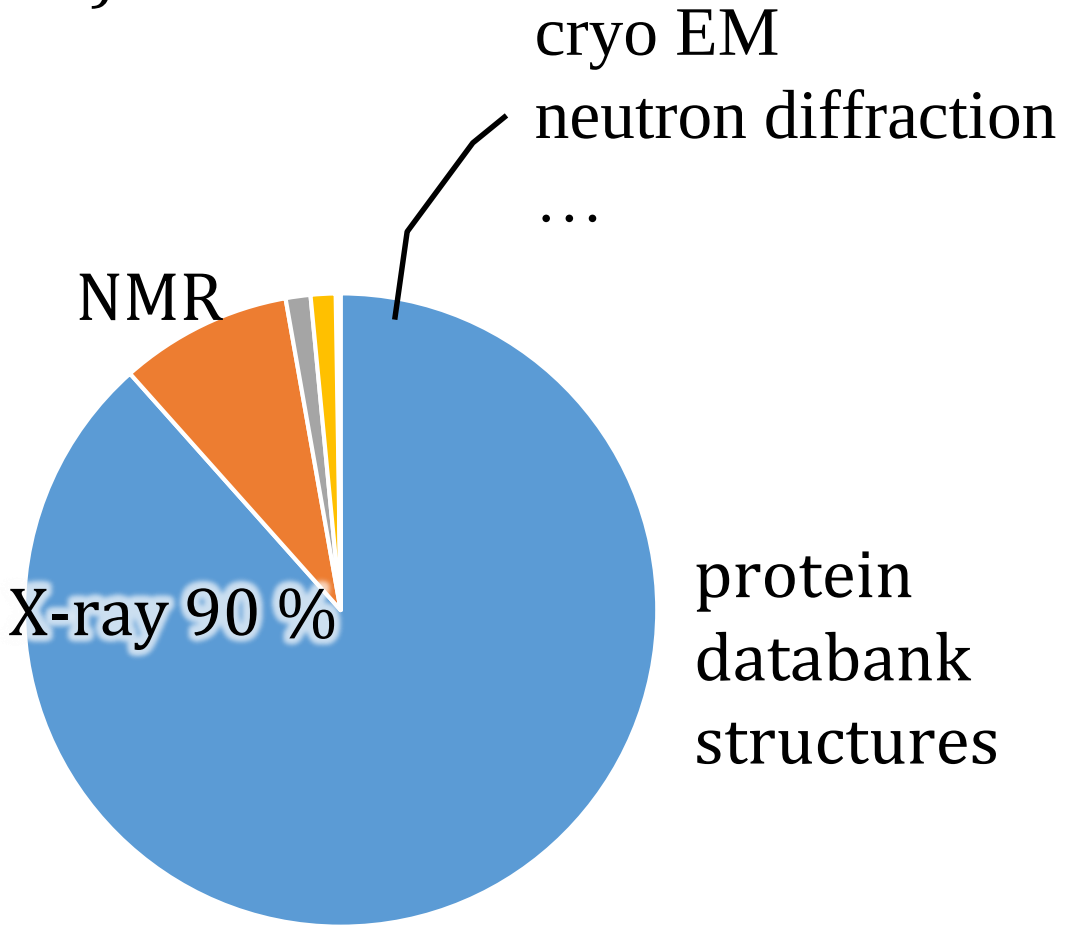
X-ray sociology

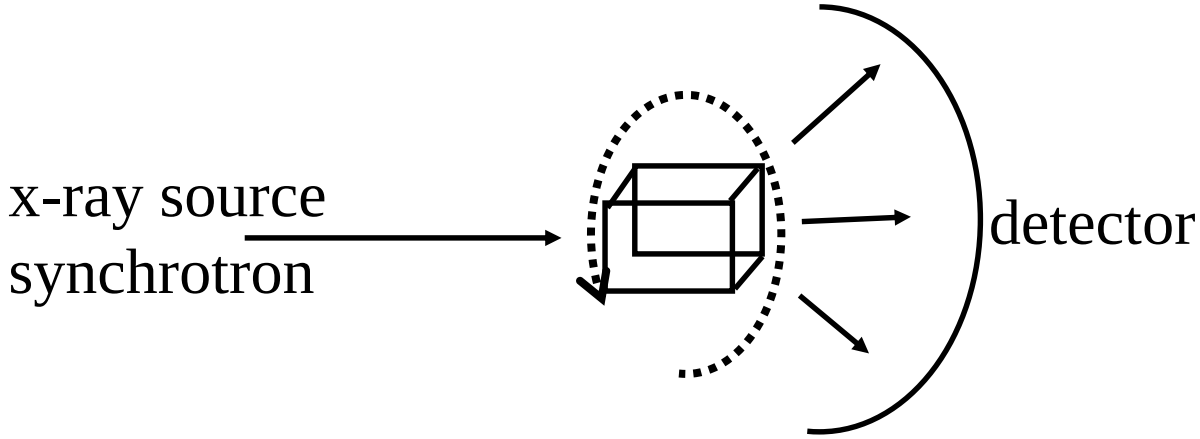
- more exact than NMR
- lots of Nobel prizes
- tells you where atoms are to a fraction (⅕ - ¼) of an Å
- can work on large structures

- Hamburg is full of crystallographers

