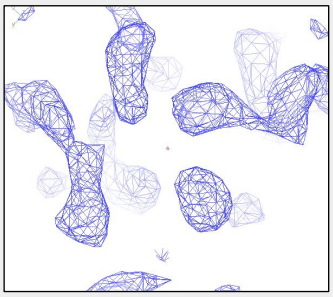
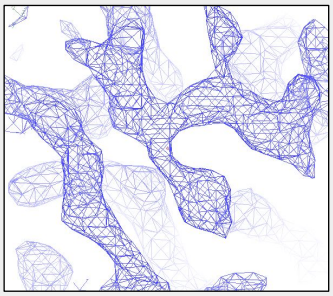
Density Modification

Paul Bond
University of York

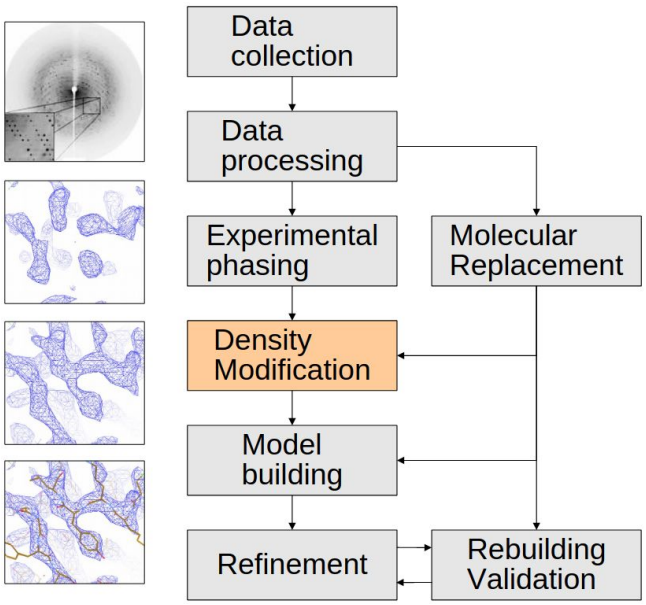
paul.bond@york.ac.uk





This presentation is on density modification, but it will also cover some background information on phase probability distributions and non-crystallographic symmetry (NCS).

X-ray Crystallography



This is a rough overview of the structure solution process for X-ray crystallography. After the experimental data has been collected and processed, some phase estimates are required to produce an initial density map. The phases can come from experimental phasing, molecular replacement or a combination of both. Initial maps from experimental phasing (like the one shown on the left) are often poor due to uncertainty in the phases. In this case density modification is used to improve the phases to give a better map for model building. Density modification can often be skipped after molecular replacement if the model is very complete and accurate. However, it can still help to improve the density when the model is incomplete or contains errors.

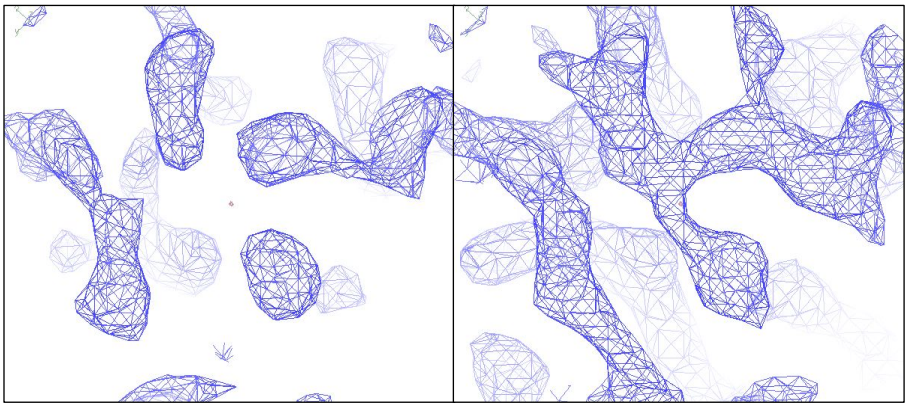
Density Modification

Classical

- DM
- Solomon
- CNS
- ACORN
- Parrot

Statistical

- RESOLVE



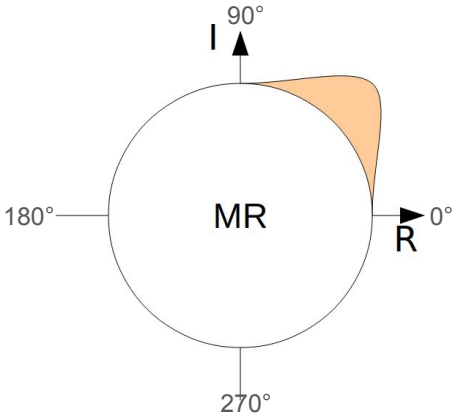
Density modification programs generally fall into one of two classes depending on how they work. The methods will be explained in more detail later in the presentation, but the brief explanation is that classical density modification programs modify the density map to update the phases and statistical density modification programs create a map of density probability distributions to update the phases. ACORN and Parrot are the programs that are available in CCP4i2 and CCP4 Cloud. RESOLVE is the most widely used statistical density modification program and is available in Phenix.

Phase Probability Distributions

Start with structure factor amplitudes and phase estimates for each reflection

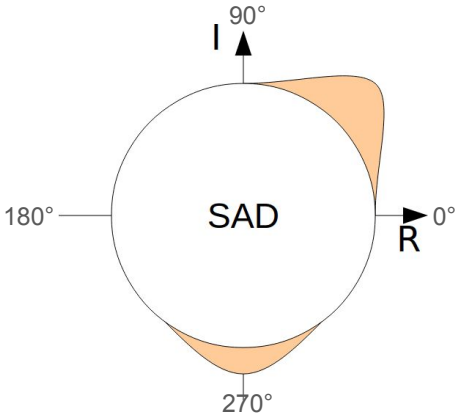
Unimodal distribution

Phase, figure of merit - Φ, FOM



Bimodal distribution

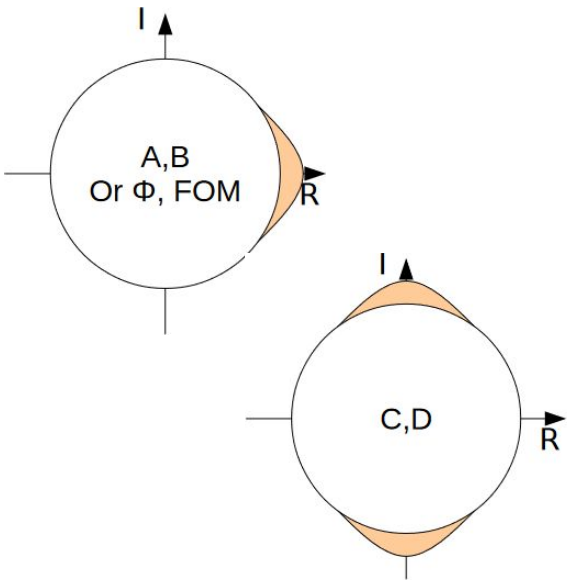
Hendrickson-Lattman coeffs - ABCD



Firstly, it is important to understand phase probability distributions. Each reflection has an amplitude from data processing and a phase estimate from either molecular replacement or experimental phasing. Here the phase estimates are shown as probability distributions on Argand diagrams where the phase goes from 0 to 360 degrees around the circle. Each technique gives a different type of phase probability distribution. For molecular replacement, we know the model will contain errors so we can't be certain about the phase. It gives a unimodal distribution with the most likely phase (Φ) and a figure of merit (FOM) between 0 and 1 describing our confidence in that phase. The more errors we think the MR model contains the lower the FOM and the wider this distribution will be. Experimental SAD phasing gives a bimodal phase distribution because the two phasing circles overlap in two places, producing two possible phases with their respective uncertainties. Bimodal phase distributions are described by Hendrickson-Lattman coefficients, which are a set of four numbers: A, B, C and D.

Hendrickson-Lattman Coefficients

- A,B represent a unimodal distribution
- C,D represent a bimodal distribution
- Relative size and sign of A,B or C,D control the direction
- Absolute size $(A^2+B^2)^{½}$ controls the sharpness
- For MR, we get A,B (or Φ, FOM) i.e. C=D=0
- Together, A,B,C,D can describe a bimodal distribution with any combination of peak height and direction

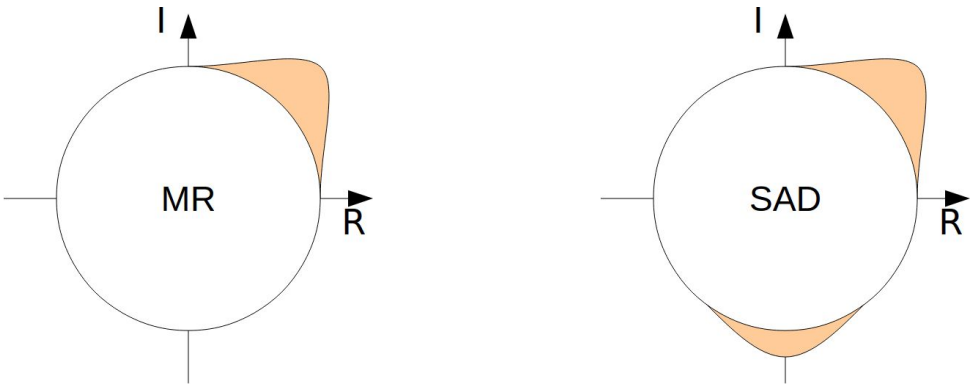


A and B represent a unimodal distribution the same as Φ and FOM do. It is possible to convert between A,B and Φ,FOM without any loss of information. C and D by themselves represent a symmetrical bimodal distribution with the two modes 180 degrees apart from each other and peaks the same height. When you combine A,B and C,D they can represent any bimodal distribution with any combination of different peak heights and directions. They are a more generic representation as they can also describe unimodal distributions (e.g. for MR) when both C and D are equal to 0.

