, D. D. Rodríguez, "Phasing Through Location of Small Fragments and Density Modification with ARCIMBOLDO", In: Advancing Methods for Biomolecular Crystallography, R. Read, A. G. Urzhumtsev, V. Y. Lunin (Eds.), Springer, Dordrecht, The Netherlands, 2013, pp. 123-132. https://doi.org/10.1007/978-94-007-6232-9_12

Posters:

I. Usón, C. Millán, M. Sammito, K. Meindl, I. M. de Ilarduya, I. De Marino, D. D. Rodríguez, K. Meindl, I. M. de Ilarduya, G.M. Sheldrick, I. Usón. "Clustering for Arcimboldo". ZCAM-Daresbury Collaborative Tutorial

C. Millán, M. Sammito, D.D. Rodríguez, K. Meindl, I. M. de Ilarduya, G.M. Sheldrick, I. Usón. "Clustering fragments for ARCIMBOLDO phasing". International School of Crystallography 45th Course Present and Future Methods for Biomolecular Crystallography

M. Sammito, C. Millán, D.D. Rodríguez, K. Meindl, I. M. de Ilarduya, G.M. Sheldrick, I. Usón. "ARCIMBOLDO structure solution with customized and clustered fragment libraries from Borges". International School of Crystallography 45th Course Present and Future Methods for Biomolecular Crystallography

C. Millán, M.Sammito, K. Meindl, I. Martínez de Ilarduya, I. Usón. "Reciprocal space clustering of BORGES-ARCIMBOLDO partial solutions: Practical cases"

ECM28 (28th European Crystallography Meeting) Awarded with the IUCr Biology Poster Prize.

G. Schoch, M. Sammito, C. Millán, I. Usón and M. Rudolph. “Structure of a thirteen-fold superhelix (almost) determined from first principles” The Biophysical Society Annual Meeting

R. Borges, N. Lemke, M. Sammito, C. Millán, I. Usón, and M. R. M. Fontes. “PLA$_2$s-like membrane perturbation mechanismo: extracting the most of crystallography data”. ECM30 (European Crystallographic Meeting), August 2016, Basel, Switzerland.

C. Millán, M. Sammito, A. Nascimento and I. Usón. "ARCIMBOLDO_SHREDDER's contribution to MR: Phasing with fragments from distant homologs". ECM30 (European Crystallographic Meeting), August 2016, Basel, Switzerland.

M. Sammito, C. Millán, R. Borges, A. Nascimento, G. M. Sheldrick and I. Usón. "Solving protein structures without a model or experimental phases". 66th Annual Meeting of the American Crystallographic Association (ACA), Denver, Colorado, U.S.A., July 2016

N. Soler, C. Millán, M. Sammito, I. Caballero, R. Borges, I. Usón. "Recent advances in ARCIMBOLDO towards low resolution". 50th International School of Crystallography: Integrative Structural Biology. June 2017, Erice, Sicily, Italy.

M.Sammito, C. Millán, A.Medina, I. Caballero, N. Soler and I. Usón. "Using BORGES_MATRIX for X-ray *ab initio* phasing and structure interpretation". Understanding biology trough structure.May 2017, Santa Fe, New Mexico, U.S.A

C.Millán , M.Sammito, A. Medina, I. Usón. “Mapping protein structure to graphs: Application to phasing using community clustering algorithms”. IUCr2017 Computing School, August 2017,Bangalore, India.

Talks:

Claudia Millán as coauthor:

R. Borges, M. Sammito, C. Millán, M. R. M. Fontes and I. Usón. “SEQUENCE SLIDER: a multi sequence evaluator and its application in venomics”. ECM29 (European Crystallographic Meeting), Rovinj, Croatia. Oral communication held on Wednesday 26th of August, 2015.

R. Borges, M. Sammito, C. Millán, J. Juanhuix and I. Usón. “Phasing your XALOC data with ARCIMBOLDO” . VII Congress of the Spanish Synchrotron User Association (AUSE) and the II ALBA User's Meeting. Oral communication held on Thursday 18th of June, 2015.

M. Sammito, C. Millán, R. Borges, G. M. Sheldrick and I. Usón. “BORGES_MATRIX: a tool to generate models for *ab initio* phasing and for structure interpretation”. ECM30 (European Crystallographic Meeting), Basel, Switzerland. Oral communication held on the 30th of August, 2016.

