Planning and carrying out automated structure determination using SAD phasing
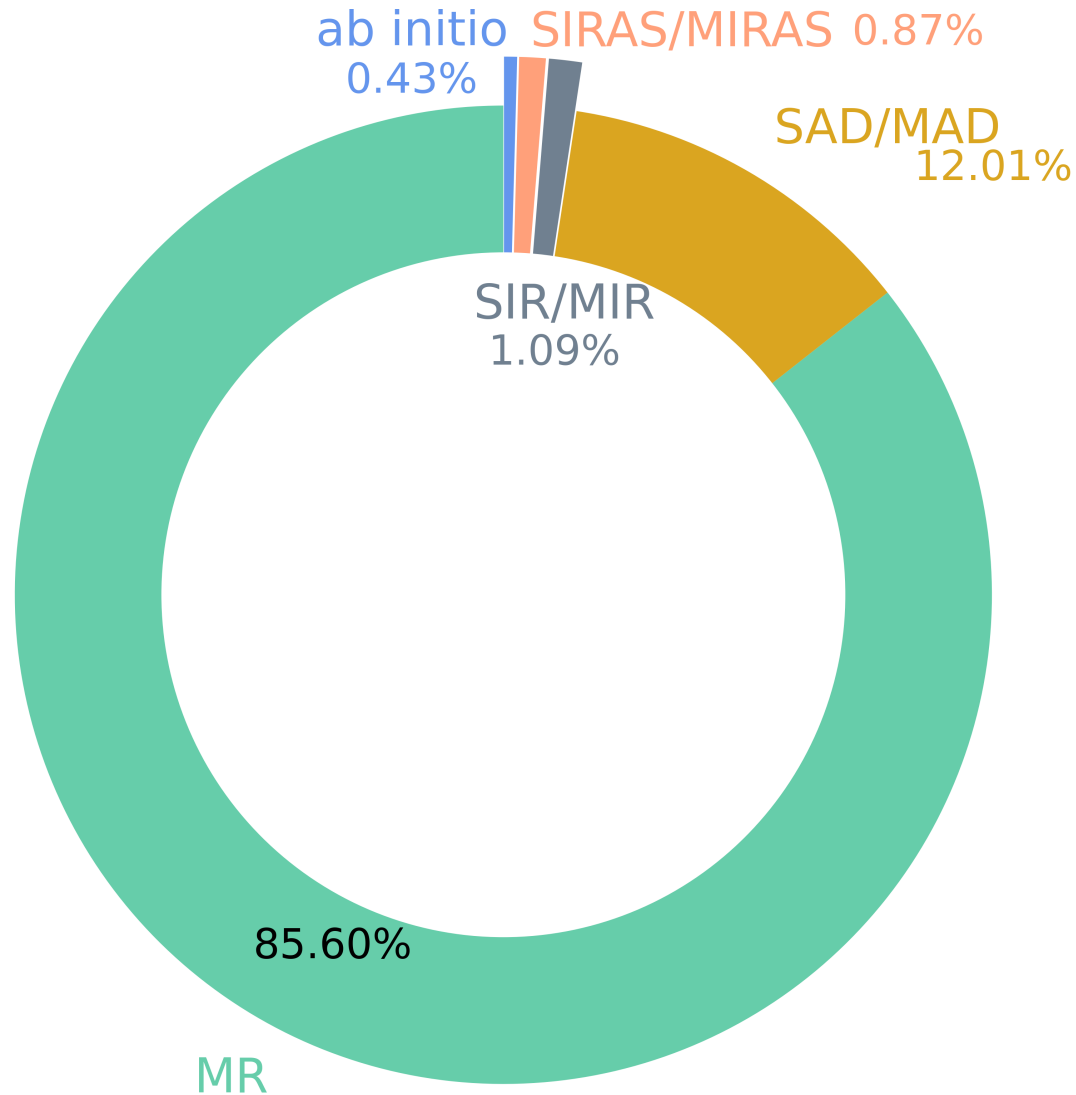
Single wavelength Anomalous Diffraction

Dorothee Liebschner, Lawrence Berkeley National Laboratory
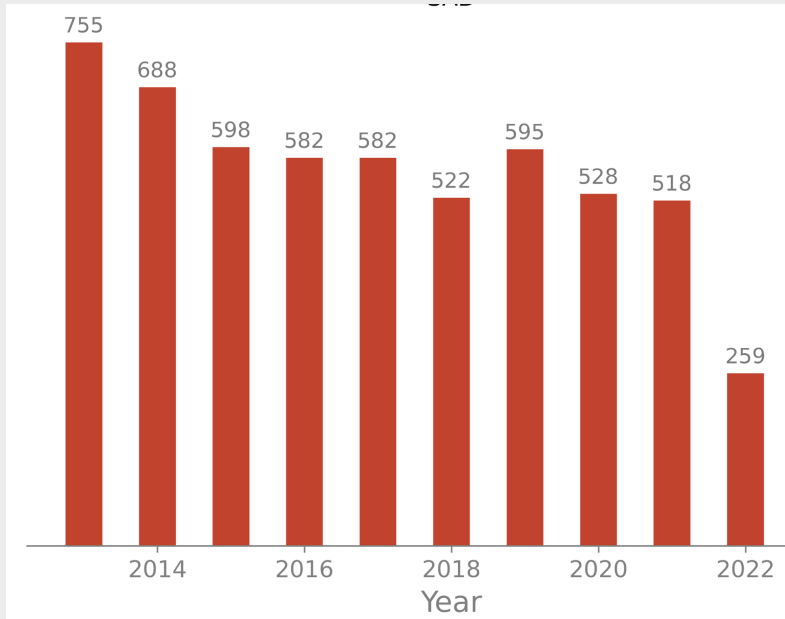Tom Terwilliger, Los Alamos National Laboratory

Phasing methods in the PDB

ab initio 0.43%

SIRAS/MIRAS 0.87%
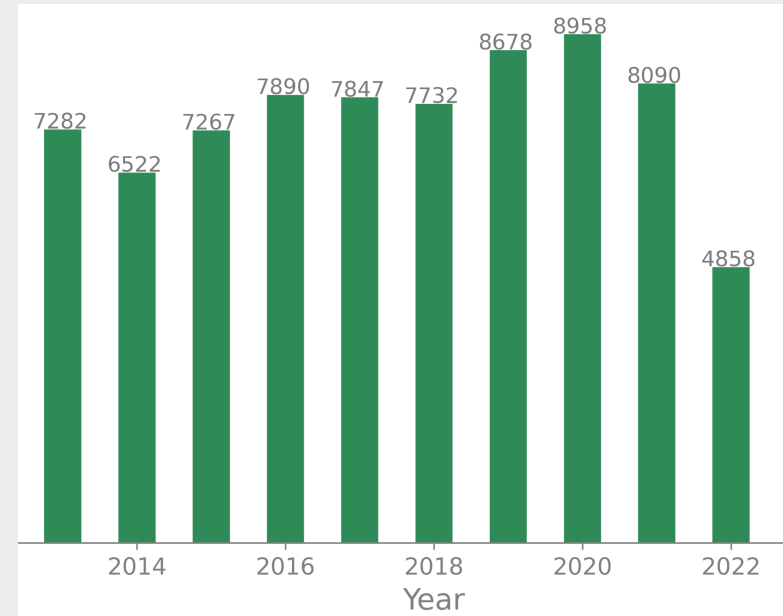
SAD/MAD 12.01%

SIR/MIR 1.09%

MR 85.60%

Note: Not all models in the PDB have (correct) info

Phasing methods in the PDB

SAD

Molecular replacement

- Less experimental phasing

- More and more MR structures

- Trend will continue with predicted models (AlphaFold etc.)

Some limitations of predicted models

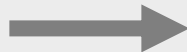- Predicted models are great to jumpstart structure determination
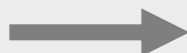
- Limitations:

Only protein ➡ No water, ions, covalent modifications, carbohydrates, ligands, DNA, RNA

Little information about residues that are far apart ➡ Models may have distortions and incorrect domain relationships
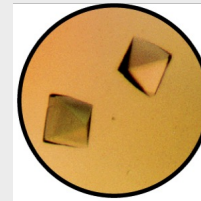
Trained on good and poor structures ➡ Parameters may systematically include poor geometry
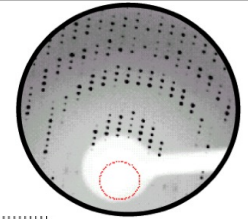
→ We still need experimental phasing.

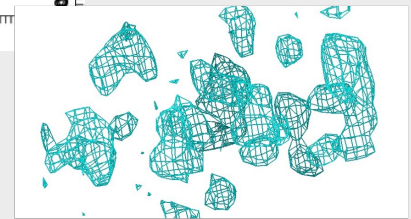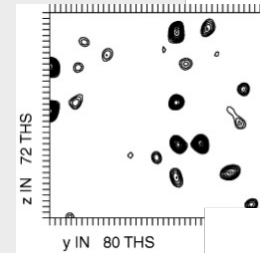# Solving a structure with SAD phasing (Se)

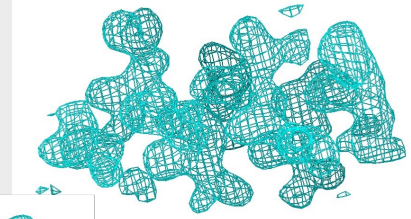Crystals with SeMet

↓

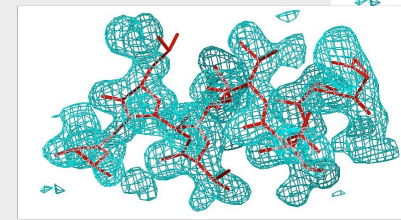Collect anomalous SAD data

↓

Locate Se atoms

↓

Phasing (calculate density map)

↓

Density modification (improve map)

↓

Model building

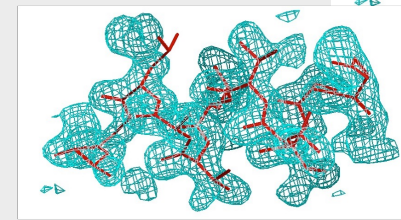# Solving a structure with SAD phasing (Se)
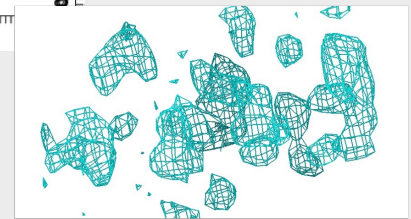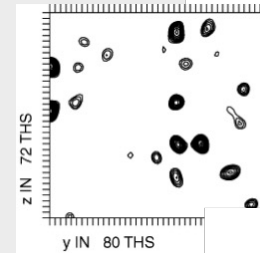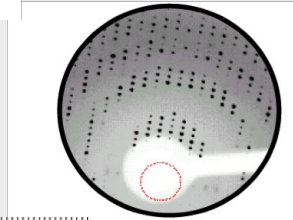
Planning the experiment
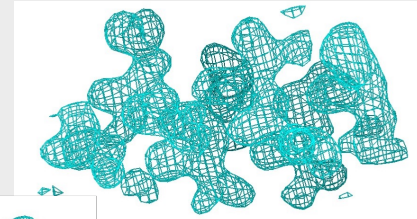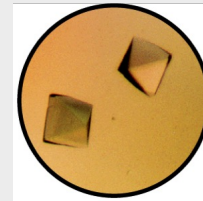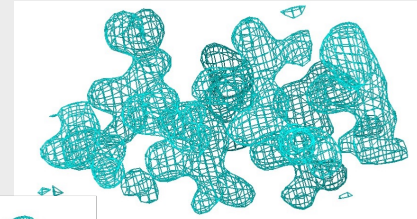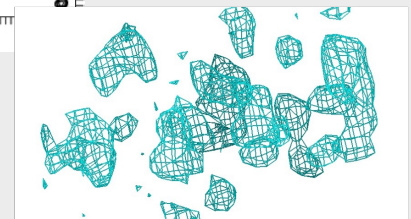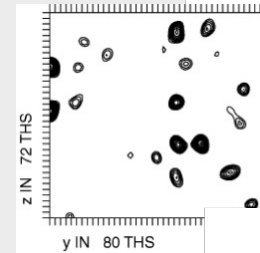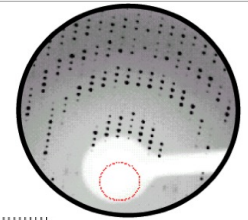
Automating the analysis

Improving the map

Will I solve my SAD structure?

Planning the experiment

Automating the analysis

Improving the map
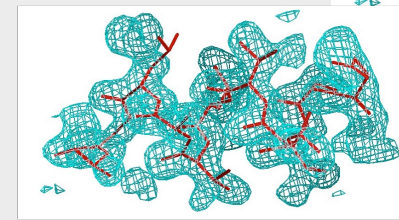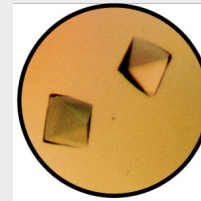
Will I find the anomalous substructure?

How many sites?

Are sites ordered?

Anomalous atom?

Wavelength?

Accurate data?