Calculating a Map

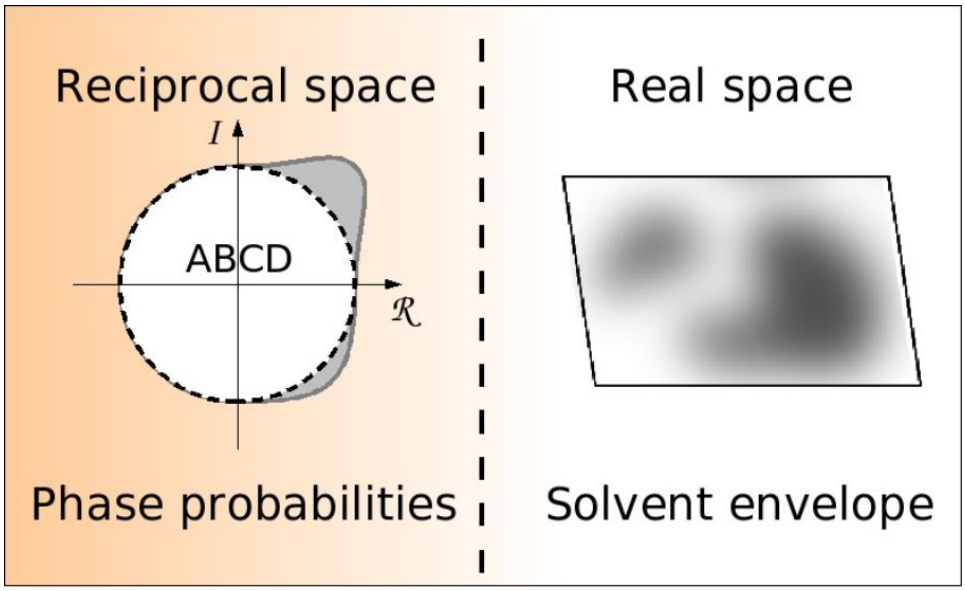
Need to reduce phase probability distributions to single phases

Phase and amplitude at the centre of mass produce the least noisy map



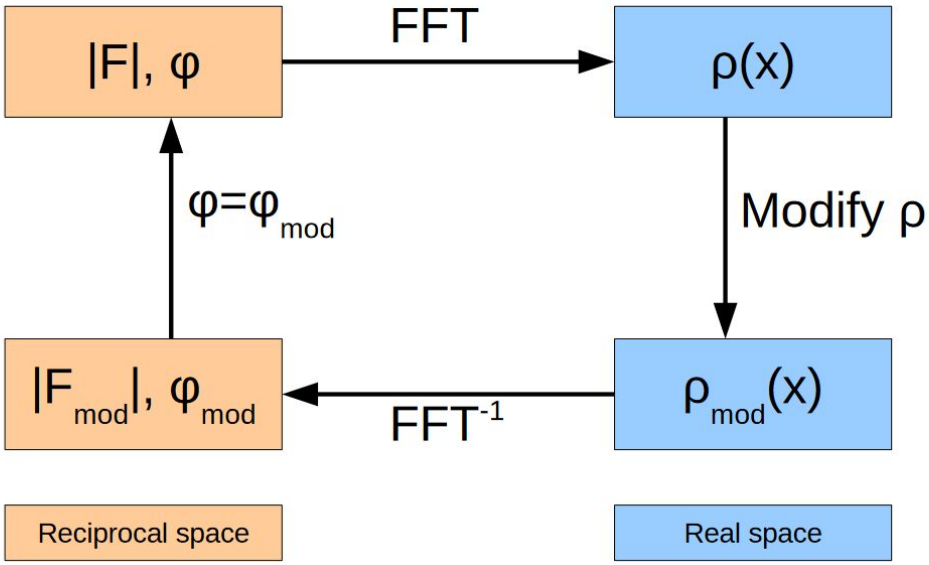
To calculate an electron density map we need an amplitude and a single phase. The map that is least noisy with respect to the uncertainty in the phase can be calculated by using the amplitude and phase at the center of mass of the phase distribution. In these Argand diagram representations the observed amplitude is the distance from the origin given by the radius of the circle. For a unimodal distribution the map phase is the same as the center of the distribution, but the amplitude gets reduced depending on the figure of merit. As the uncertainty increases and the distribution spreads out, the center of mass moves towards the center, and that structure factor becomes less important in calculating the map. For a bimodal distribution the map phase is somewhere between the two peaks and the center of mass is usually much closer to the center.

Reciprocal and Real Space



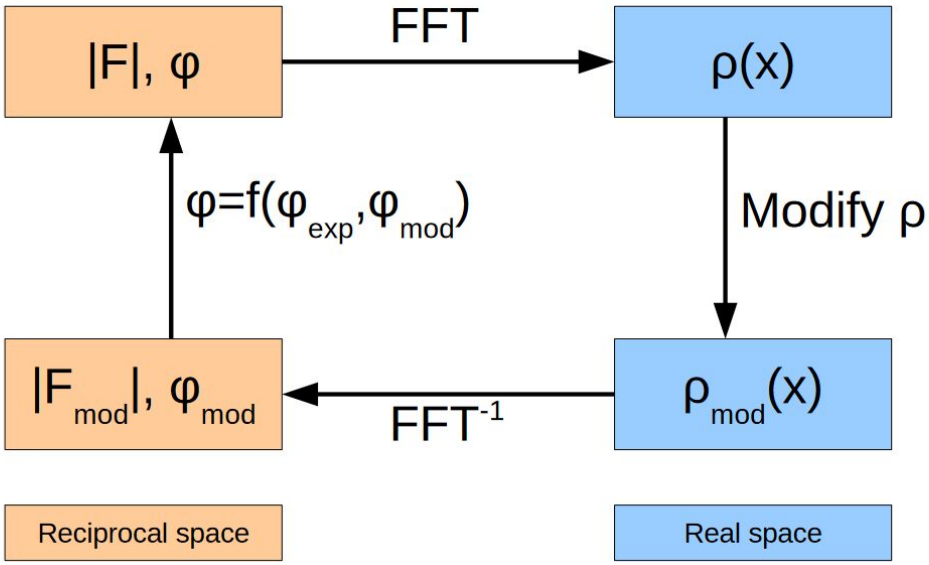
The initial electron density map will be noisy, especially in the SAD case where there is an ambiguity between two possible phases, but density modification is able to improve the map by combining information in real and reciprocal space. We start from the phase probability distributions and use them to create an electron density map in real space, then we add some information in real space in order to narrow down the possible phases.