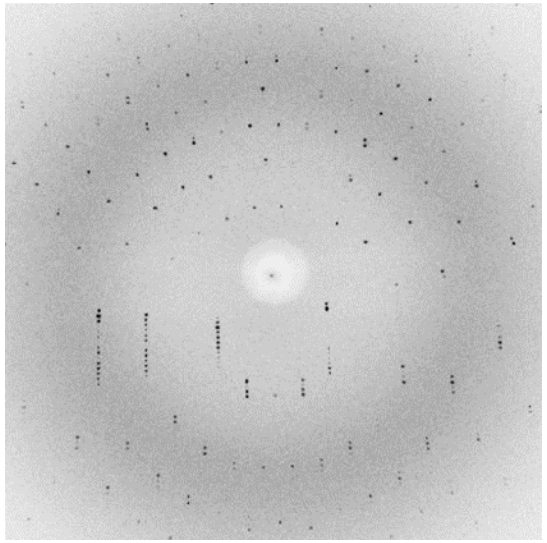
- not all proteins / nucleotides
  are happy to crystallise



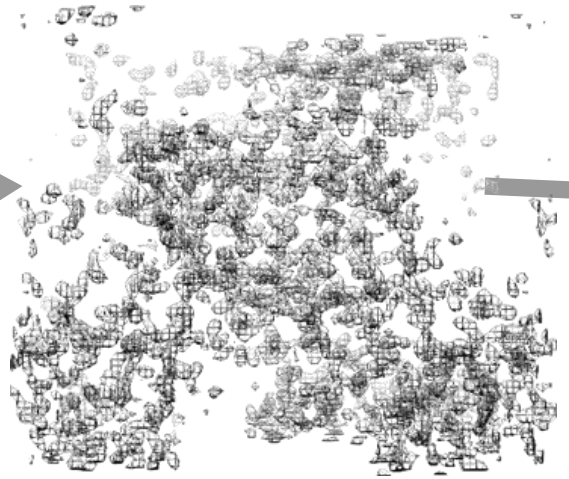
Summary of story



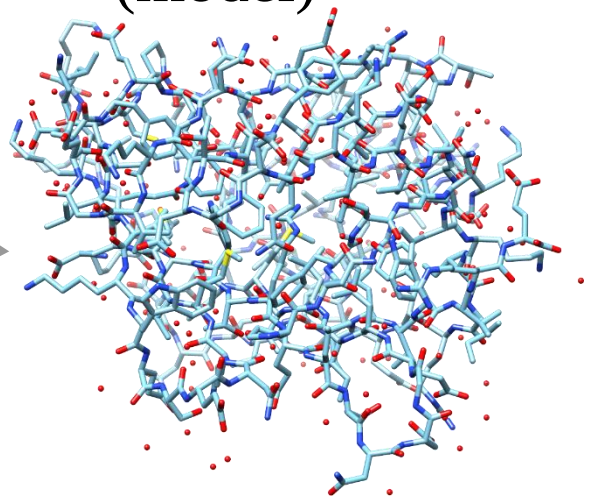
measured reflections



electron density



coordinates (model)