How many reflections?
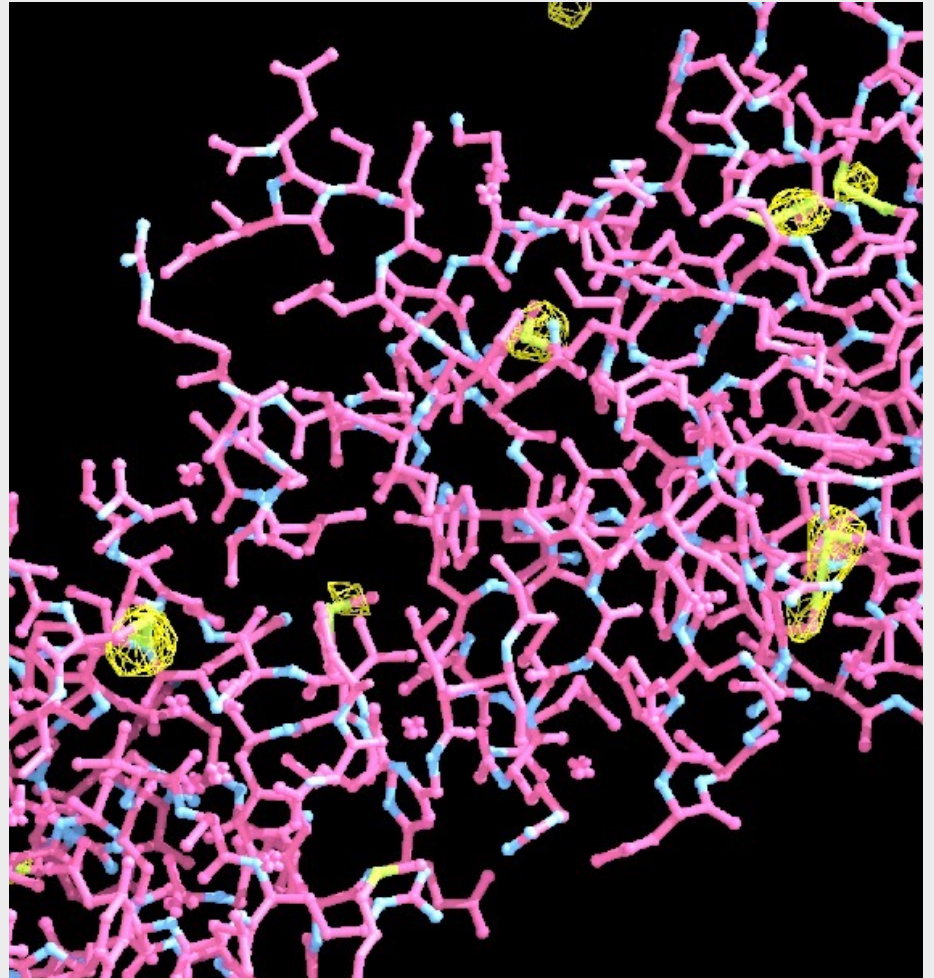
Key steps for SAD phasing

1. Find the substructure
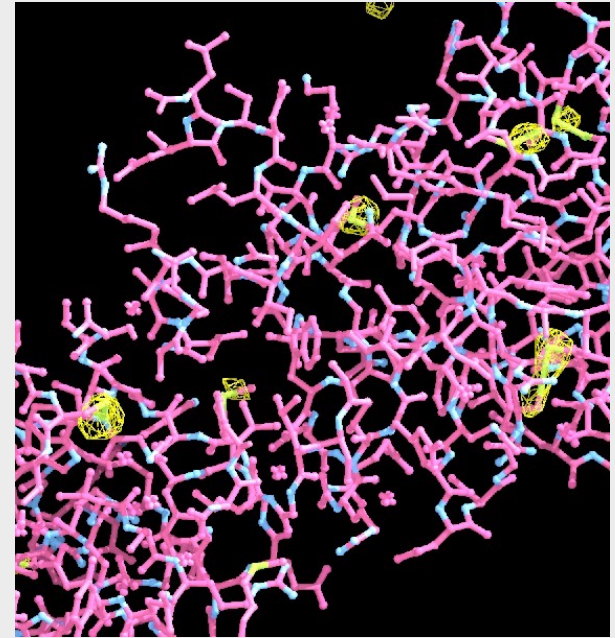
Anomalous signal S

2. Calculate an interpretable map

Anomalous correlation CC_{ano}
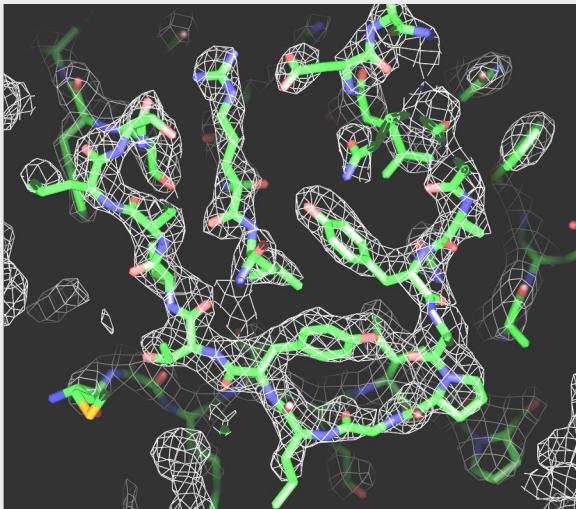
Anomalous signal

1. Find the substructure

Anomalous signal S
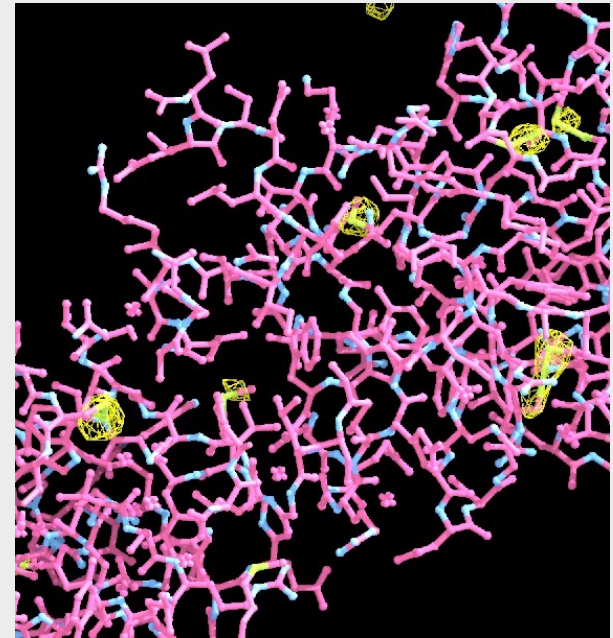
- Peak height in anomalous difference Fourier

- "Information per site"

- Substructure likely to be found if S > 10

Anomalous correlation
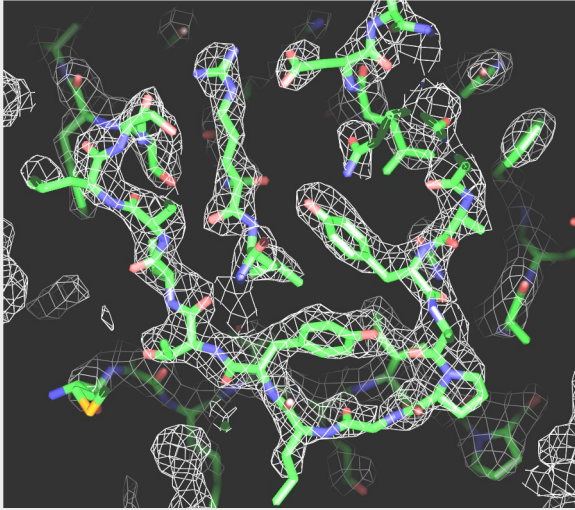
2. Calculate an interpretable map

Anomalous correlation CC_{ano}

- Correlation of anomalous differences with ideal

- Accuracy of anomalous data

- Accuracy of phasing

Anomalous signal: key to finding substructure

Anomalous signal S

Will I find sites?

← Peak height in anomalous difference Fourier

We can only calculate this once we solved the structure.

→ Estimate it

Anomalous signal: key to finding substructure
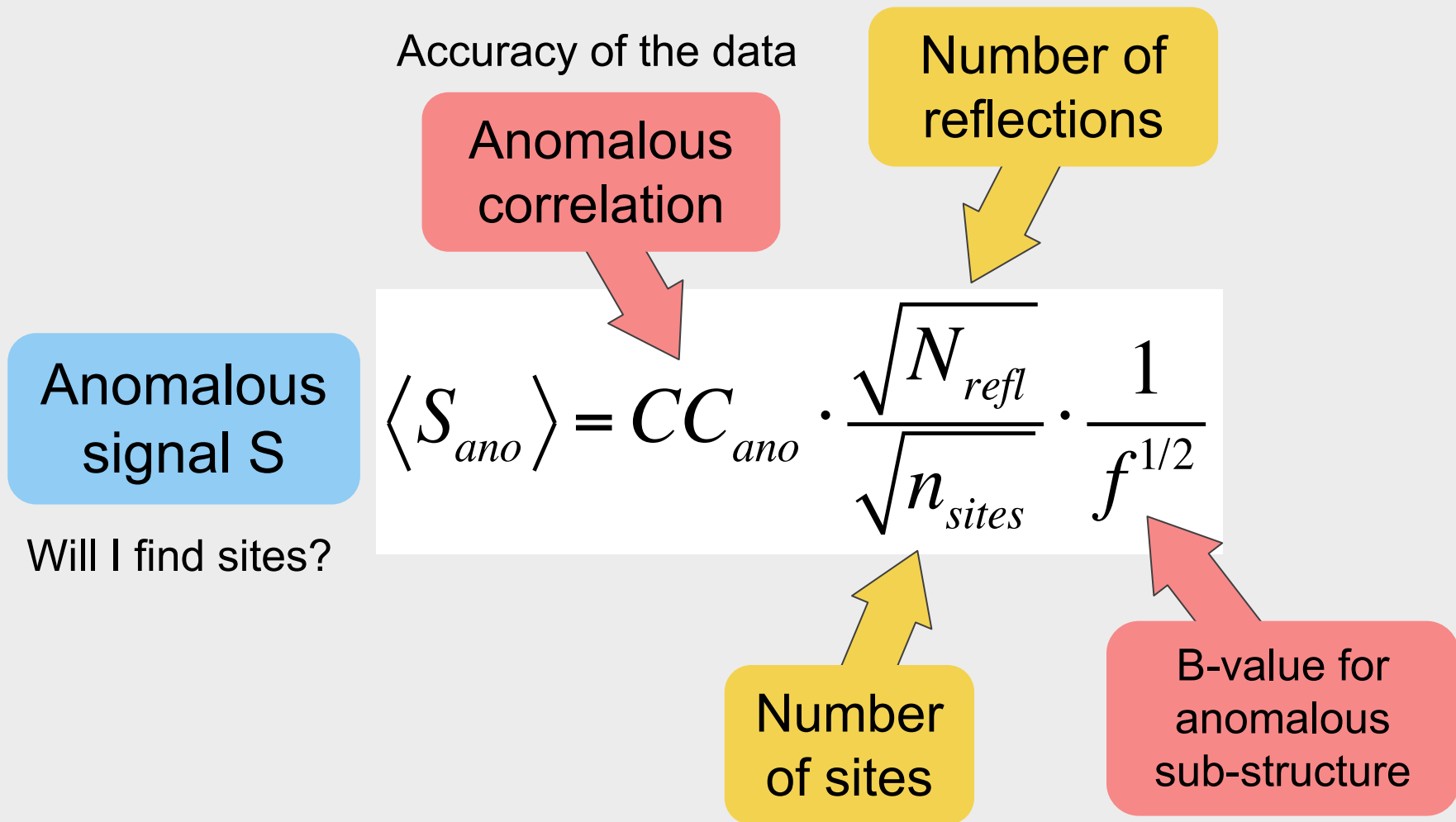
Accuracy of the data

Anomalous correlation

Number of reflections

Anomalous signal S

$$\langle S_{ano} \rangle = CC_{ano} \cdot \frac{\sqrt{N_{refl}}}{\sqrt{n_{sites}}} \cdot \frac{1}{f^{1/2}}$$

Will I find sites?

Number of sites

B-value for anomalous sub-structure

Anomalous signal: key to finding substructure

Relationship between the anomalous signal and the solution of the anomalous substructure

"What value of S (anomalous signal) do I need to solve the substructure?"

Anomalous signal: key to finding substructure

Relationship between the anomalous signal and the solution of the anomalous substructure
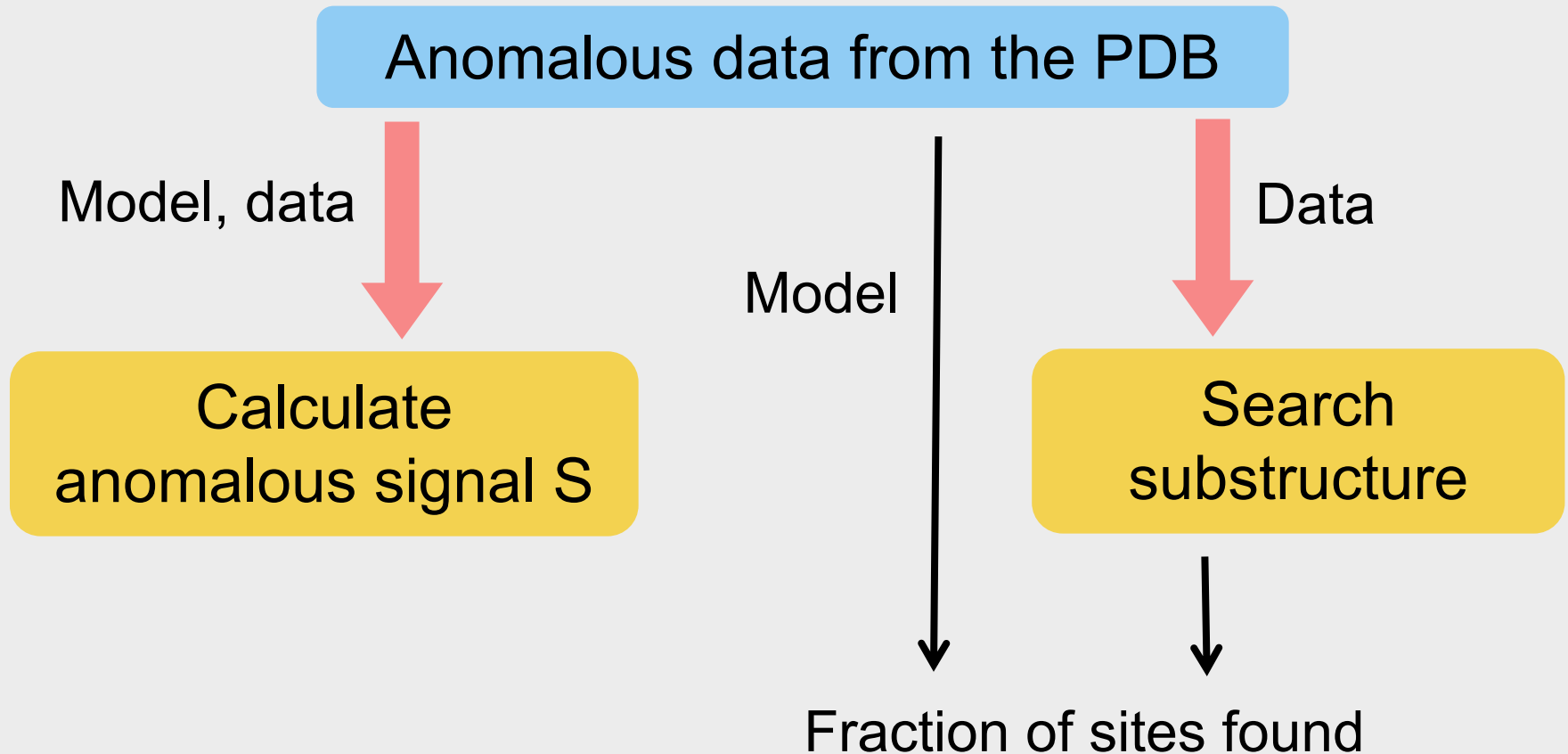
Anomalous signal: key to finding substructure