Rudimentary Classical Method

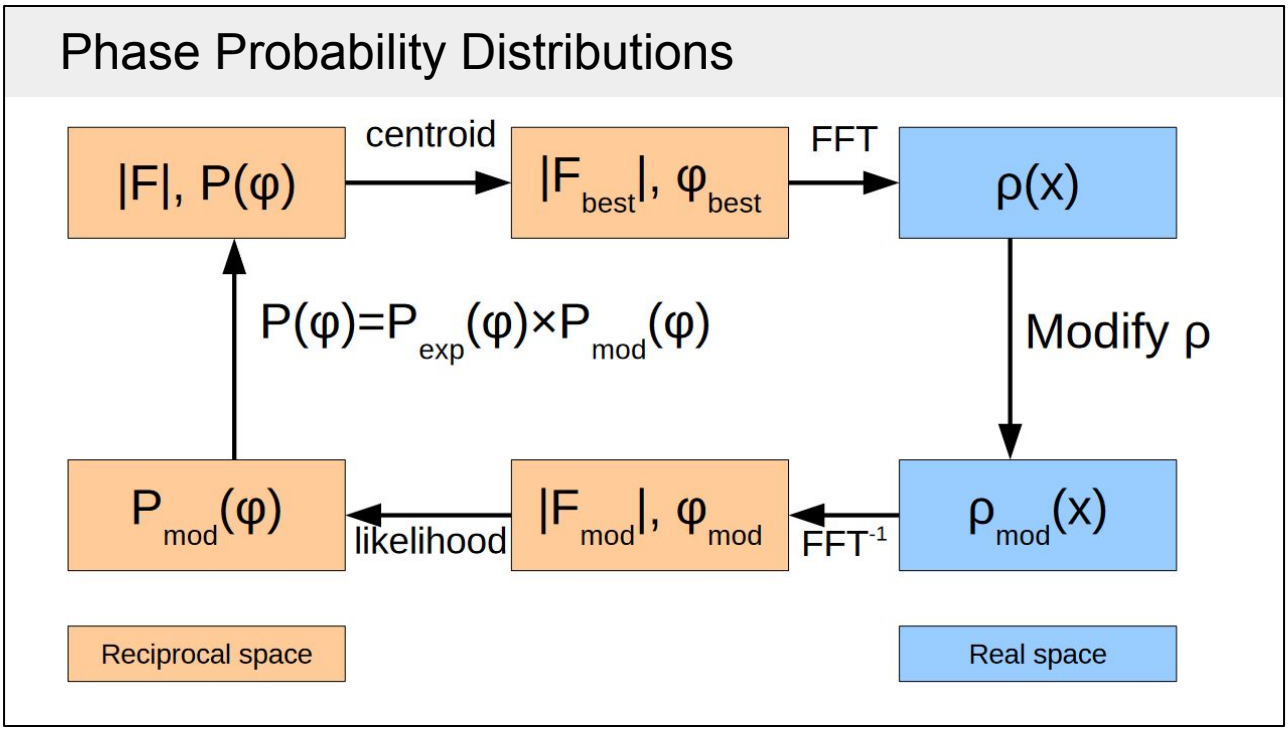


This is a classical density modification calculation in a very rudimentary form. It starts with a set of structure factor amplitudes (|F|) and phases (φ). It then calculates an electron density (ρ) map using a fast Fourier transform (FFT) and modifies the density in real space. For example, basic image processing algorithms could be used to reduce the noise in the map. An inverse Fourier transform is then used to get amplitudes and phases for the modified map, where both the amplitudes and phases will have changed. This information could be used by just replacing the original phases with the modified ones. By combining the updated phases with the original (experimentally determined) amplitudes the calculation can be repeated cyclically to improve the map.

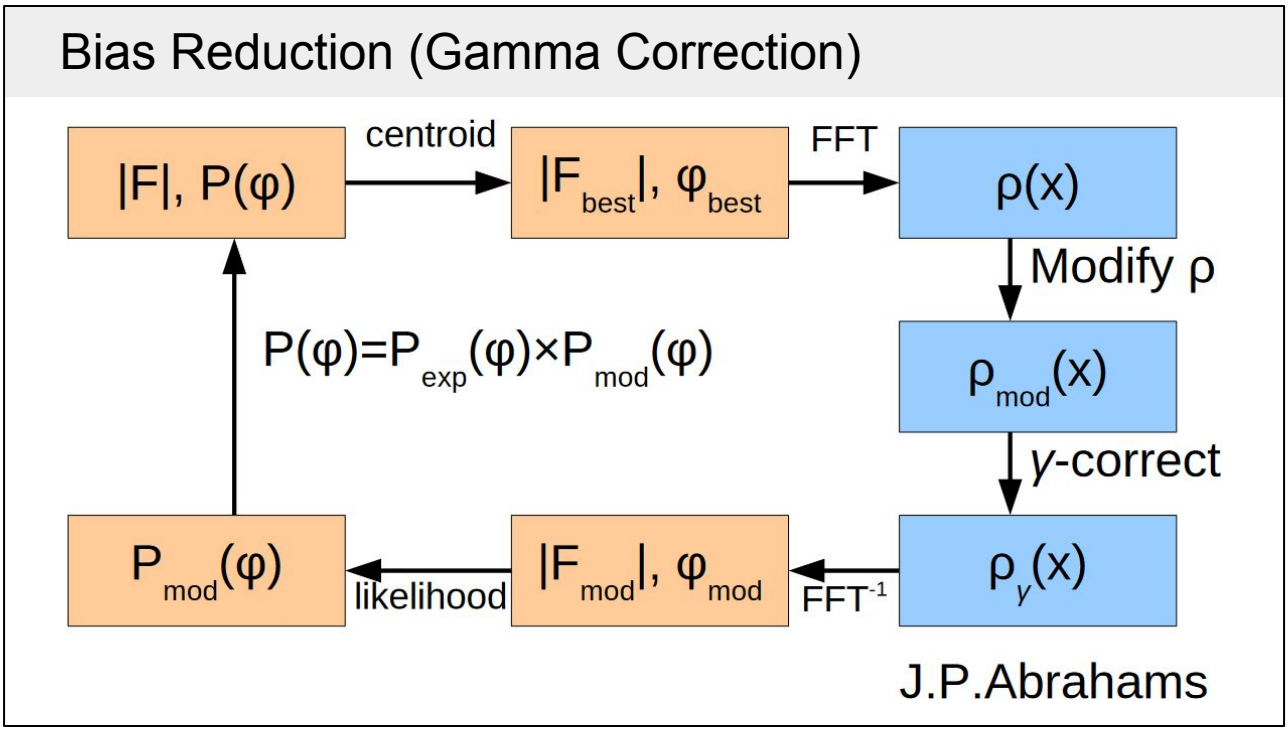
Phase Weighting



In practice, instead of the original phases being simply replaced, there is some combination of the original phases and the modified phases.

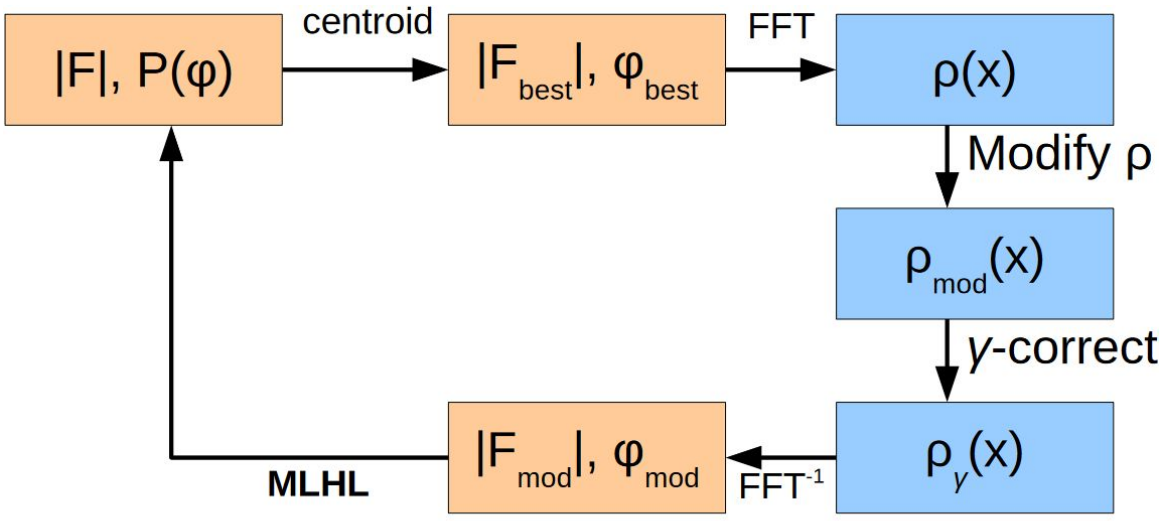


As discussed earlier, we are not dealing with single phases but with phase probability distributions. Starting from the experimental amplitudes and phase distributions in the top left, we find the center of mass to get best phases and weighted amplitudes that depend on the confidence in the best phases. As before, a Fourier transform is used to get a map, the map is modified, then an inverse Fourier transform produces modified amplitudes and modified single phases. These single phases need to be changed back into phase probability distributions, which is done by comparing the modified amplitudes with the experimental amplitudes. If there is a larger difference in amplitude then there is less confidence in the modified phase and the distribution is wider. The original distributions can then be updated through the product of the experimental distributions and the modified distributions. The main flaw in this approach is that you can only multiply probabilities together if they are independent. Unfortunately, the modified distributions are strongly dependent on the experimental distributions because they are derived from them. This leads to bias in the calculation. Repeated cycles increase the figures of merit to 1 to say we are certain about the phases, but the errors in the phases do not decrease to 0.