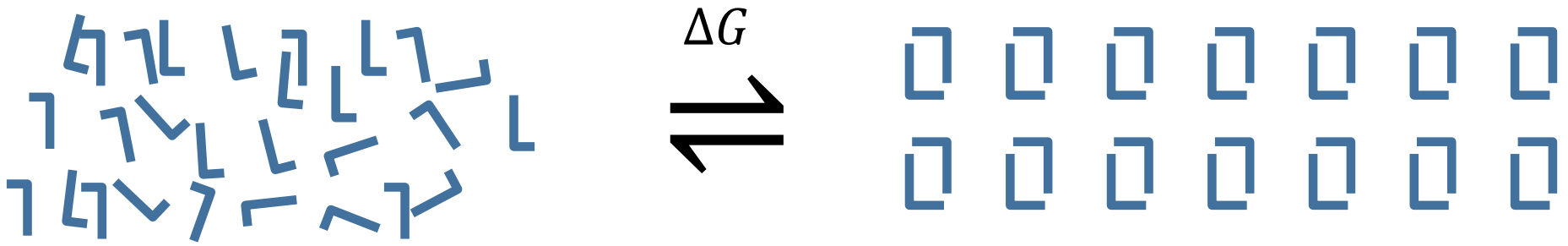
Topics

- tiny bit about chemistry / crystals
- calculating electron density
- placing coordinates in density

forming crystals

Familiar crystals – Kochsalz, sugar…

- crystal formation – small molecules
  - rigid, regular
  - soup of unordered molecules ⇌ ordered crystal, ΔG favourable



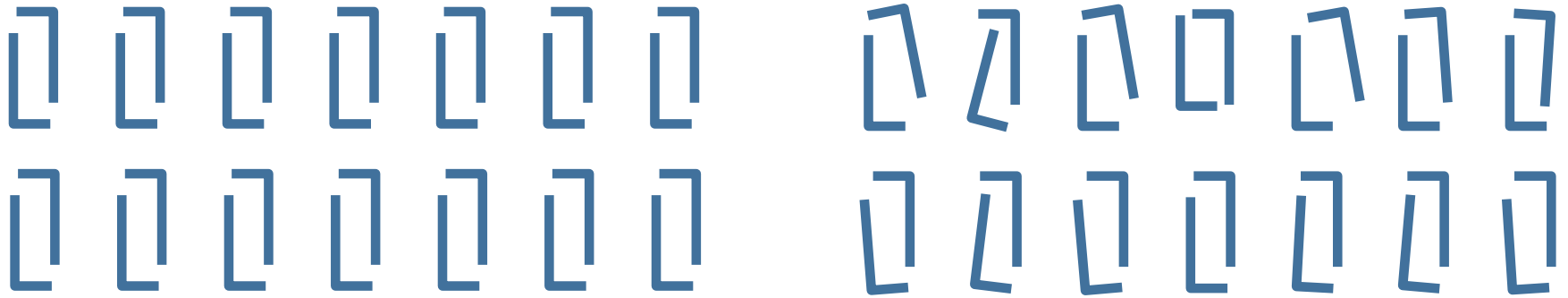
Proteins

- lots of internal mobility – not rigid
- not nice regular shapes
- soup molecules ⇌ ordered crystal, ΔG borderline