Relationship between the anomalous signal and the solution of the anomalous substructure
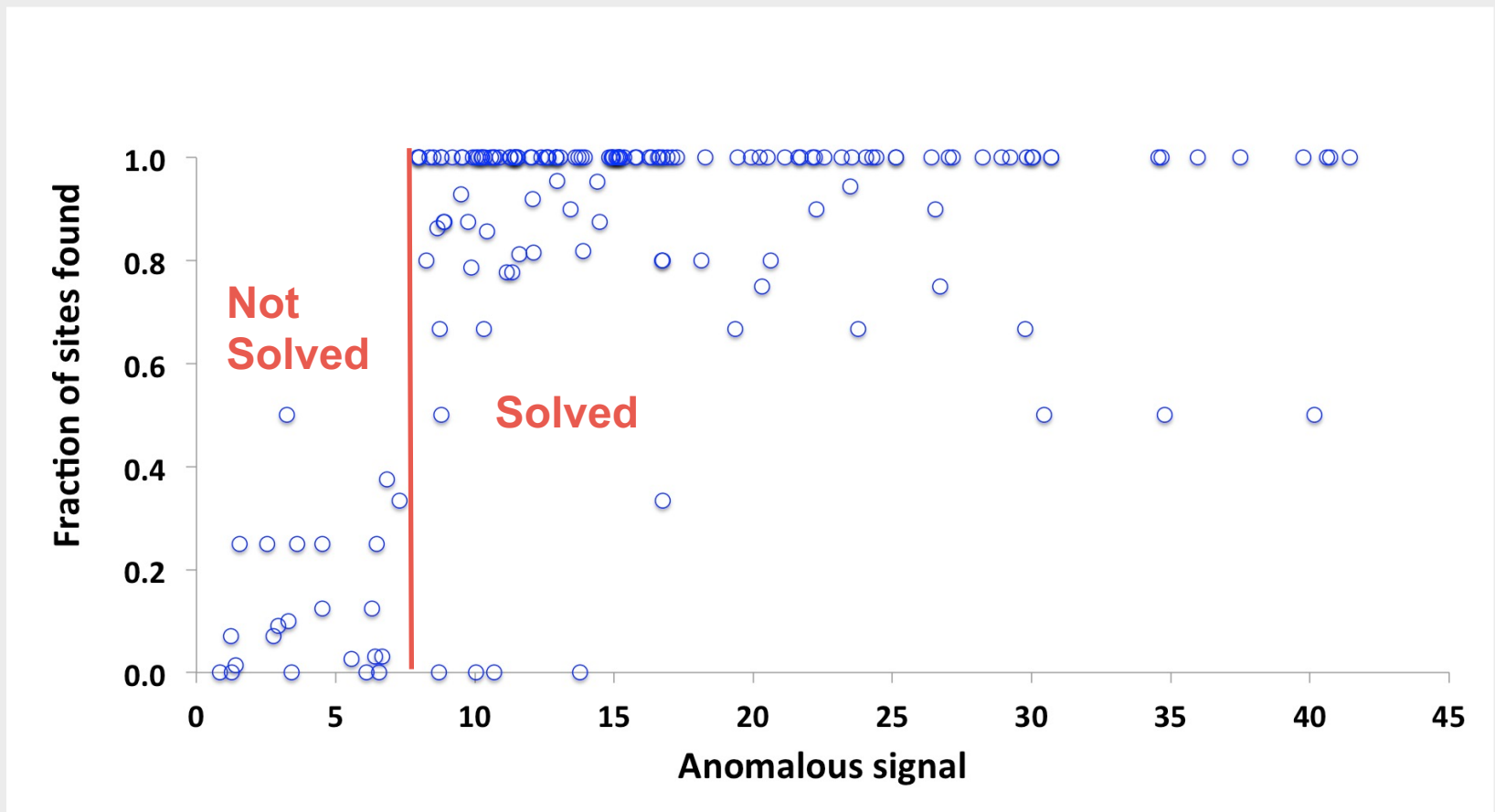
Estimating the anomalous signal

I/sigma (accuracy of data)

Anomalous atom (wavelength, f")

Number of sites

Resolution

Sequence

Anomalous signal S

Can I solve my structure by SAD phasing? Planning an experiment, scaling data and evaluating the useful anomalous correlation and anomalous signal. Terwilliger TC, Bunkóczi G, Hung L-W, Zwart PH, Smith JL, Akey D, Adams PD Acta Cryst. D72, 359-374 (2016).
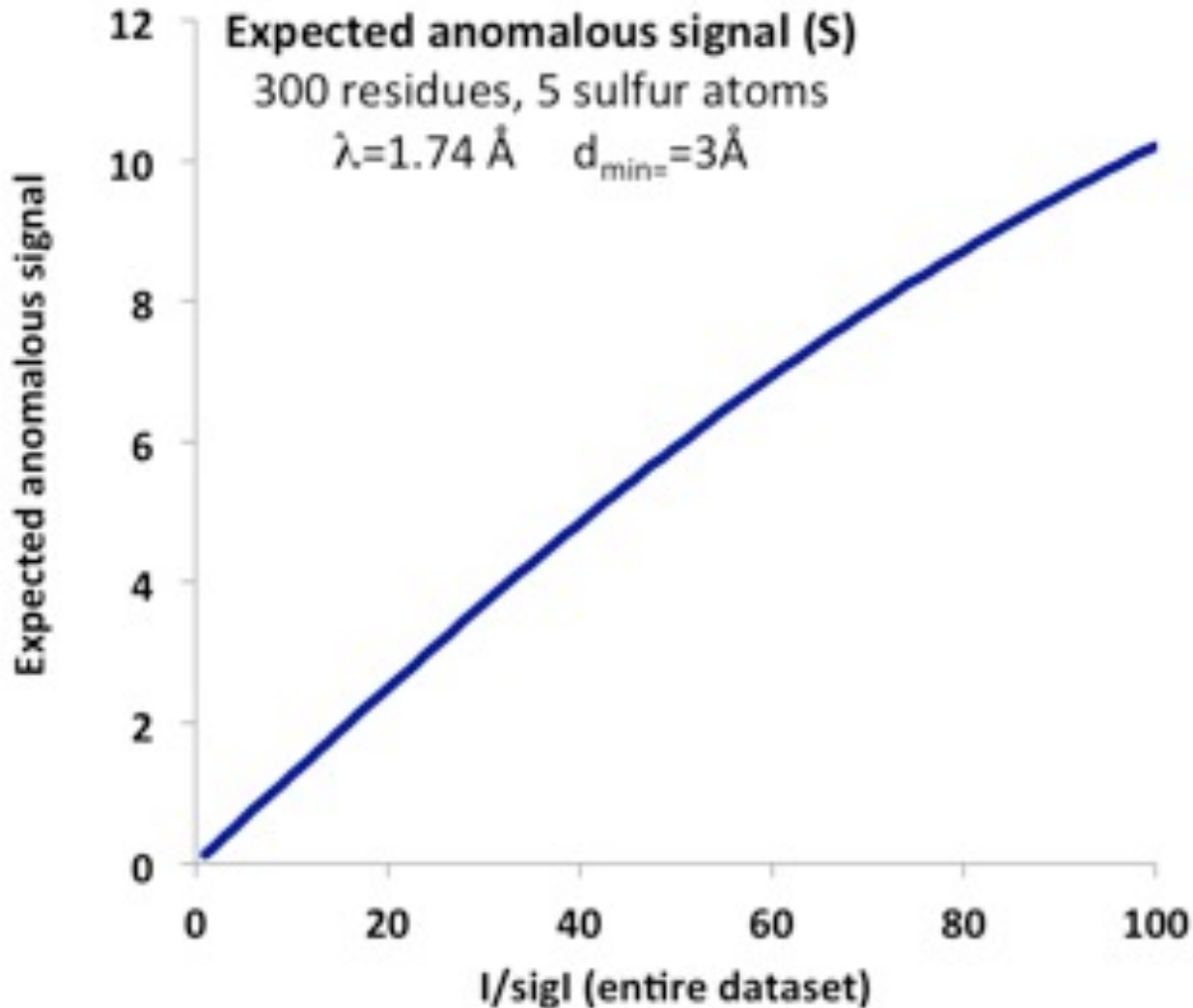
Estimating the anomalous signal

Expected anomalous signal (S)
300 residues, 5 Se atoms
$\lambda=0.9792$ Å $d_{min=}=3$Å
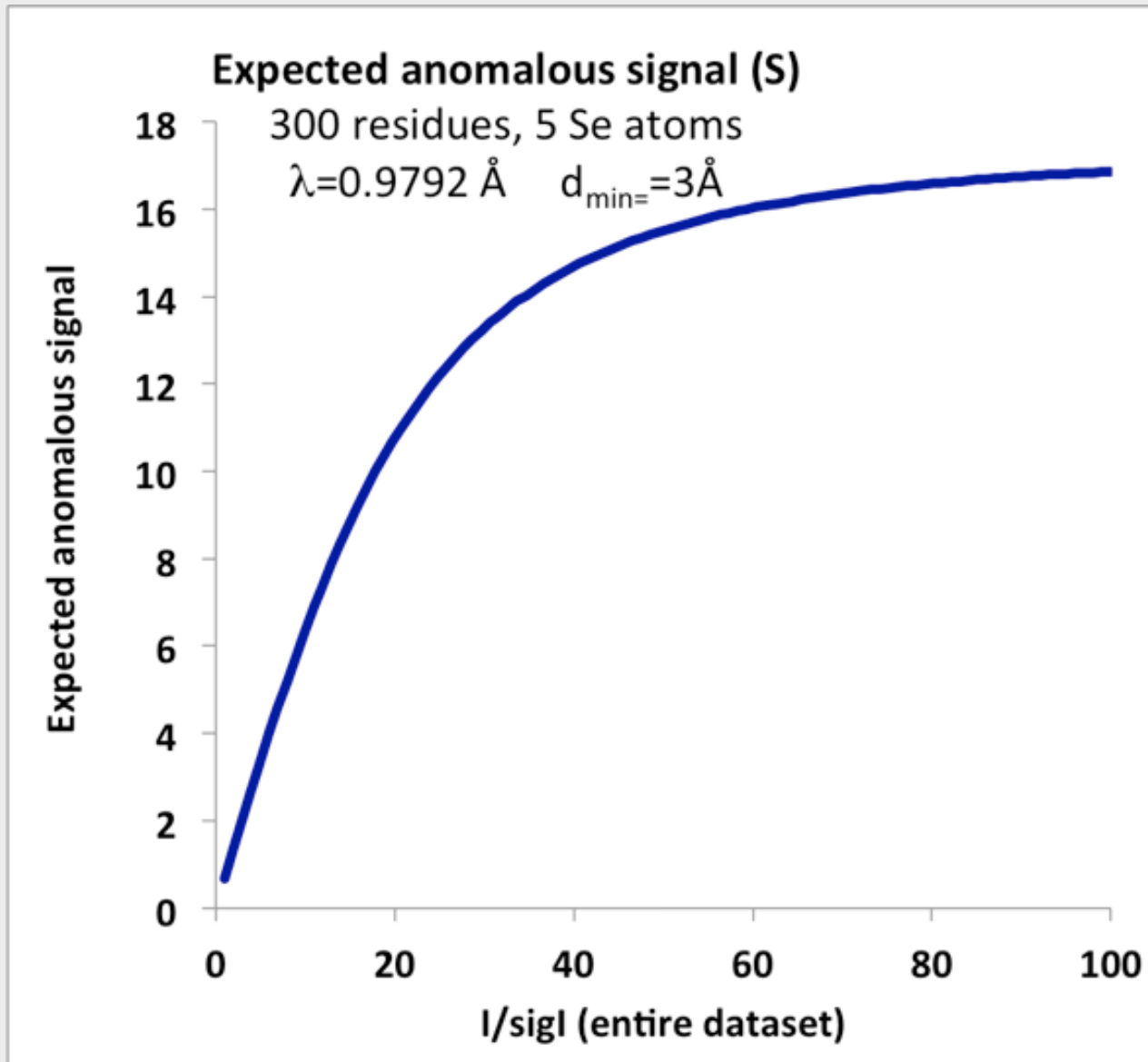
Se-SAD

5 sites

3 Å data

300 residues

0.9792 Å

Estimating the anomalous signal

Summary

- We can estimate the anomalous signal S from the data
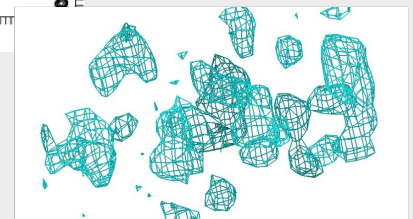
- If S > 10 $\rightarrow$ substructure is likely to be found

- We can simulate the anomalous signal (before doing the experiment)

Will I solve my SAD structure?