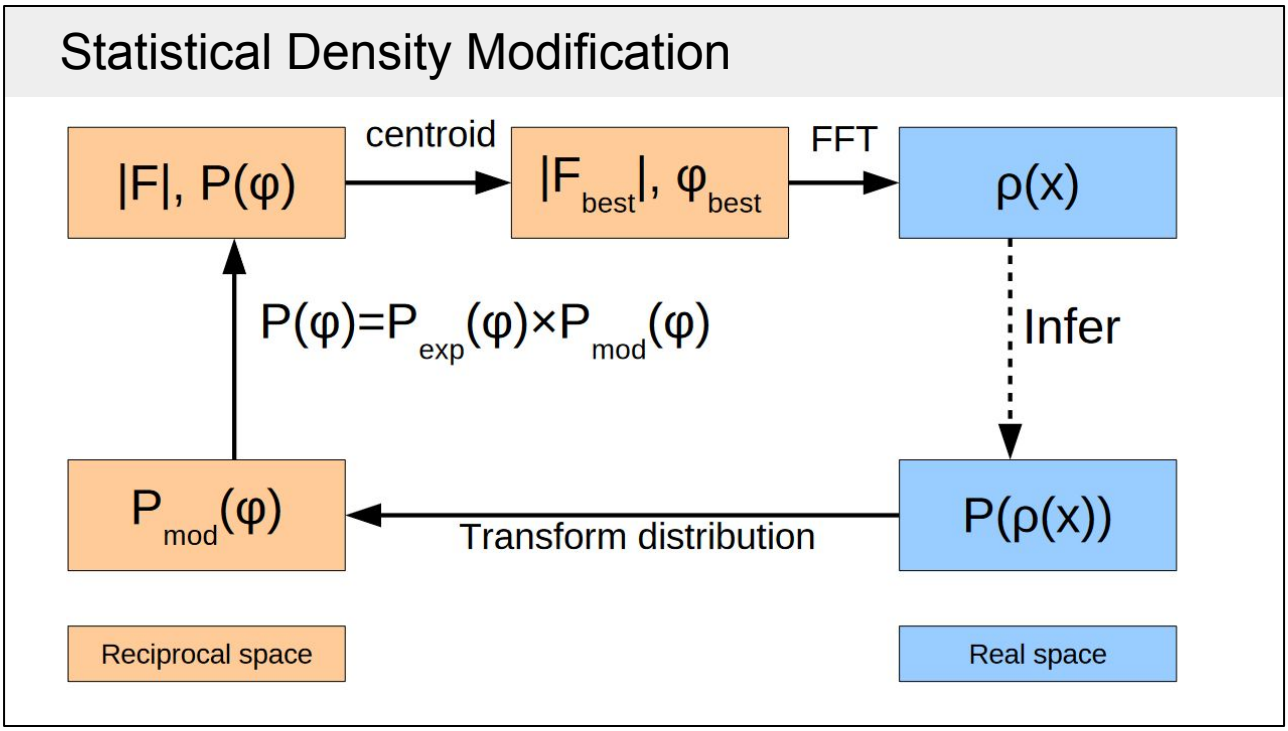


A method called gamma correction was developed by Jan Pieter Abrahams in order to reduce the amount of bias in the calculation. This is an extra step after modifying the density to subtract out part of the initial density map. The amount of the initial map you subtract out has to be right to remove the bias. This is the classical density modification cycle used in the programs DM, SOLOMON and CNS.

Maximum Likelihood H-L

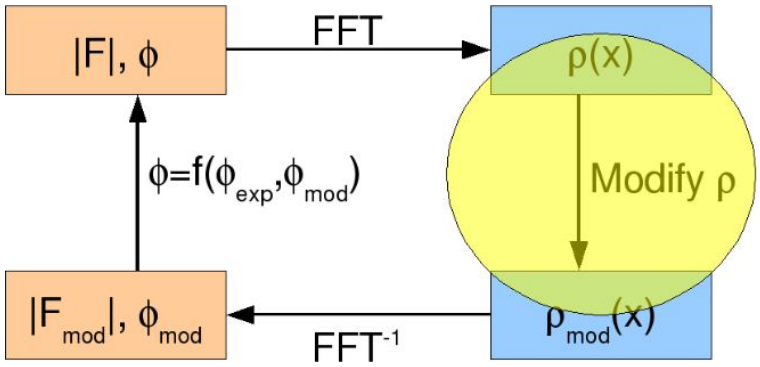


A further development to this classical density modification cycle has been to merge the last two steps into a single likelihood calculation that includes the experimental phase distribution (i.e. Maximum Likelihood using Hendrickson-Lattman coefficients). This improves the method slightly more and is available in the program Parrot.



This is a flow chart for statistical density modification, which works very differently to classical density modification. The initial density map is produced in the same way, but the map is not modified. Instead, a map of density probability distributions is created. An inverse Fourier transform converts these real-space density probability distributions into reciprocal-space phase probability distributions and, as before, those are used to update the original phase probability distributions.

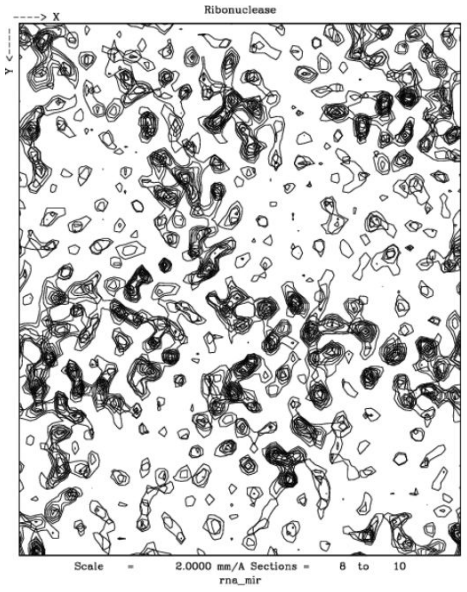
Classical Techniques

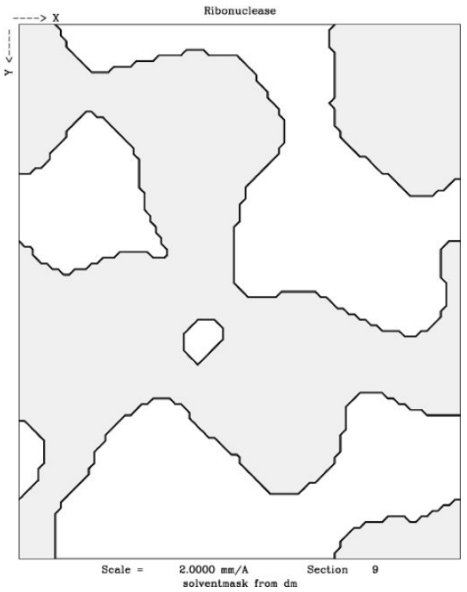


- Solvent flattening
- Histogram matching
- Non-crystallographic symmetry (NCS) averaging

For classical density modification, the most important thing is how the map is modified in real space to improve it. It is a similar problem to standard image processing, e.g. when photo processing software is used to remove noise from an old photograph. In this case it is working in 3D instead of 2D and uses prior knowledge about well-phased electron density maps. There are three very widely used techniques that aim to make the density map more realistic: solvent flattening, histogram matching and NCS averaging.

Solvent Flattening





Solvent flattening is the oldest technique and probably the most powerful one if they were used in isolation. The image on the left shows a few stacked 2D slices through a contoured electron density map. There are regions with lots of features that belong to the protein and sparse regions without many features that belong to the solvent. Using the knowledge that there are fewer features in the solvent regions (because they are disordered) and the fraction of solvent in the crystal, programs are able to draw a mask that divides the map up into protein regions and solvent regions. The solvent mask for the example above is shown in the image on the right. If the map has very good phases then the features in the solvent region will be real ordered features, i.e. mostly ordered waters close to the protein. However, early on in the calculation the phases are poor and there is more noise in in the solvent than real ordered features. Solvent flattening removes all the density in the solvent regions, which will remove both ordered features and noise. If it removes more noise than ordered features then the phases improve overall and the next cycle will produce a less noisy map.

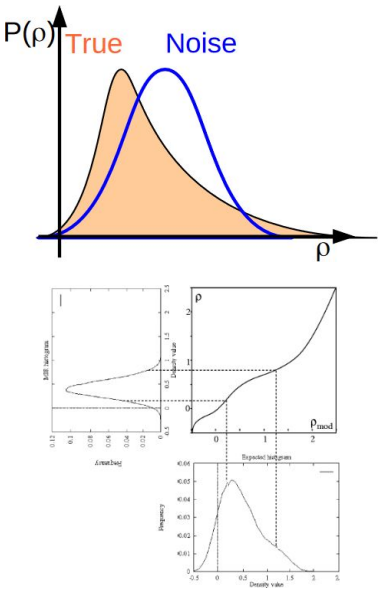
Histogram Matching

A technique from image processing for modifying the protein region.

- Noise maps have Gaussian histogram.
- Well phased maps have a skewed distribution: sharper peaks and bigger gaps.

Sharpen the protein density by a transform which matches the histogram of a well phased map.

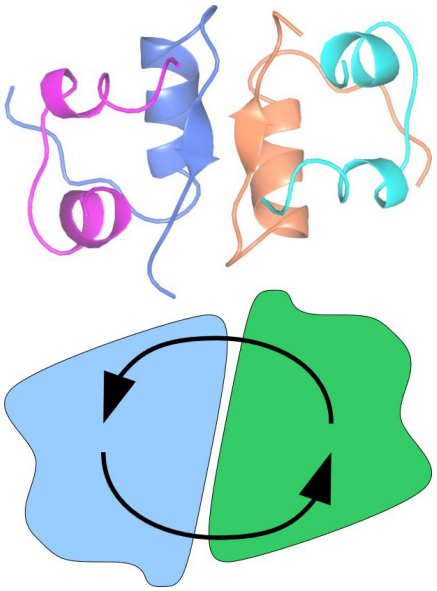
Useful at better than 4Å.