  Often can not be crystallised

Protein crystals

- small, 1 to 100 μm
- not as well ordered as small molecules

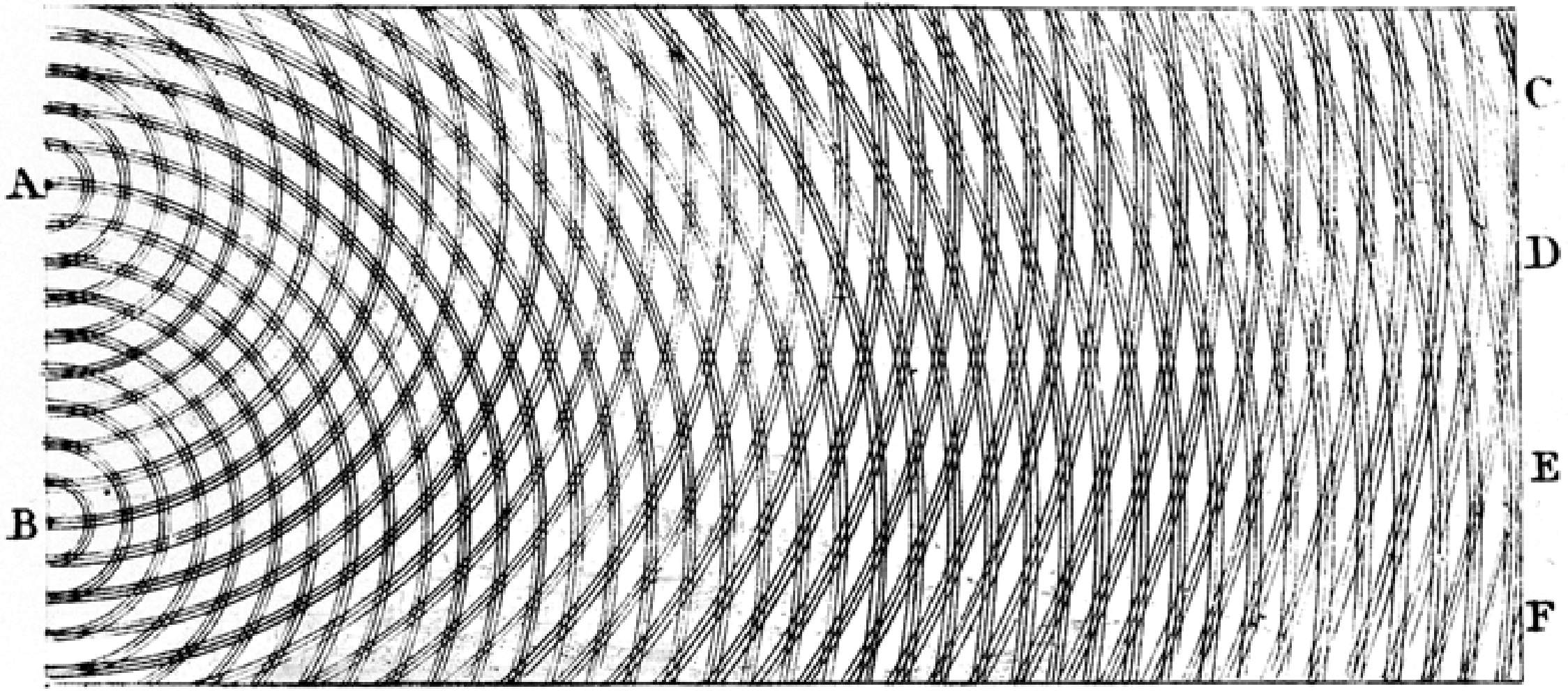


We measure an average over all molecules – consequence ..

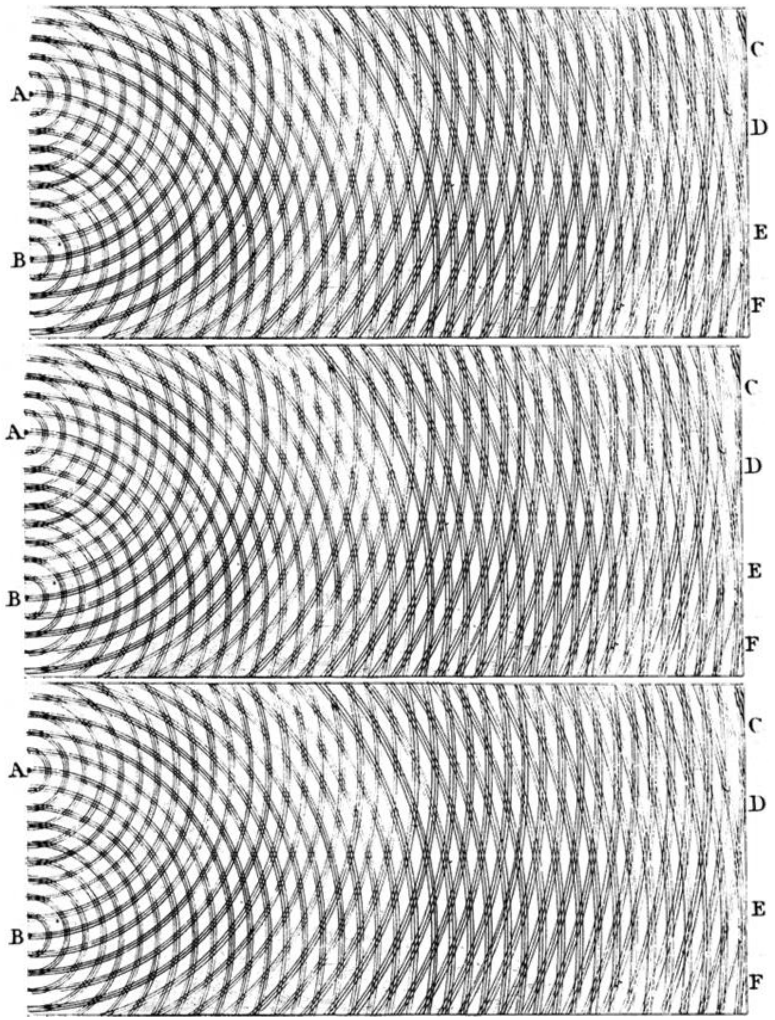
- right hand side – the average is smeared