Planning the experiment

Automating the analysis

Improving the map
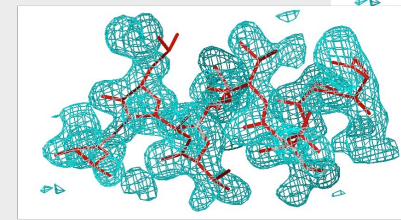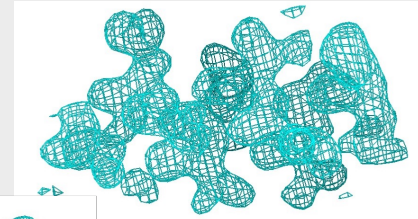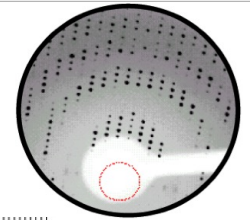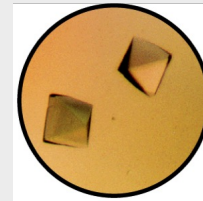
Why automate structure determination?

Makes straightforward cases easier

… and difficult cases feasible for experts

Speeds up the process

Reduces errors

Allows you to try more possibilities

Decision-making in automation

What does a good electron density map look like?

Using expected features of maps to make decisions

Decision-making in automation

Which map is better?

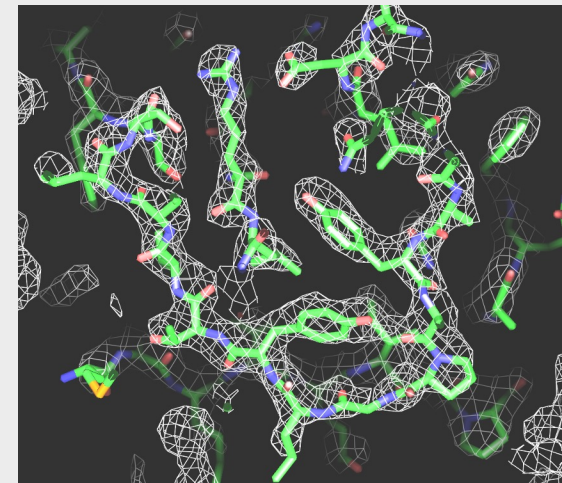Density looks like protein

Flat solvent region

Connected density
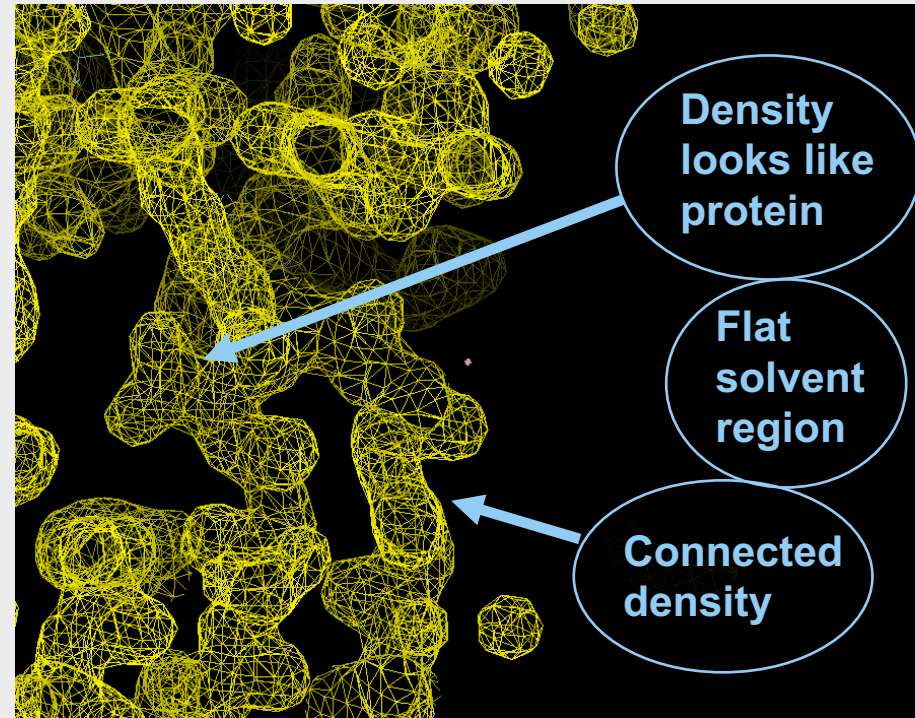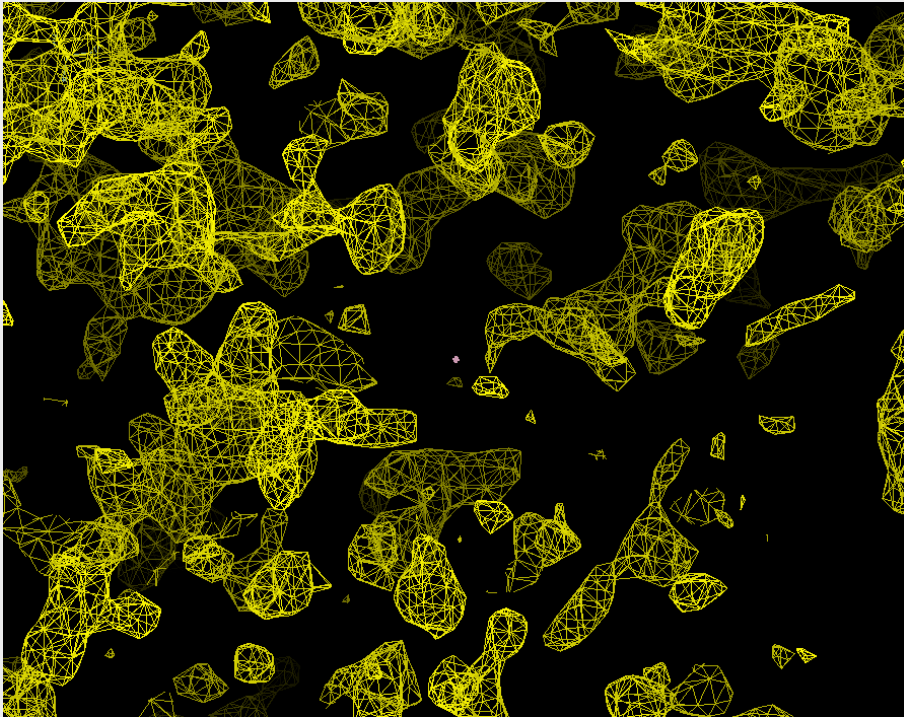
Histograms of density have positive skew

Skew = measure of the asymmetry

Typical histogram of electron density

Low density: Points between atoms and in solvent region

High density: Points on top of atoms

Histogram skewed to the right

Histograms of density have positive skew

Poor map
(inverse hand)

Poor map:
nearly-perfect
Gaussian
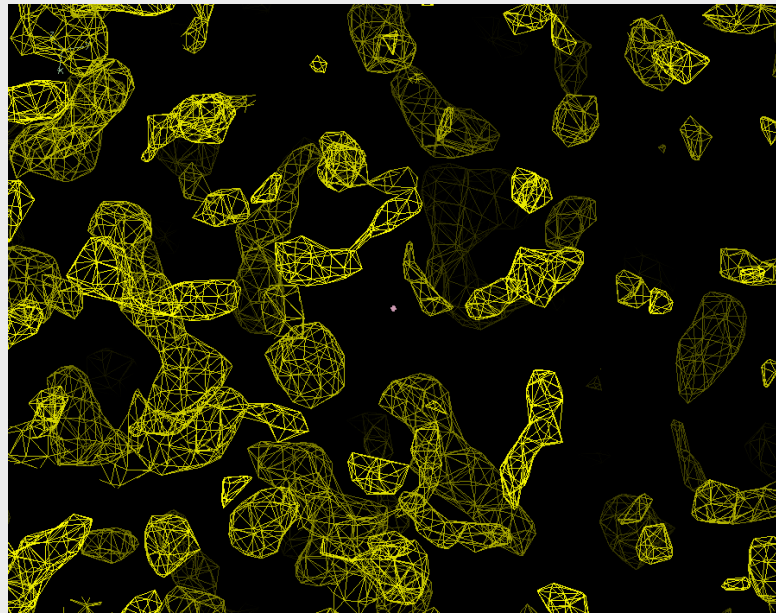
Good map:
slight positive
skew
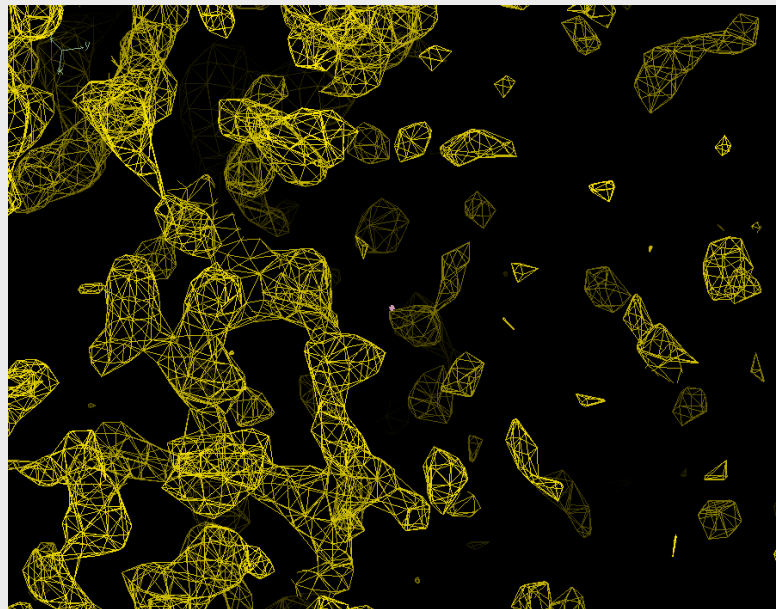
Good map

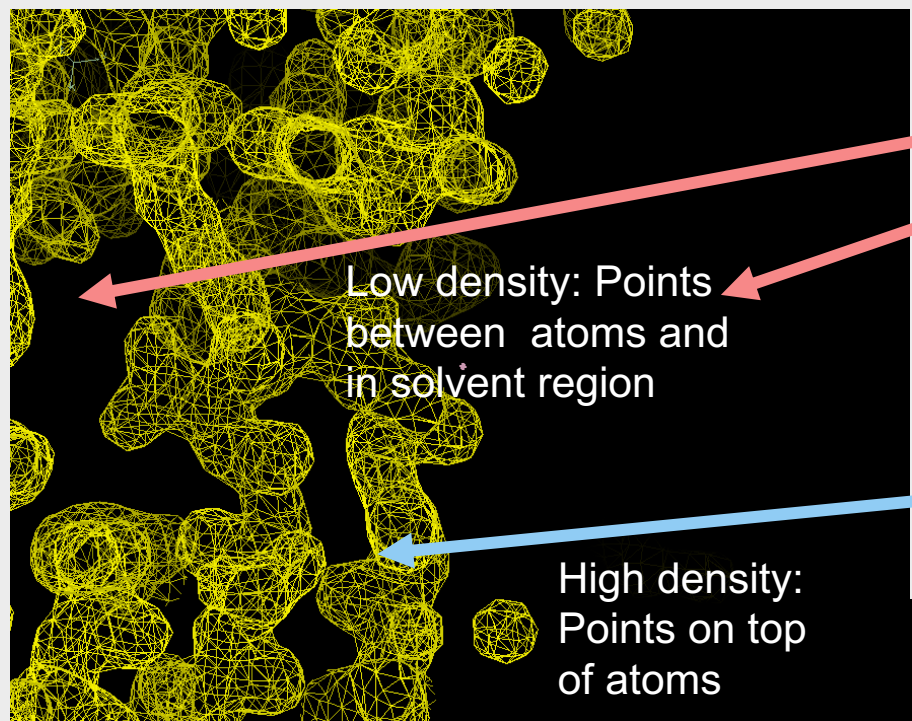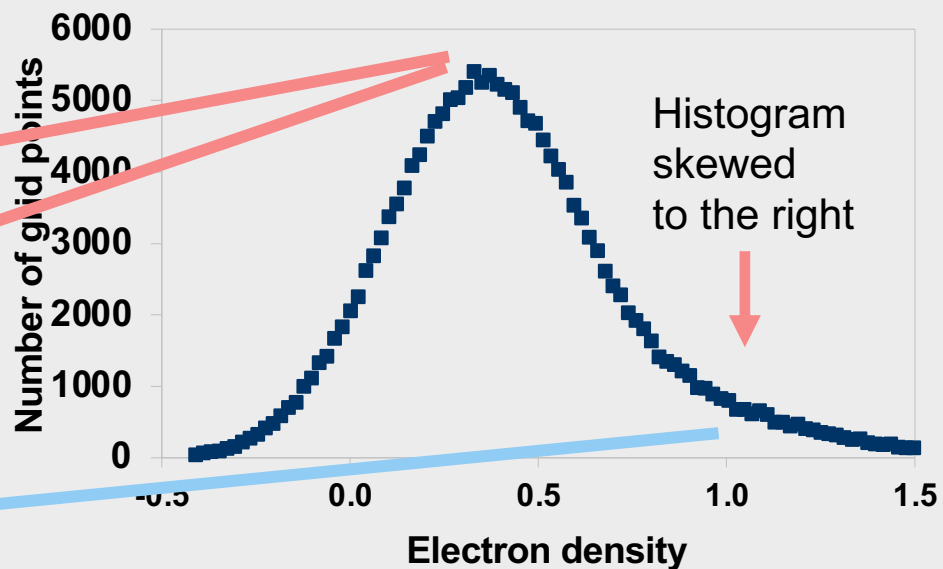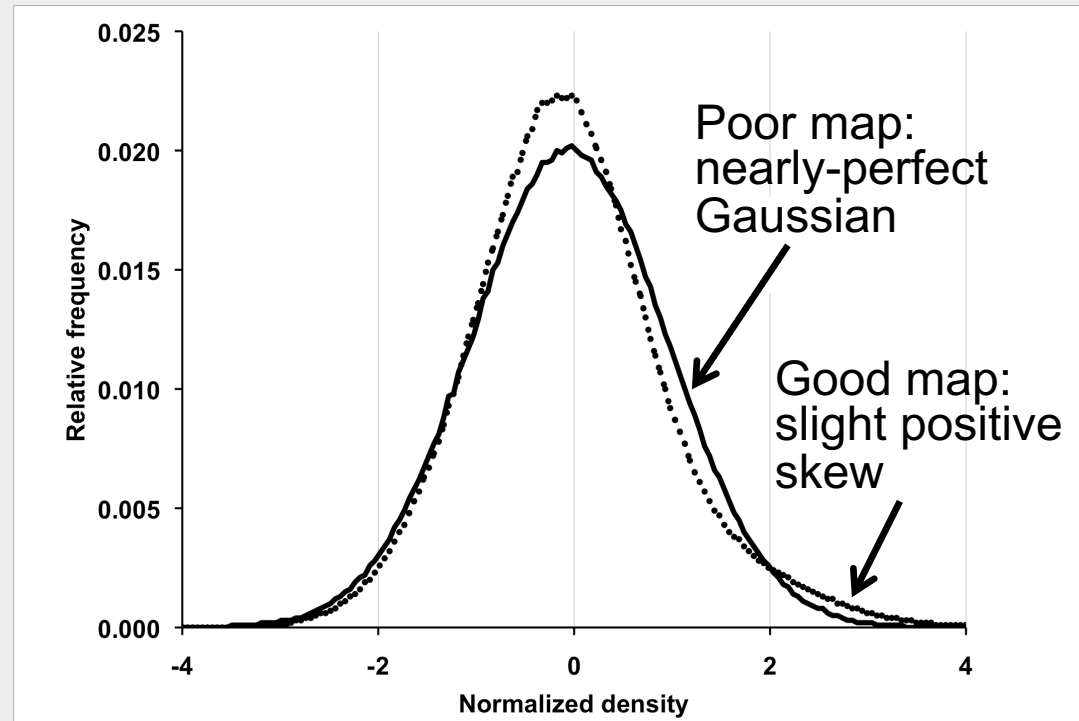# Histograms of density have positive skew

How well does the skew reflect map quality?