Histogram matching is a technique used to modify the density in the protein region which, even with a badly phased map, should have more real ordered features than noise. It uses the knowledge that pure noise maps have a gaussian distribution of density values but well-phased maps without noise have a more skewed distribution with fewer high density values (corresponding to atom positions) and more low density values (corresponding to the gaps between atoms). The protein density then gets modified using a monotonically increasing curve so that the density distribution looks like that of a well-phased map. This technique is generally useful at resolutions better than 4Å.

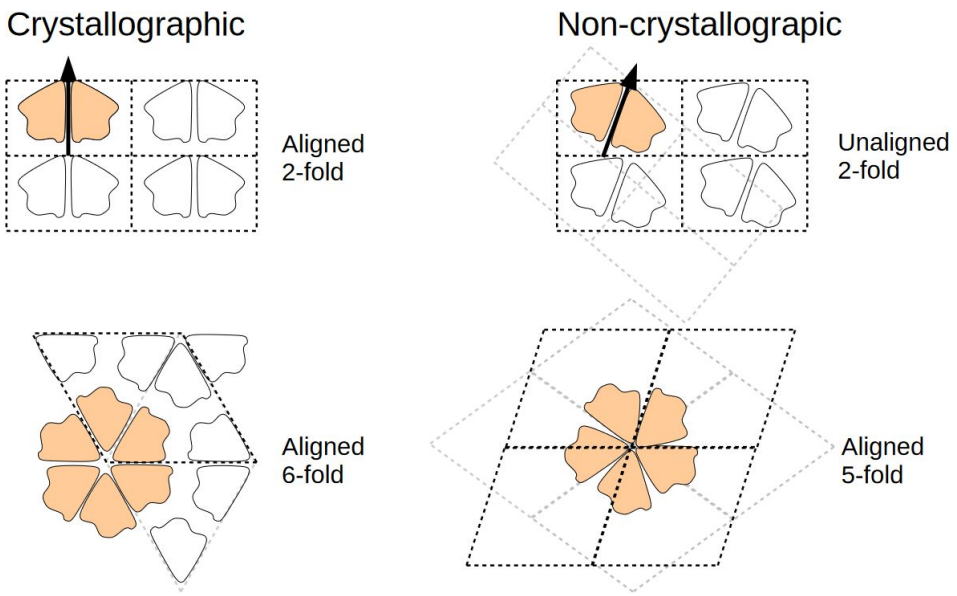
Non-crystallographic Symmetry

- If the molecule has internal symmetry, we can average together related regions.
- In the averaged map, the signal-noise level is improved.
- If a full density modification calculation is performed, powerful phase relationships are formed.
- With 4-fold NCS, can phase from random!



Non-crystallographic symmetry (NCS) averaging is a more complicated technique than solvent flattening or histogram matching. If there are multiple copies of the same molecule within the asymmetric unit then it is possible to average together related regions of the density map. The example on the right shows an insulin dimer. It forms a hexamer in the crystal with three-fold crystallographic symmetry and two molecules in each asymmetric unit related by two-fold non-crystallographic symmetry (the difference between crystallographic and non-crystallographic symmetry is explained on the next slide). By averaging the density of the two molecules together, the real signal will be reinforced and the noise will sometimes reinforce and sometimes cancel out, which improves the signal-to-noise level. When the density modification calculation is performed cyclically, with the modified phases updating the previous ones for the next cycle, then relationships between the phases of different reflections are formed to give a big improvement. With 4-fold NCS, it is even theoretically possible to solve the phase problem starting from random phases, although this does not work in practice because we would not know which parts of the map to average.

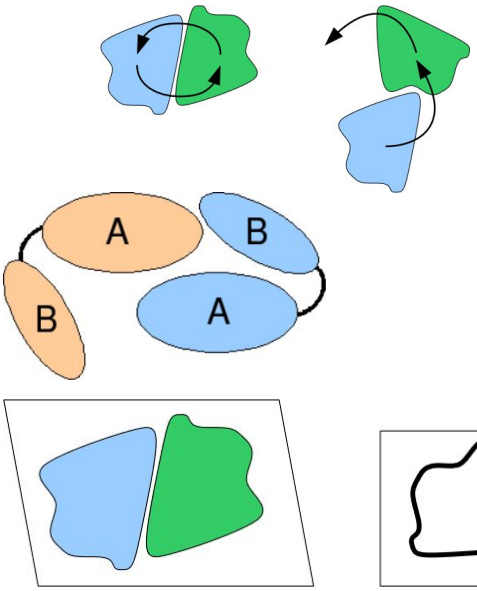
Non-crystallographic Symmetry



The difference between crystallographic and non-crystallographic symmetry is that crystallographic symmetry operations map the whole crystal onto itself and NCS operations only work locally and not for the whole crystal. This slide shows four examples: (top left) a 2-fold rotation axis aligned along a unit cell direction, (bottom left) a 6-fold rotation axis aligned along a cell dimension perpendicular to the image, (top right) a 2-fold axis that is not aligned with a cell axis that maps the two highlighted molecules onto each other but not the molecules in other unit cells, and (bottom right) a 5-fold axis perpendicular to the image that has to be local because a unit-cell lattice cannot contain 5-fold symmetry.

Non-crystallographic Symmetry

- Proper and improper NCS: (closed and open)
- Multi-domain averaging:
- Multi-crystal averaging:



This slides introduces other useful concepts for NCS averaging. NCS operations can be proper (closed) or improper (open). Proper NCS operations map every copy of the molecule onto another copy to form a closed loop. Improper NCS operations only map some molecules onto others. Multi-domain averaging is needed when molecules are flexible as the density cannot be averaged over the whole molecule. In this example the two A domains are related by a proper 2-fold rotation but this operation does not work for the two B domains, for which separate improper rotations are needed. Multi crystal averaging is a technique to average density between molecules in different crystal forms. For example, if there is one crystal with two copies in the asymmetric unit (ASU) and another crystal with one copy in the ASU it is possible to average between all three copies.

Non-crystallographic Symmetry

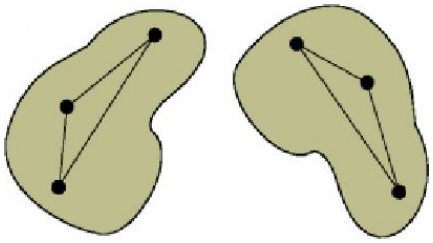
How do you know if you have NCS?

- Cell content analysis – how many monomers in ASU?
- Experimental information
- Self-rotation function
- Difference Pattersons (pseudo-translation only)

How do you determine the NCS?

- From heavy atoms
- From initial model building
- From molecular replacement
- *From density MR (hard)*

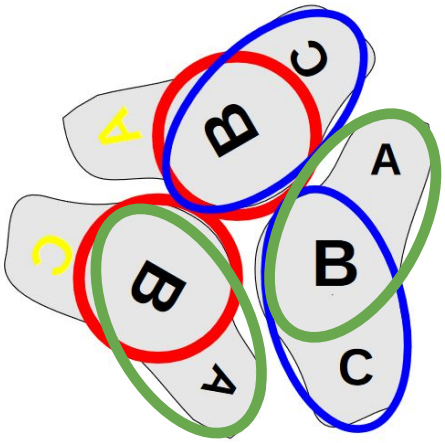
Region determined automatically



Cell content analysis is the primary method of estimating the number of molecules in the ASU. Given a sequence or a molecular weight, it estimates the protein volume and combines that with a solvent content probability, as protein crystals usually have around 50% solvent. There may also be experimental information that will make a number of copies more likely, for example if the protein forms a trimer in solution it is likely to form a trimer in the crystal as well. A self rotation function performs a molecular replacement search of the data against itself, which may show rotations. A difference Patterson can detect pseudo-translations, where the molecule moves without rotating. Although, in practice, these last two techniques are not often used as it is usually possible to proceed with structure solution without being certain of the NCS at first. Determination of the NCS operators depends on the phasing technique used. Experimental phasing produces a heavy atom substructure. If each molecule has 3 or more heavy atoms associated with it, then it should be possible to determine the NCS operators through superposition. Otherwise, if the map is good enough to build part of each molecule (either automatically or interactively) then the partial models could also be superposed. If molecular replacement has placed multiple copies then determining the NCS operators is straightforward. It may also be possible to determine NCS from the density alone. Phenix has a program called *find_ncs_from_density* that works automatically. As a last resort, a more manual approach would be to find a region of strongly varying density that is likely to correspond to a molecule, isolate it using a mask and use that for molecular replacement. Because NCS is local the averaging is only applied to certain parts of the map. Parrot detects this region automatically by looking at where there is agreement in the density.