Diffraction

Start with 1-dimension



Think of a grid



r_{AB}

waves
(light, sound, ..)

λ

a detector

x

Spacing of peaks on the detector ?
Will depend on

- distance r_{AB}
- wavelength λ

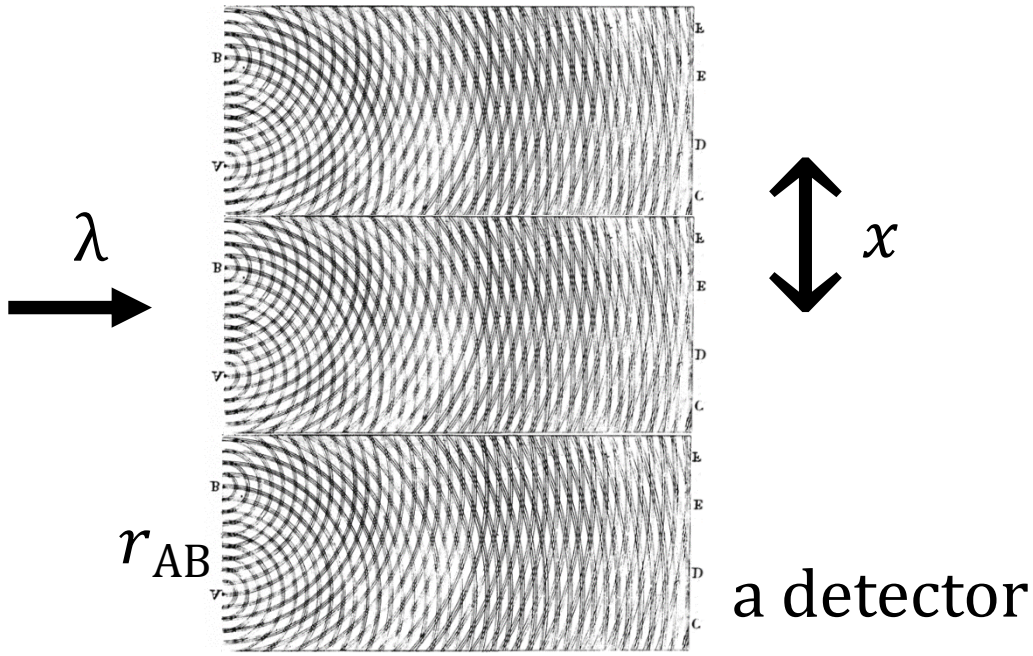
What we measure will be

- periodic (cosinus) in x
- change slowly if λ is big

$$I = \cos\left(\frac{2\pi}{\lambda}x + \alpha\right)$$

- α phase – important later

- relate this to proteins

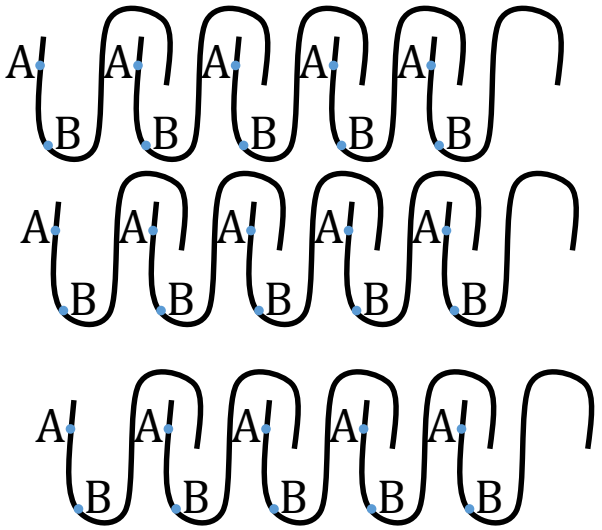


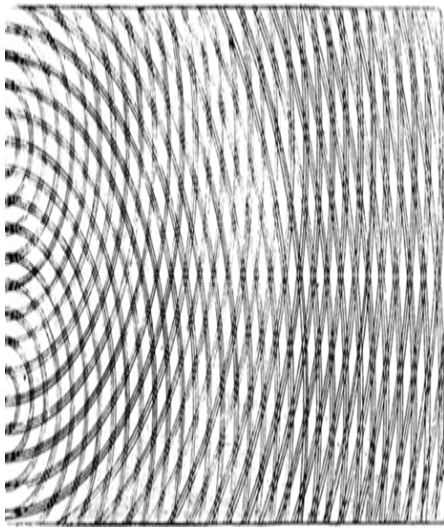
protein crystal and grids

- wavelengths are we talking about ? X-ray ≈ 1 Å
- do we have grids ?

Work in 3D

- makes maths more difficult
- have to rotate and add up pictures
- we have lots of pairs of atoms (AB)





X-rays interact with electrons

- hydrogen is almost invisible

What one measures

- sum over many pairs of electron clouds