➢ 247 MAD, SAD, MIR datasets with final model available

➢ Run *phenix.autosol* on each dataset → maps

➢ Score the maps based on skew

➢ Compare the scores with the actual quality of the maps (correlation to model map)

Positive skew in good maps
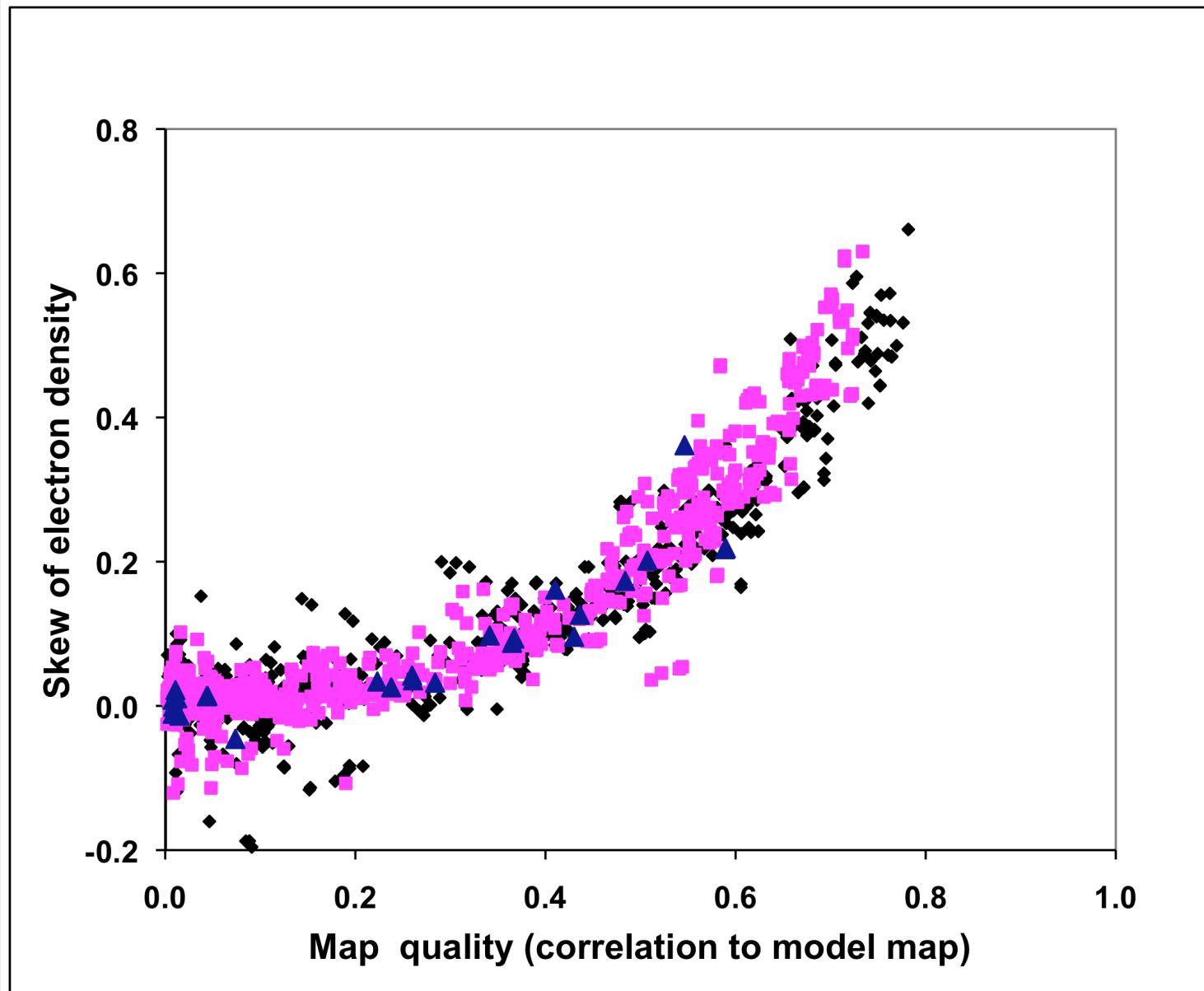
Estimate map quality from skew

Skew depends on map quality   →   Estimate map quality from skew
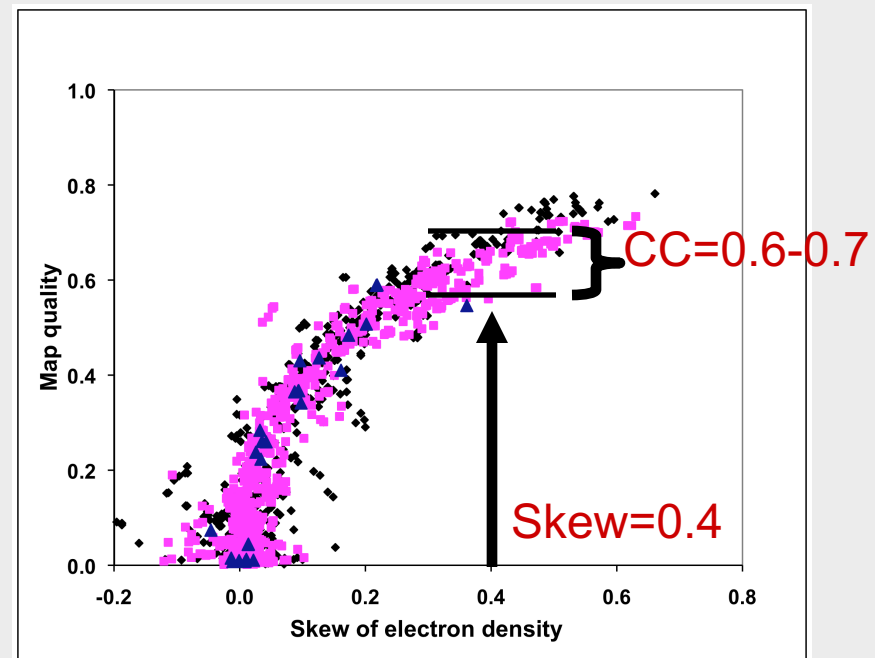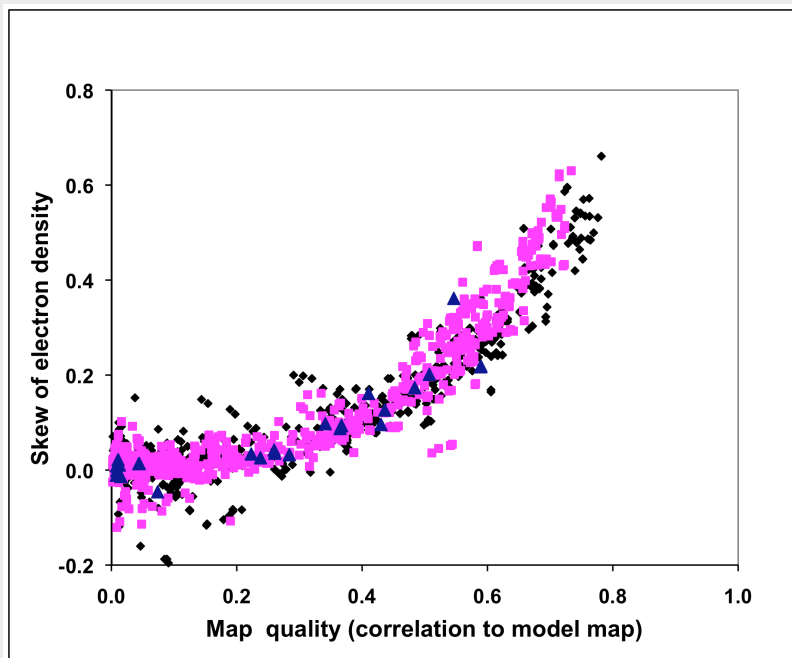
Summary

- The skew reflects map quality

- We can estimate map quality from the skew

- Use the skew for decision making (automation)

Will I solve my SAD structure?

Planning the experiment

Automating the analysis

Improving the map
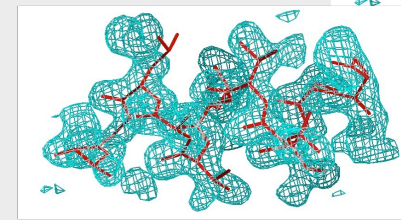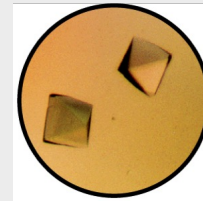
Map improvement by density modification

What does a good electron density map look like?

↓

Using expected features of maps to improve maps

Density modification = "phase improvement"

Experimental Data

Initial phases

Improved phases
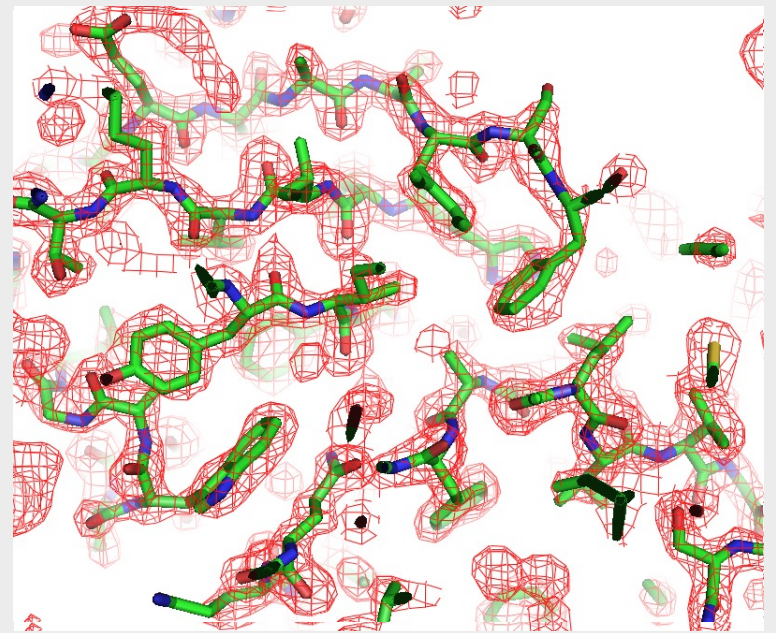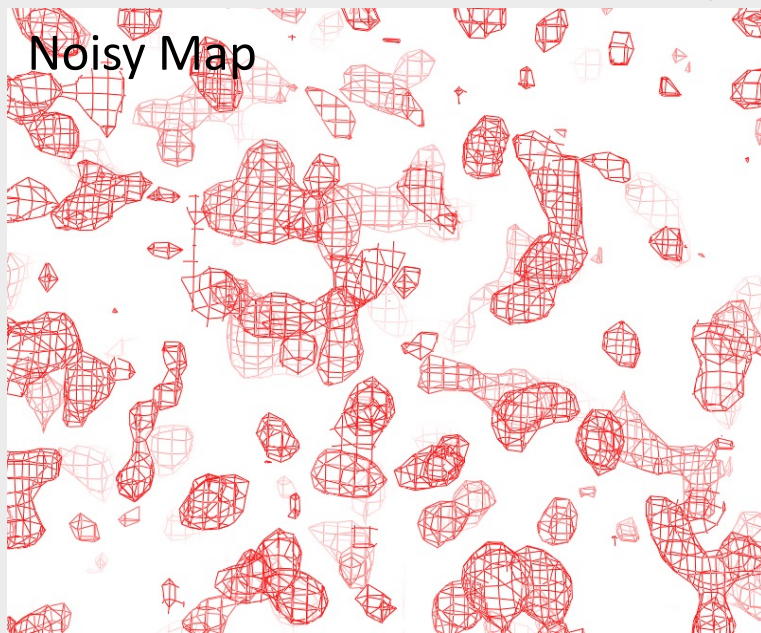
Noisy Map