Pairwise-weighted NCS Averaging

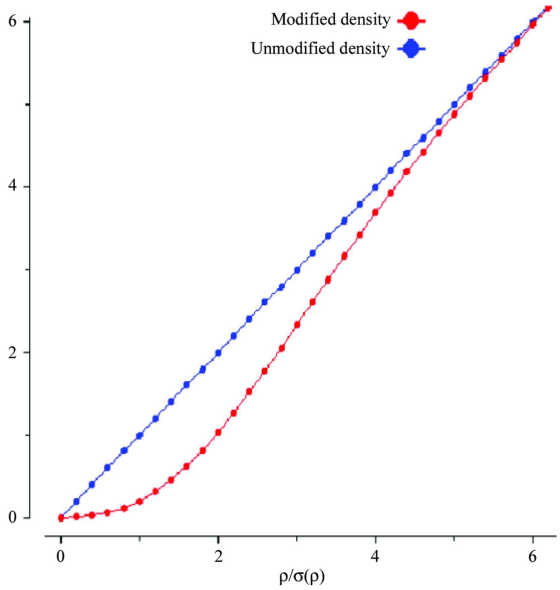
- Average each pair of NCS related molecules separately with its own mask.
- Generalisation and automation of multi-domain averaging.



Another development in Parrot is the pairwise-weighted NCS averaging function, which deals with multi-domain averaging by creating a separate mask (that determines the NCS averaged region) for each pair of molecules. In this example there are three copies of a molecule, each with three domains. The A and C domains in yellow are different from the others, so the density should not be averaged over the whole molecule. Instead, three separate masks (in red, green and blue) are determined automatically by examining the agreement in the density between that pair of molecules.

Acorn

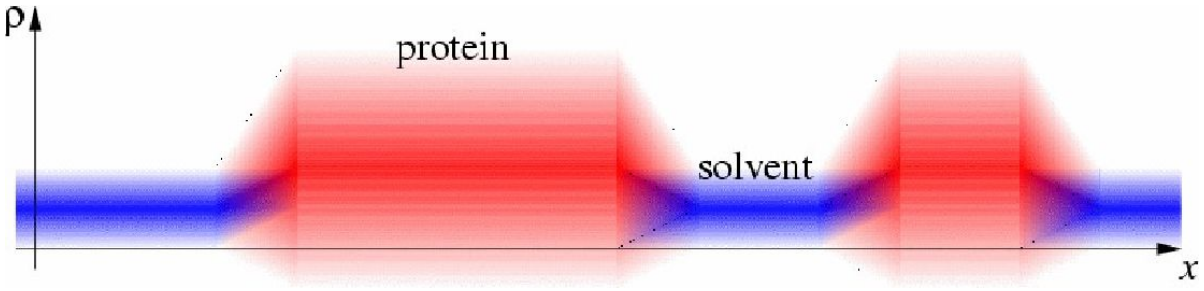
- Refine the map using dynamic density-modification (DDM)
- Previously needed atomic resolution data, but now works to slightly lower resolutions (~1.7Å) by artificially extending the data
- Works even with very poor phases (especially at higher resolution)
- Measures success through the correlation coefficient of weaker reflections (should go up)



ACORN is another program that uses a classical density-modification cycle, but with a technique it calls dynamic density-modification. It works on the assumption that higher density has more chance of being real density from the structure and lower density has more chance of being noise, so it applies the function on the right to reduce the lower density. It works with high resolution data (e.g. better than 1.7Å), although there is a trade-off between resolution and phase quality. With 1Å data it is likely to work even with very poor phases, but with 1.7Å data it may only help with better initial phases. ACORN splits the reflections up into groups and uses the stronger reflections for the calculation and the weaker reflections to assess its progress using a correlation coefficient. The number of cycles is determined automatically.

Statistical Density Modification

Form a statistical description of expected map features.



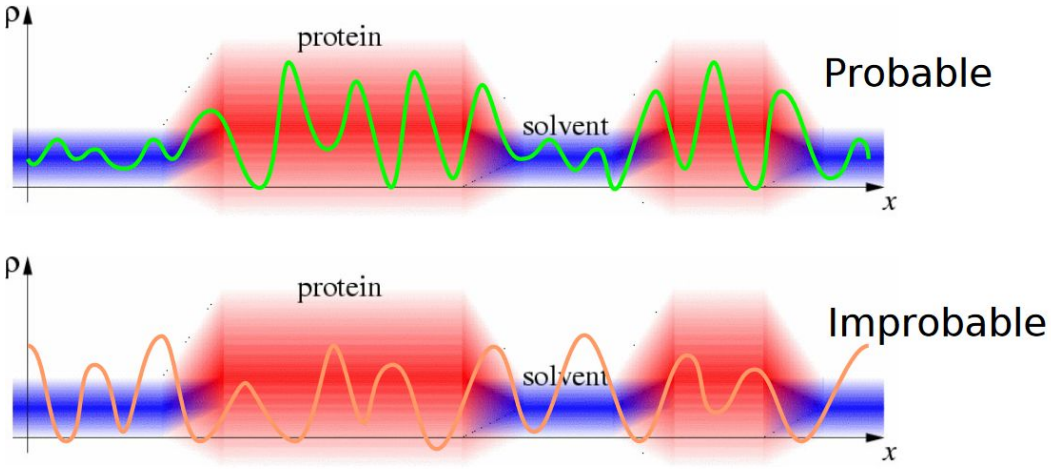
E.g.

- Protein has higher mean, and is more peaky (higher variance)
- Solvent has lower mean, and is flatter (lower variance)

These next few slides give a more detailed description of statistical density modification, which is based around forming a statistical description of what the density is expected to look like. It is usually possible, even with badly phased maps, to identify which are the protein and solvent regions. For well-phased maps, the protein regions have a slightly higher mean density but also a much higher variance because they are ordered. Conversely, the solvent regions have a lower mean and are much flatter.

Statistical Density Modification

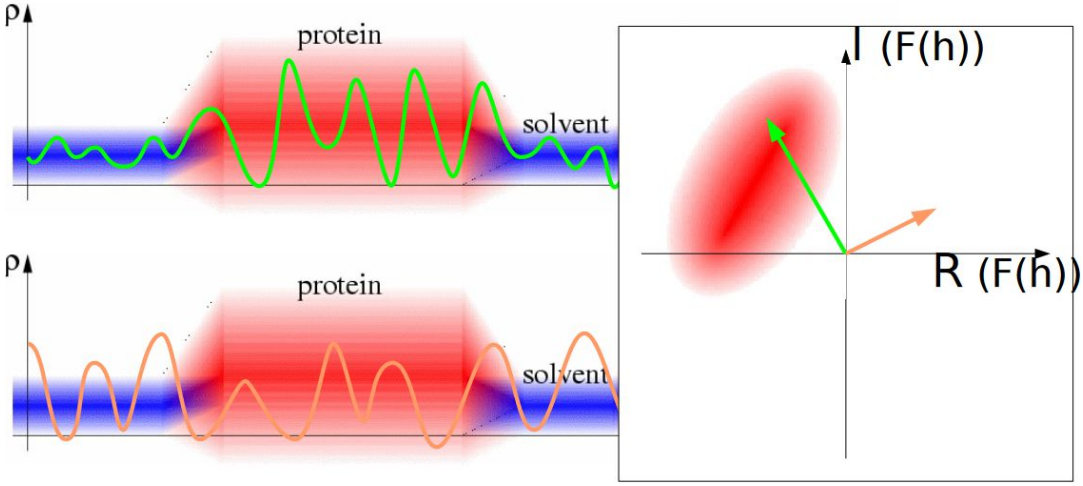
Probability of a map is determined by how well it fits these distributions:



These expected distributions can be used to quantify the probability of a map. In this example there are two maps. The top map has a high mean and variance in the protein and a low mean and variance in the solvent so it is quite probable, i.e. likely to be correct. The bottom map is improbable because it does not fit the expected distributions well.

Statistical Density Modification

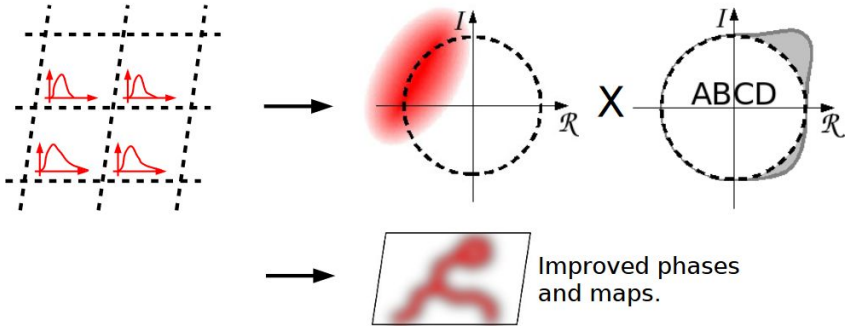
Probability of each structure factor is given by the probability of the corresponding map.