I. Usón, A. McCoy, C. Millán, M. Sammito, A. Nascimento, R. Oeffner, y R. Read. “Molecular Replacement with Small Fragments ARCIMBOLDO and PHASER". Gordon Research Conference on Diffraction Methods in Structural Biology. Bates College, Lewiston (Maine, U.S.A.). July 2016

M. Sammito, C. Millán, A. Medina, I. Caballero, N. Soler and I.Usón. "Using BORGES_MATRIX for X-ray *ab initio* phasing and structure interpretation". Understanding biology through structure. Santa Fe, New Mexico, U.S.A

Presenting author Claudia Millán:

C. Millán, "Clustering ARCIMBOLDO partial solutions in reciprocal space", Structural & Computational Biology Programme Seminar, Barcelona, Spain. May 2012

C. Millán, "BORGES", oral communication and tutorial at the Software Fayre, 28th European Crystallographic Meeting (ECM28). Warwick, UK. August 2013

C. Millán, M. Sammito, R. Borges and I. Usón. “Use of clustering algorithms to combine partial solutions in reciprocal space” . ECM29 (European Crystallographic Meeting), Rovinj, Croatia. Oral communication held on Thursday 25th of August, 2015.

C. Millán, "ARCIMBOLDO", oral communication and tutorial at the Software Fayre, ECM30 (European Crystallographic Meeting), Basel, Switzerland. August 2016

C. Millán, "ARCIMBOLDO: phasing with fragments combining MR (PHASER) and density modification (SHELXE)". Oral communication held at the 20th August 2017 at the 'Phasing and Model Building' 24th Congress and General Assembly of the International Union of Crystallography, Hyderabad.

C. Millán, M. Sammito, A. F. J Nascimento, I. Caballero, N. Soler, R. Borges, G. Petrillo, and I. Usón. "New in the ARCIMBOLDO toolbox for phasing with small fragments". Oral communication on the 26th of August, 24th Congress and General Assembly of the International Union of Crystallography, Hyderabad, August 2017.

C. Millán, "ARCIMBOLDO", oral communication and tutorial at the Software Fayre, 24th Congress and General Assembly of the International Union of Crystallography, Hyderabad, August 2017

C. Millán. "ARCIMBOLDO: phasing with fragments combining MR (PHASER) and density modification (SHELXE)" and “A hitchhikers guide to structure solution using ARCIMBOLDO”. One-day full workshop about the ARCIMBOLDO software and its use for structure solution. Institute of Molecular Biosciences, Graz, May 2018.

Oral presentation accepted for ECM31 in Oviedo, Spain, August 2018, entitled “ARCIMBOLDO_SHREDDER: making the most of good data despite having only a poor homolog”.

Invited speaker for CCP4 Study Weekend on Molecular Replacements, which will take place in January 2019

Schools and meetings

- Organization of a one-day Arcimboldo workshop at the University of Graz (Graz, Austria, 2018)
- Participation in the ccpem 2018 Spring Symposium IV, (Keele, U. Kingdom, 2018)
- Participation in the Conference on methods and applications in the frontier between MX and CryoEM, (Barcelona, Spain, 2017)
- Attendance and grant awarded (IUCr Young Scientist Award) to participate in the 24th Congress and General Assembly of the International Union of Crystallography (Hydebarad, India, 2017)
    - Oral communication held on August 25th, entitled “New in the ARCIMBOLDO toolbox for phasing with small fragments”
    - Tutorial about the use of ARCIMBOLDO in the Software Fayre, held on August 27th
    - Oral communication in the 'Phasing and Model Building' workshop, a satellite meeting held on the 20th and 21st of August.
    - Participation in the Computational Crystallography School (Bangalore, August 2017)
- Attendance and grant awarded for participation in PyData Barcelona 2017 (Barcelona, 2017)
    - Oral communication held on May 21st, entitled “Using network community clustering algorithms to aid determination of protein structures”
- Attendance and grant awarded at the 30th European Crystallographic Meeting (Basel, Switzerland, August 2016)
    - Tutorial about the use of ARCIMBOLDO in the Software Fayre, held on 1st September
- Attendance and grant awarded to participate in the Methods in Crystallographic Computing school in Loßburg-Wittendorf, Black Forest, Germany, August 2016)
    - Oral communication held on August 26th , entitled “Multiprocessing versus threading in a single workstation in Python 2.x”
- Attendance and grant awarded to participate in the CCP4 Study Weekend 2016: Protein-Ligand Complexes: Understanding Biological Chemistry (Nottingham, 2016)
- Attendance and grant awarded to participate at the 29th European Crystallographic Meeting (Rovinj, Croatia, 2015)
    - Oral communication entitled “Use of clustering algorithms to combine