Summing and Fourier transforms

Simple formula in 1D and one pair

$$I = \cos\left(\frac{2\pi}{\lambda}x + \alpha\right)$$

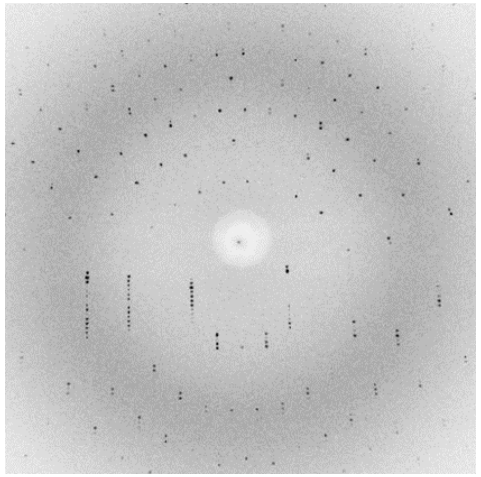
- what we measure is a sum $\sum_{reflections} \cos(\dots)$
- better nomenclature Σ_{hkl} - h, k, l are indices of detected spots

How do I break a signal into frequencies ?

- fourier transform – think of pictures of media player
  - from sound in real time to picture of frequencies $\cos(\omega)$

- Can you just take a signal and fourier transform ?
- only if you know the α 's
- There is phase for every reflection, but you cannot measure it

phase problem



- lots of "reflections" (10^5)
- you measure their size / intensity
- you cannot measure their phase α

- if you know the α for each reflection
  - just calculate density

Where do the phases (α) come from ?

Ways to find phases

Can you do it directly ?