The maps correspond to sets of structure factors that have different values in the probable and improbable maps. Because the top map is probable, the structure factors are probable. Because the bottom map is improbable then the structure factors are too. Real-space density probability distributions can be converted to reciprocal-space probability distributions for individual structure factors in this way.

Statistical Density Modification

- Obtain per-grid density probability distributions.
- Transform to reciprocal space.
- Combine with experimental phases.
- Map probability becomes phase probability distribution.



Bricogne (1992) Proc. CCP4 Study Weekend
Bricogne (1997) Methods in Enzymology

The overall process starts by assigning density probability distributions to each grid point in real-space. Those are transformed into reciprocal space distributions for individual structure factors, which are then combined with the experimental phase distribution to produce an updated phase distribution for the next cycle. There is a complication that the probability distribution for one structure factor depends on all the other structure factors. In practice this is handled by looping through each structure factor and changing one at a time. This is the method used in the program RESOLVE and the theory comes from Gerard Bricogne.

Statistical Density Modification

Advantages

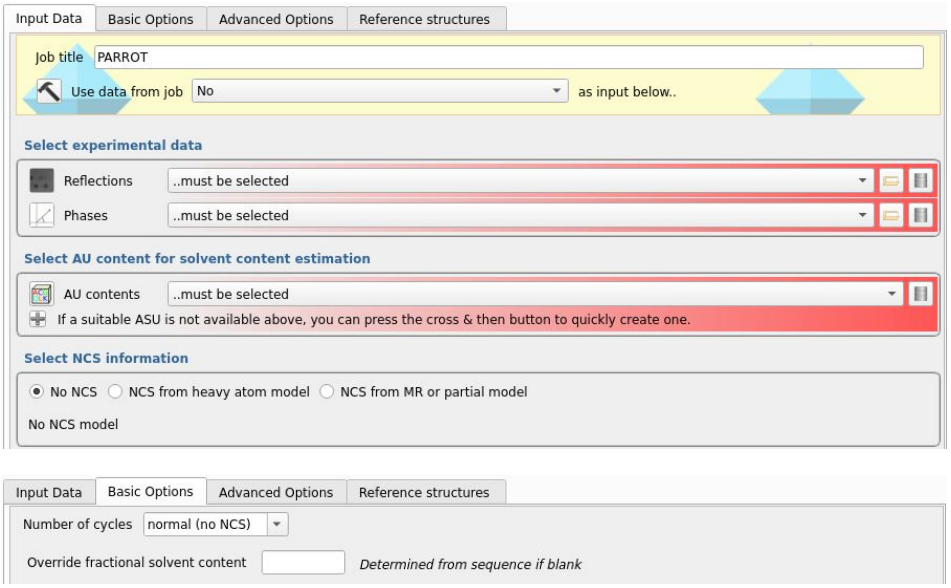
- Reduced bias
- Better phases

Disadvantages

- Slower
- Latest classical methods comparable

Historically at least, statistical density modification has less of a problem with bias and can produce better phases than classical density modification, at the cost of taking longer to run. However, the latest classical methods are comparable in terms of bias reduction and performance. There are cases where classical methods will work better and others where statistical methods will work better so it is worth trying both in difficult cases. In easy cases both methods should give acceptable results.

CCP4i2 - Parrot



This is the input page for Parrot in CCP4i2. It requires a set of reflections, starting phases that you want to improve and the asymmetric unit (AU) contents in order to estimate the solvent content. To determine the operators for NCS averaging, it must also be provided with either a heavy atom substructure or an atomic model (e.g. from MR or initial model building). The basic options tab allows you to specify the the number of cycles as either normal (without NCS), normal (with NCS) or many (with NCS), which correspond to 3, 10 or 100 respectively. Using NCS averaging means that more cycles can be performed without the resulting phases being too biased. The number of cycles option only changes the number of cycles and does not affect whether NCS averaging is performed. The tab also has the option of overriding the solvent content estimation. Note that the CCP4i2 Parrot task does not use the number of copies specified in the AU contents and will estimate the number of copies itself. If the number of copies is known and Parrot estimates it incorrectly then it is important to use this option as it affects the determination of the solvent mask. There are more advanced options in the other tabs that do not normally need to be changed.

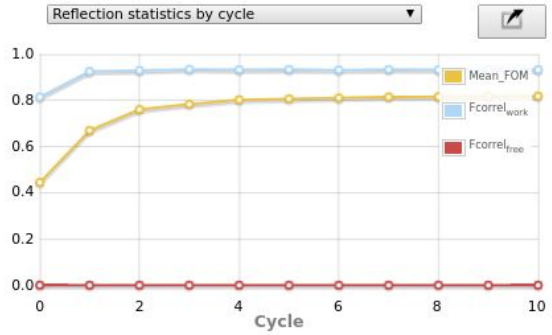
CCP4i2 - Parrot

Results

The solvent content is set to 50.3% corresponding to 2.0 copies of the sequence in the asymmetric unit.

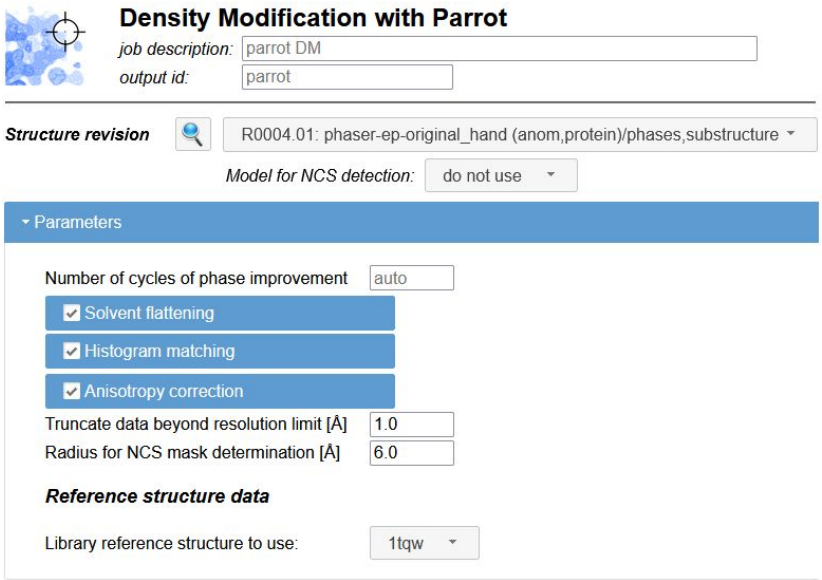
The final figure-of-merit is **0.82**, which suggests that the map is good enough for model building. However the figure-of-merit from density modification can be seriously overestimated.

Solvent content	0.5034
Initial FOM	0.445
Final FOM	0.819



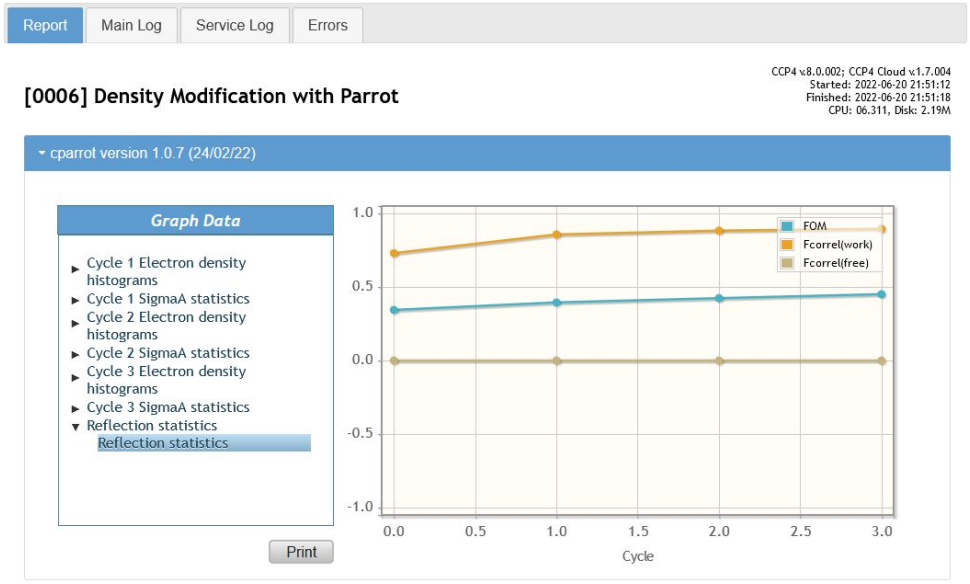
This is the results page for Parrot in CCP4i2. The text at the top states the estimated number of copies along with the corresponding solvent content, and would report on any NCS that was found. It also describes the expected quality of the phases based on the final figure of merit (FOM). There should hopefully be an increase in the FOM as the calculation proceeds.