partial solutions in reciprocal space" held on Thursday 25th of August

- Attendance to IRB Barcelona BioMed Conference on "Transporters and other Molecular Machines" (Barcelona, 2014)
- Co-organization of BAC2014 (Biotech annual Congress 2014), (Barcelona, 2014)
- Attendance to Campus Gutemberg, (Barcelona, 2014)
- Attendance and grant awarded at Introduction to Software Development for Crystallographers (Warwick, 2013)
- Attendance and grant awarded at ECM28 (28th European Crystallography Meeting) (Warwick, 2013)
- Attendance and grant awarded at The CCP4 Study Weekend 2013: Molecular Replacements (Nottingham, 2013)
- Attendance and grant awarded to participate 45th Course of the International School of Crystallography "Present and Future Methods for Biomolecular Crystallography" (Erice, Sicily, 2012)
- Participation in the Macromolecular Crystallography School (Madrid, 2012)
- Participation in the ZCAM-Daresbury Collaborative Tutorial (Zaragoza, 2012)
- Contribution to organize the 43rd European Brain and Behavior Society Meeting (Seville, 2011)
- Organization of a bioinformatics workshop at the Science Week at the UPO (Seville, 2010)
- Organization of the Science Week at the UPO (Seville, 2009)
- Participation in Symbiosis, the 14th European Biotechnology Congress (Barcelona, 2009)
- Co-organization of a careers advice event : What now? (Seville, 2009)
- Instructor at the workshop Science at the UPO (Seville, 2008)
- Co-organization of ESPOU'08 (Experimental Science in Pablo de Olavide University) (Seville, 2008)
- Participation in the 3rd Biotechnology at the University Congress (León, 2008)
- Co-organization of the 2nd Biotechnology at the University Congress (Seville, 2007)
- Participation in the the summer course "Biotechnology: molecular markers, RT-PCR and DNA pooling. Structural Genomics" (Córdoba, 2007)

Popular science

Talks:

- "A través del espejo y lo que Claudia encontró allí" at the event "Cientifiques a prop III" organised by the AMIT (Women in Research and Technology association) on the 8 November 2017. https://youtu.be/exEyS6gbiHQ
- "From 3D Structure to Function" at the BCN Science Slam on the 27th of October of 2017 https://youtu.be/WPGwEqPf78s
- "The eyes of chemistry", 2nd Ellerslie talk at 9 Adams Road, Cambridge, on Saturday 29th July 2017. Article about the talk in http://biofisica.info/articles/the-eyes-of-chemistry/

Articles in popular science magazines:

- MoleQla ISSN 21730903
    - “El reto de la Castafiore: ¿Puede un cristal antibalas ser destrozado por la voz de una soprano?”. *MoleQla* nº4
    - “¿Y por qué cristalina? Bienvenida”. *MoleQla* nº4
    - “Entrevista a Juan Manuel García Ruiz”. *MoleQla* nº6
    - “Computación ciudadana”. *MoleQla* nº7
    - “Bioinformática para la cristalografía”. *MoleQla* Nº8
    - “Descubriendo como sienten las células”. *MoleQla* Nº11
    - “2014, Año Internacional de la Cristalografía”. *MoleQla* Nº13
    - “¿Estás tú tan en forma como la cristalografía en su año internacional?”. *MoleQla* Nº16
    - “Ada Yonath y la estructura de la máquina de máquinas celulares: El ribosoma”. *MoleQla* Nº17
- Arbor ISSN 0210-1963
    - Millán, C. and Usón, I. (2015). “Crystallographic Structure Solution”. *Arbor*, 191 (772): a218. doi: http://dx.doi.org/10.3989/arbor.2015.772n2004
- Biofisica ISSN 2445-4311
    - Millán, C. and Usón, I. (2018). “The eyes of chemistry”. Biofisica. http://biofisica.info/articles/the-eyes-of-chemistry/