Basis of density modification
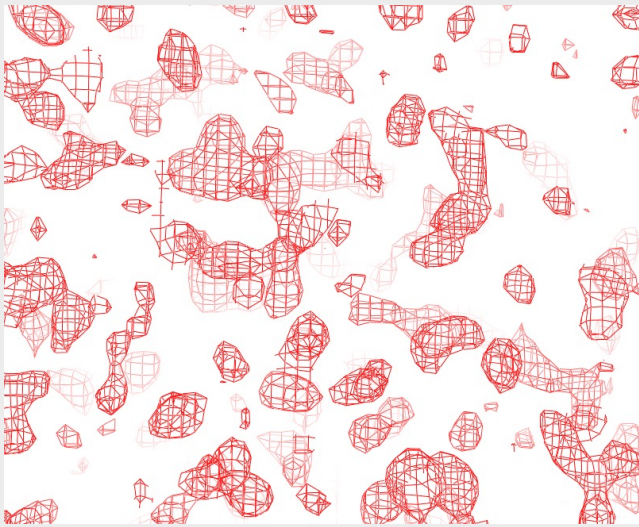
Noisy map
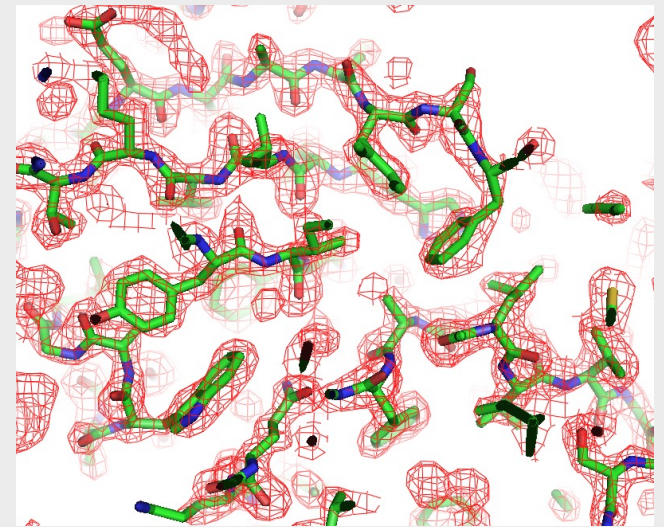
Clear map

Improve the noisy map to create the clear map using two key facts:

1. We know a good map when we see it

2. Improvement anywhere means improvement everywhere

Density modification: strategy

Identify local expected density

Find phases consistent with **experiment that** lead to **expected density**

Density everywhere is improved
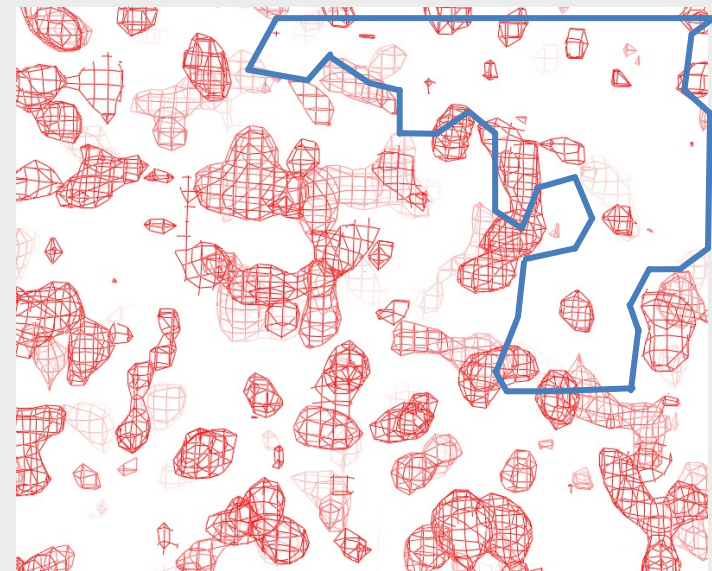
Noisy map
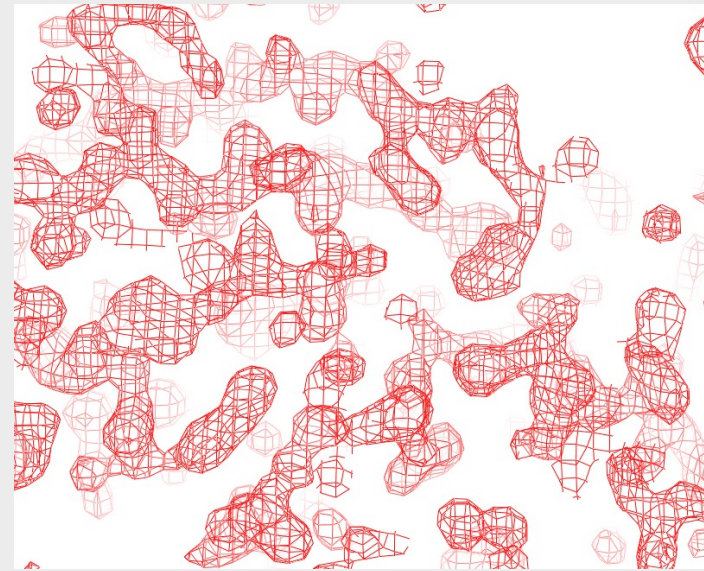
Clear map

1-dimensional example to illustrate the details of statistical density modification

crystal with one protein atom

A Fourier sum of sines and cosines

Protein atom

Flat solvent

Fourier terms
sum to make the curve

A Fourier sum of sines and cosines

Protein atom
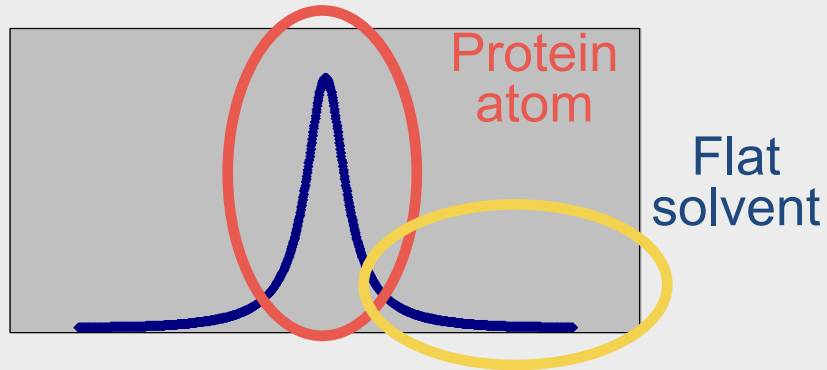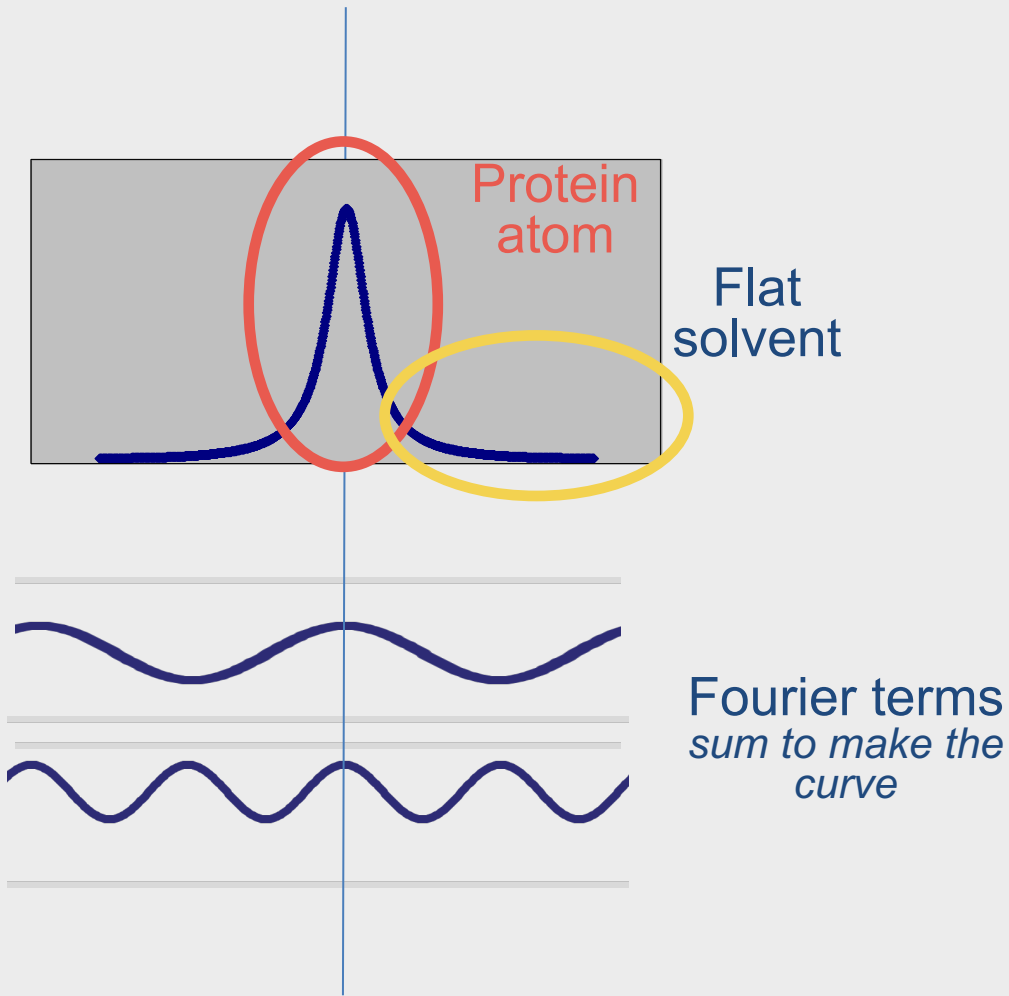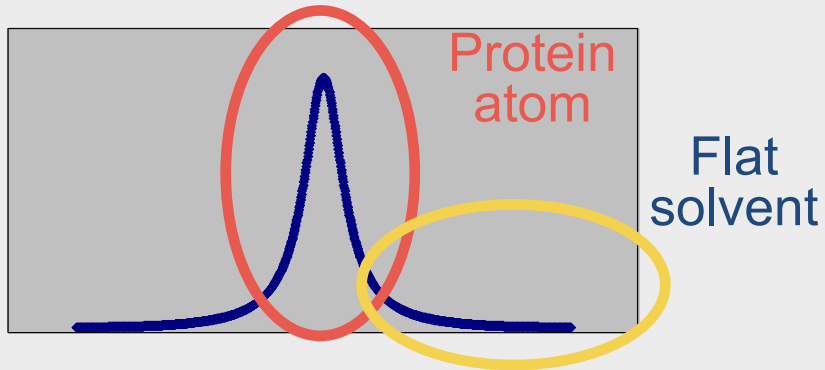
Flat solvent

Find out the phase of one Fourier term using:

1) All other Fourier terms

2) Flat solvent

A Fourier sum of sines and cosines

Protein atom
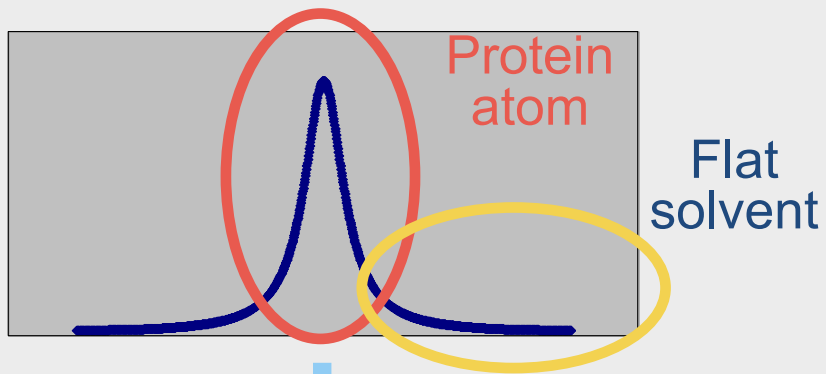
Flat solvent

Take out one Fourier term:
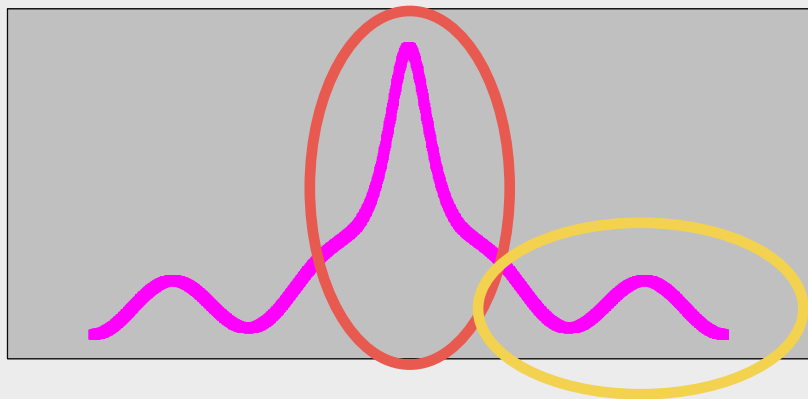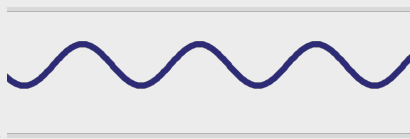
Using flat solvent to identify phase of one term
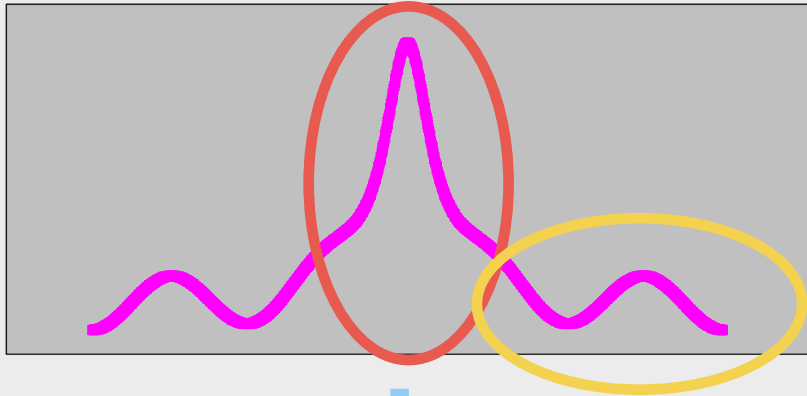
One Fourier term:
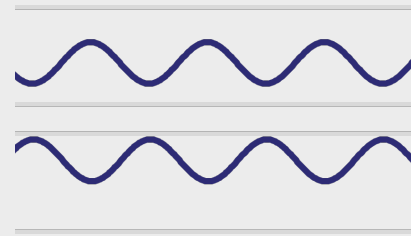
Correct phase