CCP4 Cloud - Parrot



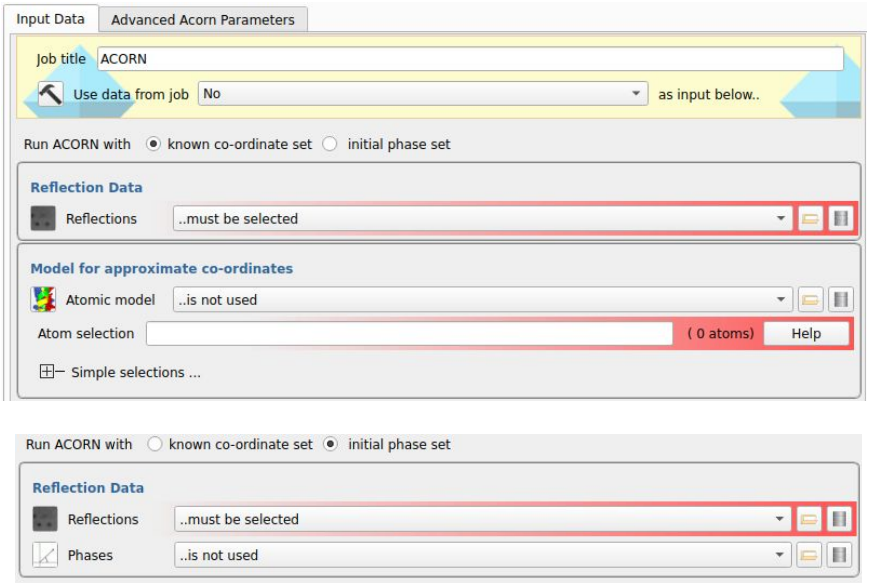
The is the interface to Parrot in CCP4 Cloud. It requires a structure revision that contains a set of starting phases. If the structure revision also has a model or a heavy atom substructure that can be used for NCS detection. The number of cycles is chosen automatically by default, but it can be overridden. There are other advanced options that would not normally be need to be changed, including options which are not in CCP4i2 to turn off the solvent flattening or histogram matching steps. Unlike the CCP4i2 task, CCP4 Cloud uses the number of copies specified in the asymmetric unit contents to calculate a solvent content that is given to Parrot. If the number of copies is not known it is recommended to create branches in the project tree with the different options, starting with the most likely.

CCP4 Cloud - Parrot



CCP4 Cloud shows the same graph of the FOM for each cycle under Reflection statistics.

CCP4i2 - ACORN

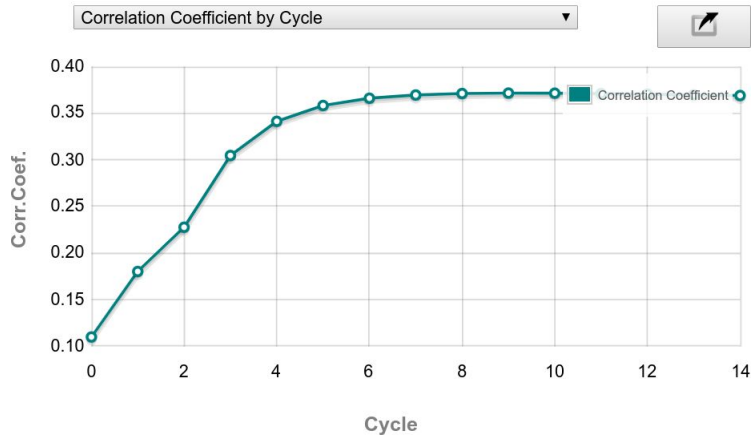


This is the interface for ACORN in CCP4i2. It requires reflections and either coordinates or phases to generate the initial density map. The interface is simpler than that of Parrot because it does not use a solvent mask or NCS averaging and the number of cycles is determined automatically. All of the other options are in the Advanced Acorn Parameters tab.

CCP4i2 - ACORN

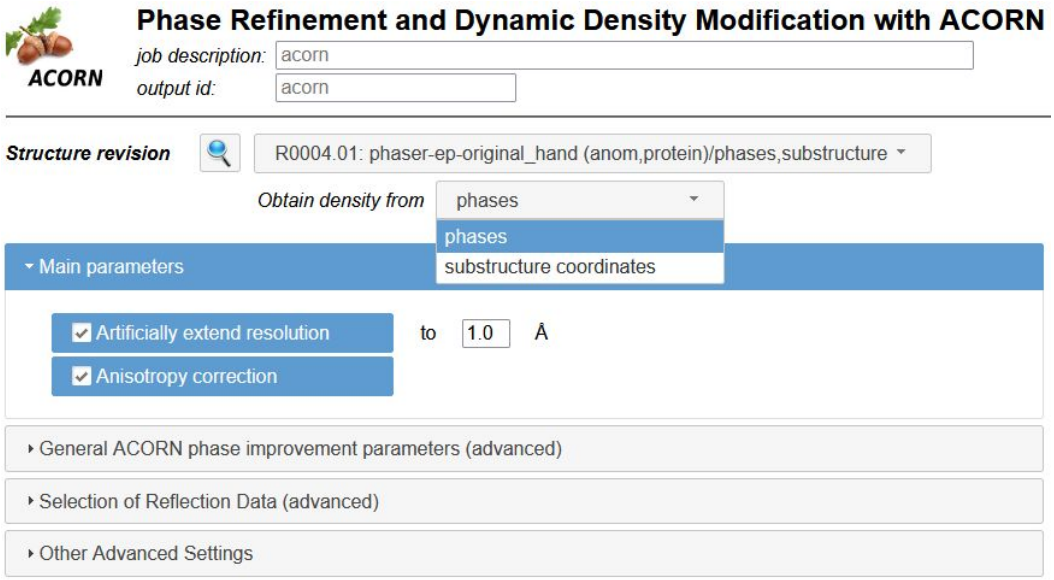
Results

Results for Acorn Run



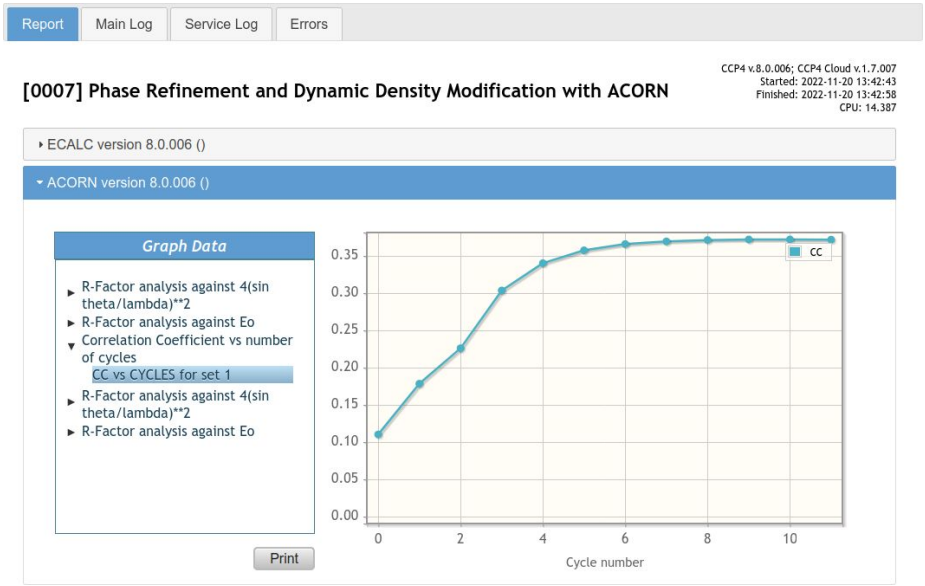
The results page shows a graph of correlation coefficient against the number of cycles, which should increase if ACORN is working.

CCP4 Cloud - ACORN



The CCP4 Cloud page for ACORN requires a structure revision with phases, but if the revision also contains model or substructure coordinates then those can be used to obtain the initial density map instead. There are options to choose whether to extend the resolution and perform anisotropy correction as well as other more advanced options.

CCP4 Cloud - ACORN



The report in CCP4 Cloud also has a graph of correlation coefficient against the number of cycles.

Concluding Tips

- Parrot is included in ModelCraft (a pipeline that combines density modification, model building and refinement) so try that first.
- With high resolution data, try ACORN to improve the phases. SHELXE is also good at improving phases and providing an initial model trace.
- NCS helps a lot. Try combining Parrot and MR or Coot interactive model building to get to the point where NCS can be detected.
- Make sure the solvent content is correct when running Parrot.
- Re-submit to ModelCraft for automated model building once the starting point has been improved.

ModelCraft is an model-building pipeline that is available in both CCP4i2 and CCP4 Cloud. It includes density modification with Parrot, so it is worth first trying ModelCraft first as it may also build most of the model automatically. If ModelCraft does not work with the input data you give it then it is worth trying to improve the starting phases through the techniques suggested in this slide.

Acknowledgements

- Kevin Cowtan
- JCSG data archive
- Garib Murshudov, Raj Pannu, Pavol Skubak
- Eleanor Dodson, Paul Emsley, Randy Read, Clemens Vonrhein
- Many others
- Funding: The Royal Society, BBSRC, CCP4