Phase 180° off

Adding the incorrect phase

Adding the correct phase
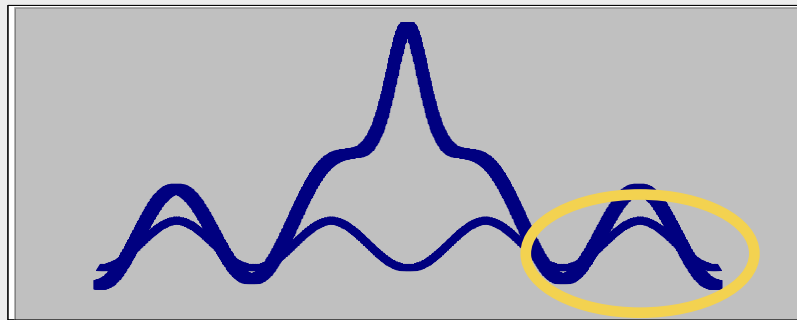
Solvent gets worse
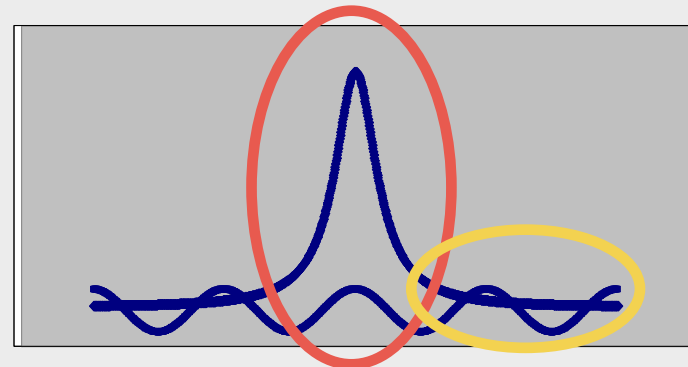
Density modification of real maps

Expectation about the flatness of the solvent

→ Identify the phase of one Fourier term
→ Improve the map in the protein region
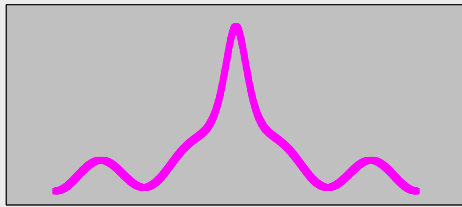
= Transfer information from one part of the map to another.
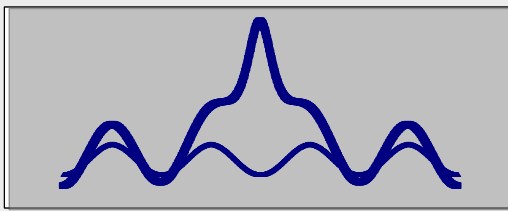
Incorrect
phase
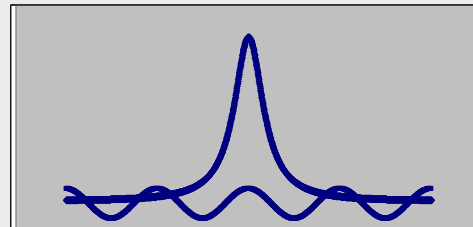
Correct
phase

Density modification of real maps
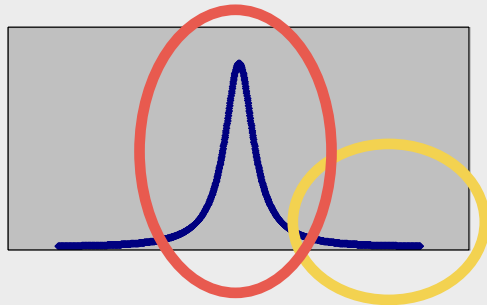
Real world:
Correct phase → $p_{map}(\varphi)$
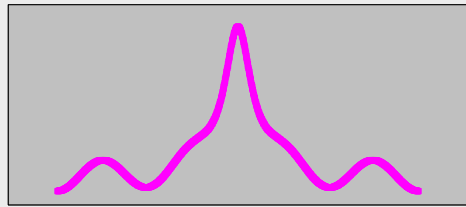
Experimental phase information = $p_{exp}(\varphi)$

Density modification phase probability:

$$p(\varphi) = p_{exp}(\varphi)\, p_{map}(\varphi)$$
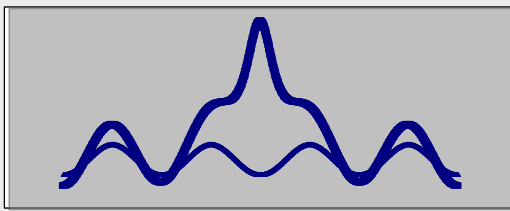
Incorrect phase
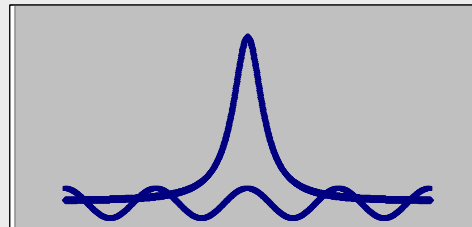
Correct phase

Summary
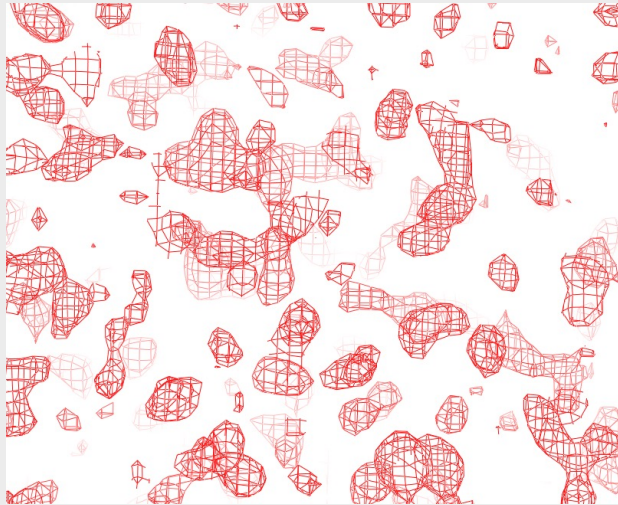
Improved phases
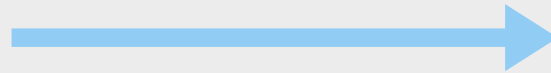
$$p(\varphi) = p_{exp}(\varphi)\, p_{map}(\varphi)$$
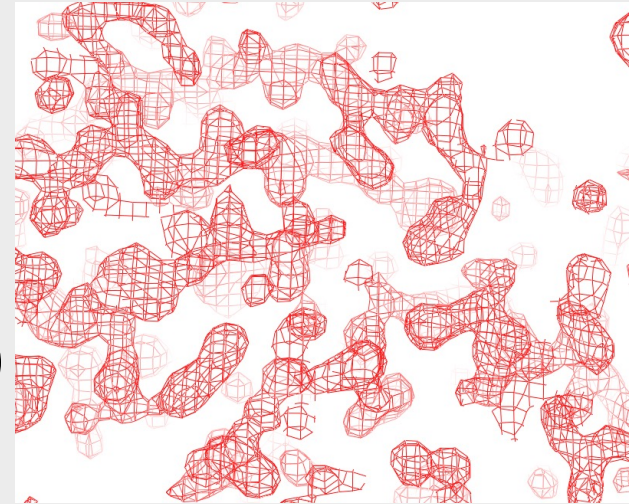
We know a good map when we see it

Improvement anywhere means improvement everywhere

Density modification transfers information from one part of the map to another

Summary

- You can simulate your SAD experiment → you can plan your experiment.

- Use prior expectations about density maps to automate the analysis.

- When you improve the map *anywhere*, the map will get better *everywhere*.

References

phenix-online.org

Phenix documentation

Tutorials with sample data

Video tutorials

Simulate a SAD experiment with...

www.youtube.com/c/phenixtutorials

Terwilliger, T. C. (2000). *Acta Cryst. D*. **56**, 965–972.

Terwilliger, T. C. (2002). *Acta Cryst. D*. **58**, 1937–1940

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 Cryst. D*. **64**, 61–69.

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 Cryst. D***65**, 582–601.

Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016). *Acta Cryst. D***72**, 346–358.

Terwilliger, T. C., Bunkóczi, G., Hung, L.-W., Zwart, P. H., Smith, J. L., Akey, D. L. & Adams, P. D. (2016). *Acta Cryst. D***72**, 359–374.

The Phenix Project

Lawrence Berkeley Laboratory

Paul Adams, Pavel Afonine, Dorothee Liebschner, Nigel Moriarty, Billy Poon, Christopher Schlicksup, Oleg Sobolev

University of Cambridge

Randy Read, Airlie McCoy

Los Alamos National Laboratory
New Mexico Consortium

Tom Terwilliger, Li-Wei Hung

UTHealth

Matt Baker

Duke University

Jane & David Richardson, Christopher Williams, Vincent Chen

An NIH/NIGMS funded Program Project

Liebschner D, *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Cryst. 2019 **D75**:861–877

Automated model-building

Planning the experiment

Automating the analysis

Improving the map

Building a model
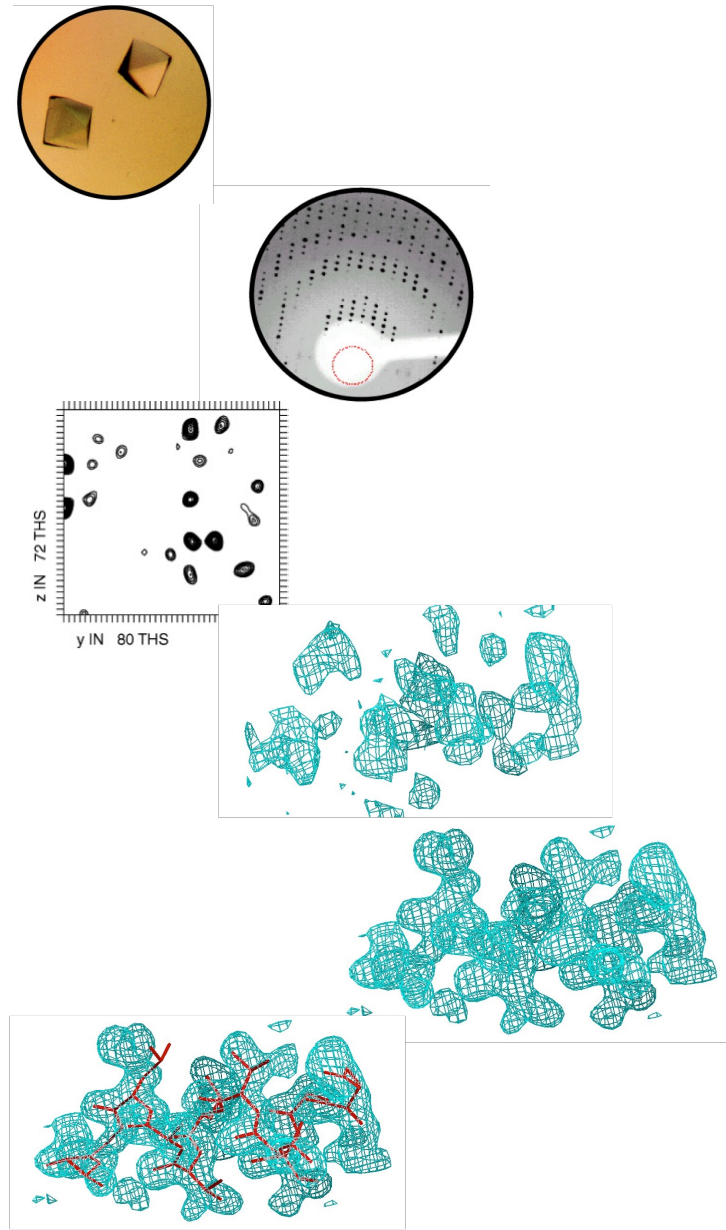
Automated model-building

Multi-step procedure:

- Locate helices and strands

- Extend helices and strands iteratively with tripeptides from libraries

- Assemble fragments into a poly-ala chain

- Build side chains and align them to the protein sequence

Finding regular protein structure

Extending with short fragments from PDB

Assembling best model

Identifying residue type at each position

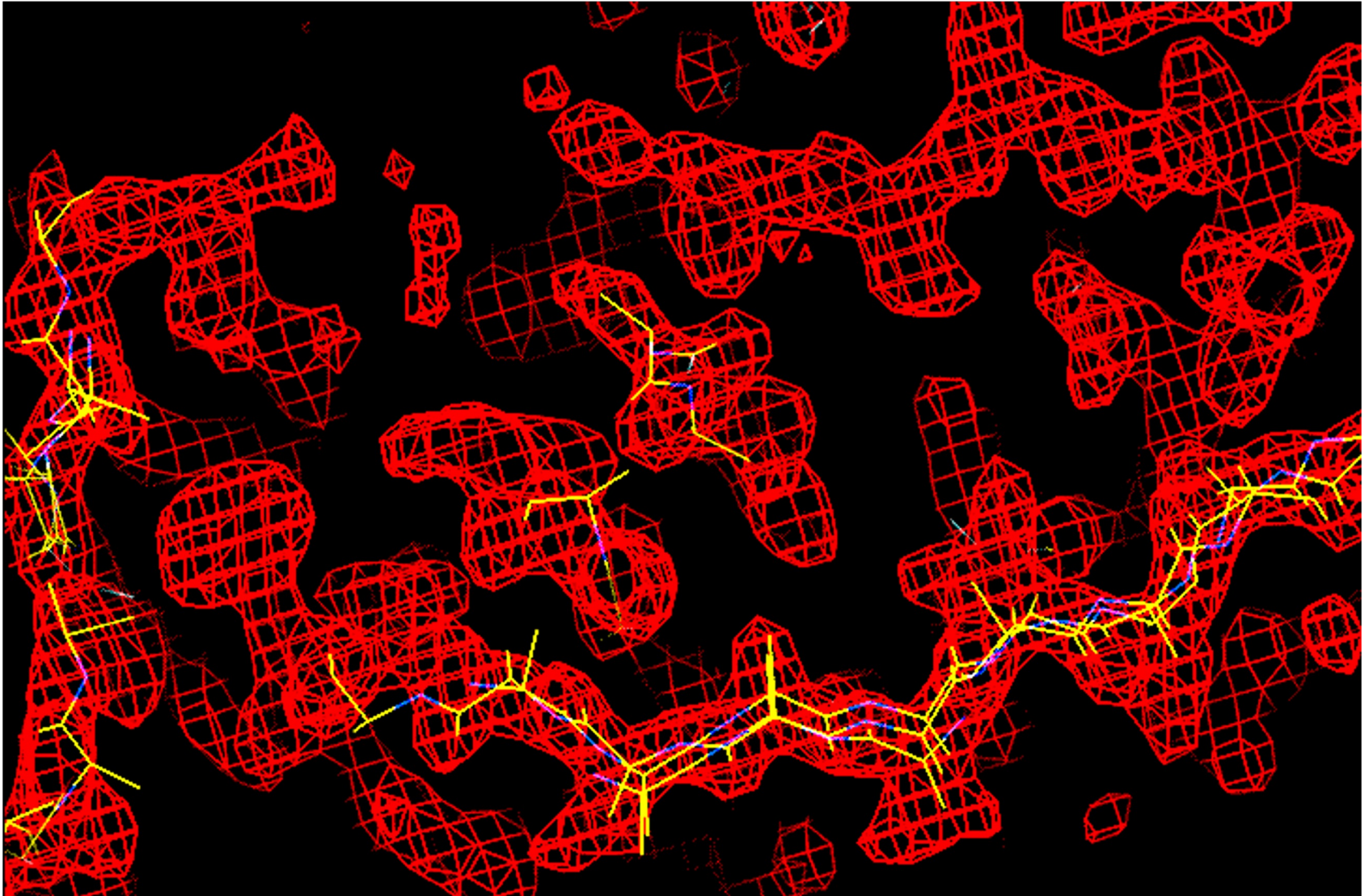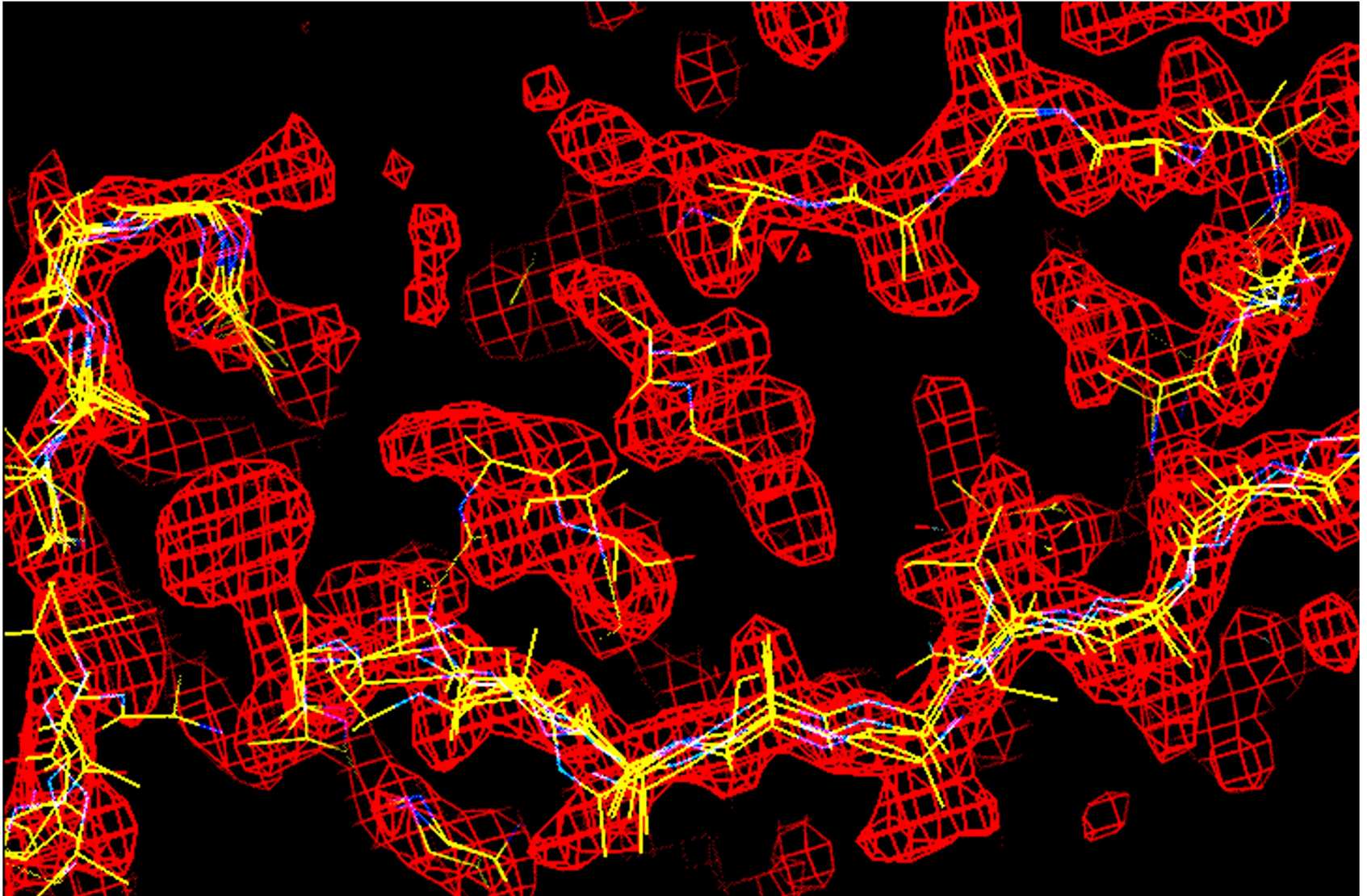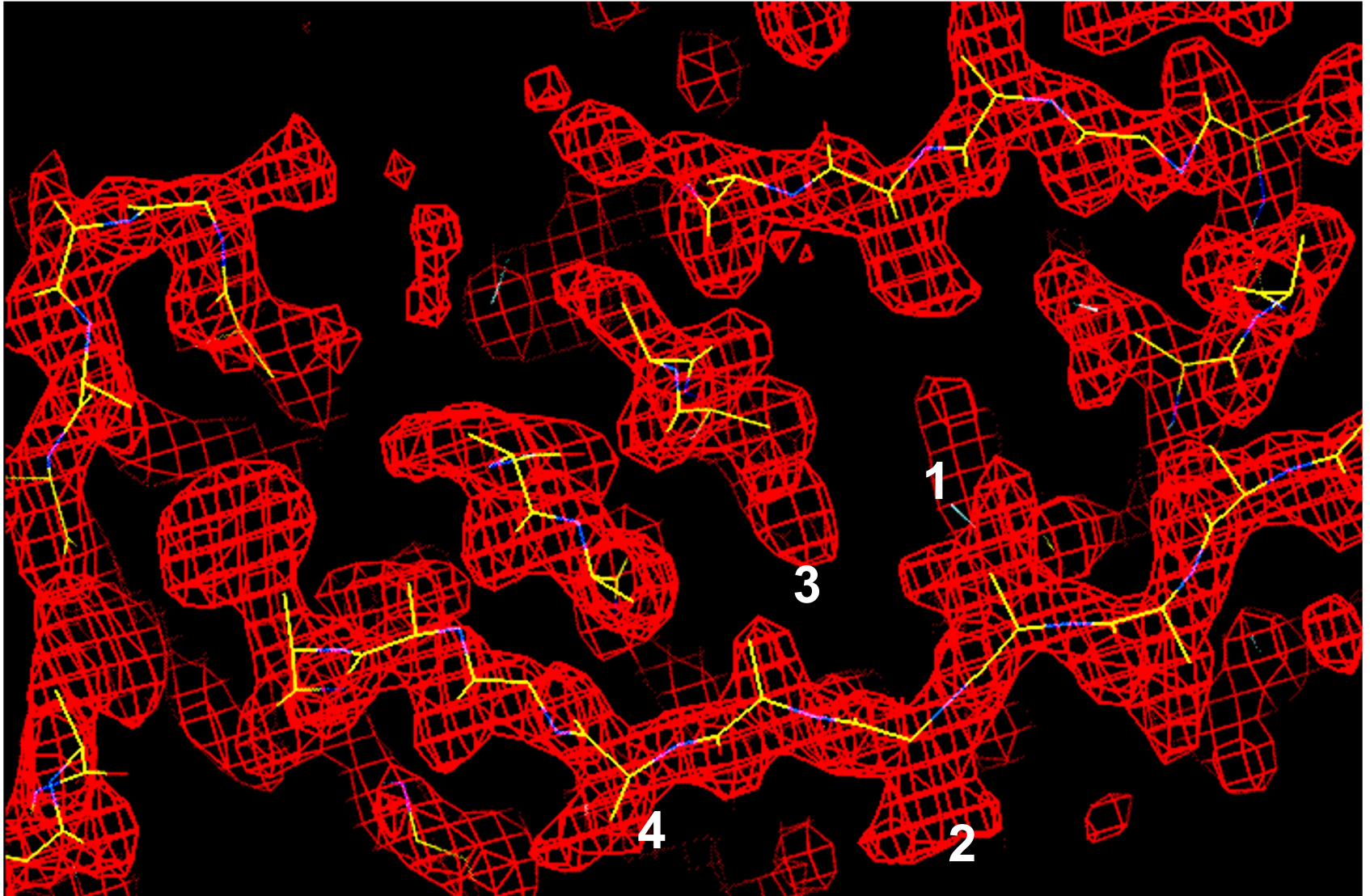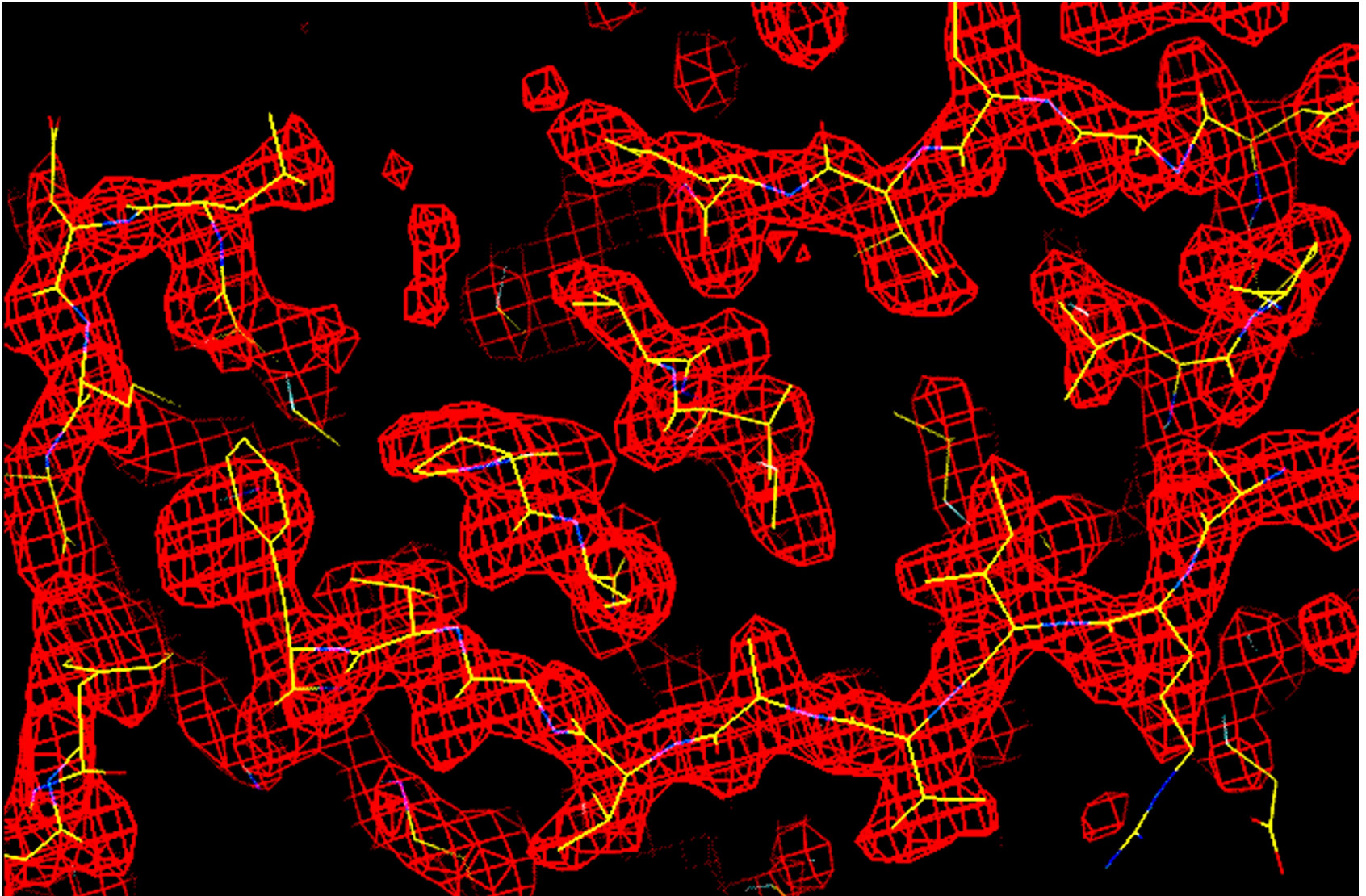| #	G	A	S	V	I	L	M	C	F	Y	K	R	W	H	E	D	Q	N	P	T
1	6	5	4	18	18	6	1	1	1	2	6	2	2	1	9	6	1	0	1	4
2	4	11	14	37	5	2	0	2	0	0	2	3	0	0	1	2	0	0	0	6
3	11	23	5	12	5	3	2	0	1	3	7	3	1	0	5	3	2	0	2	2
4	7	9	6	16	8	5	2	0	1	3	8	4	1	0	7	6	2	0	3	4
5	31	7	3	7	4	2	1	0	1	3	5	4	1	0	6	2	2	0	11	1
6	1	3	3	41	14	8	0	0	0	0	2	1	0	0	2	4	0	0	1	9
7	0	0	0	0	0	0	0	0	15	63	1	0	17	1	0	0	0	0	0	0
8	2	3	6	23	10	6	2	1	0	1	4	3	0	0	5	16	1	0	1	6
9	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Automated structure determination