- for very small molecules – yes
- protein ?
  - imagine you have 10^5 reflections
  - try just 4 possibilities in each case
  - 4^{10000} possibilities

Strategies – two most important

1. if you know some coordinates, substitute them into the equation – gets a good starting point (molecular replacement)
2. if you know some phases, easy to get the next ones (multiple isomorphous replacement)

+ if you have a reasonable initial guess, it can be optimised

molecular replacement

- most common – ¾ or more of structures in protein data bank
- you need to know some coordinates

- you do not know coordinates
  - you have the coordinates of something close (homologue)

- pretend these coordinates are responsible for the measurements
- substitute into formula and get an initial set of α 's
- refine

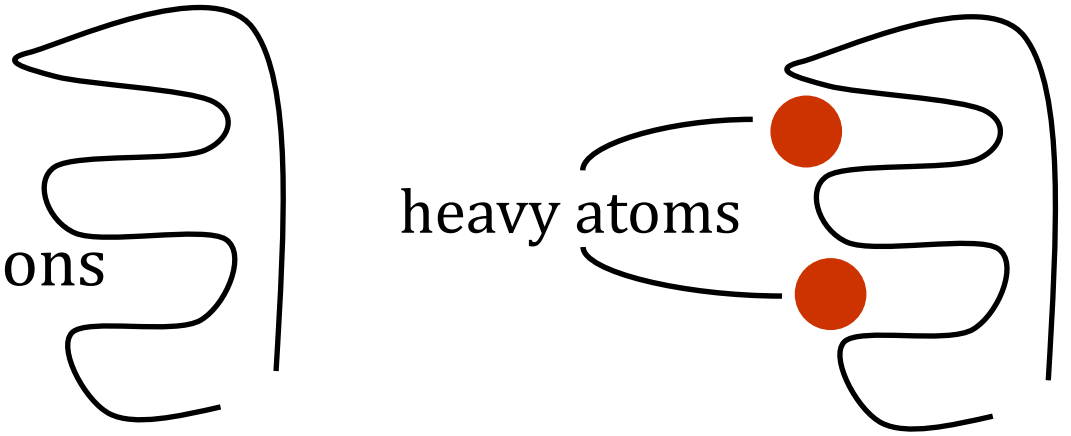
multiple isomorphous replacement (MIR)

Remember

- for small molecules you can find phases directly (few reflections)
- if you know some phases, the next are easier

How to treat protein as if is a small molecule ?

- Heavy atoms have so many electrons they dominate the observations
- give a starting point for phases



- what are heavy atoms ? Au, Pt, Hg, Br, Se, Xe

- should bind at the same position in every protein molecule

more phasing

How difficult ?

- for many proteins – plenty of data but phases take years

Other methods ?

- yes

resolution - meaning

- most common statistic for X-ray structures – resolution
- meaning - when do two points look like one ?

- Resolution in X-ray.. depends on how scattered the waves are

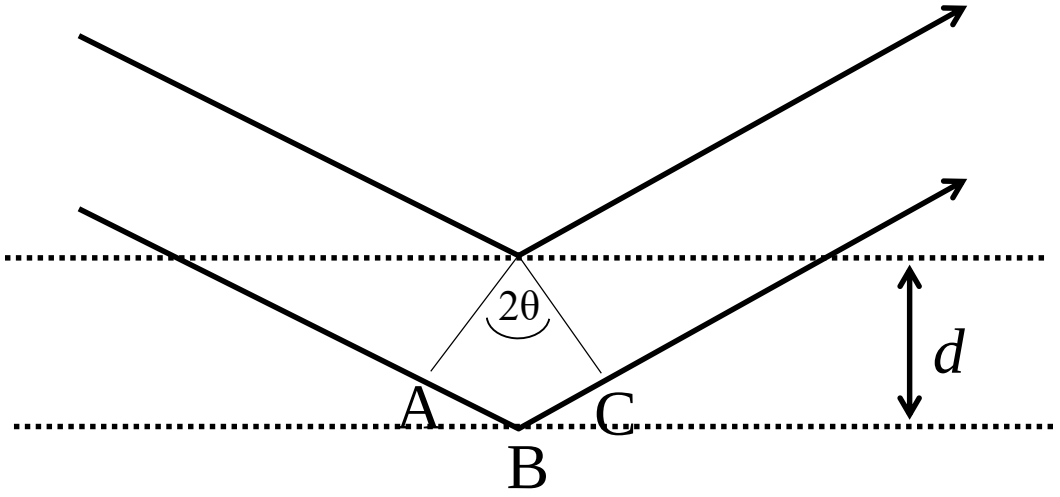
planes in a crystal

waves and resolution

- bottom path = lower path + wavelength λ then they reinforce each other
- difference in lengths is $n\ \lambda$

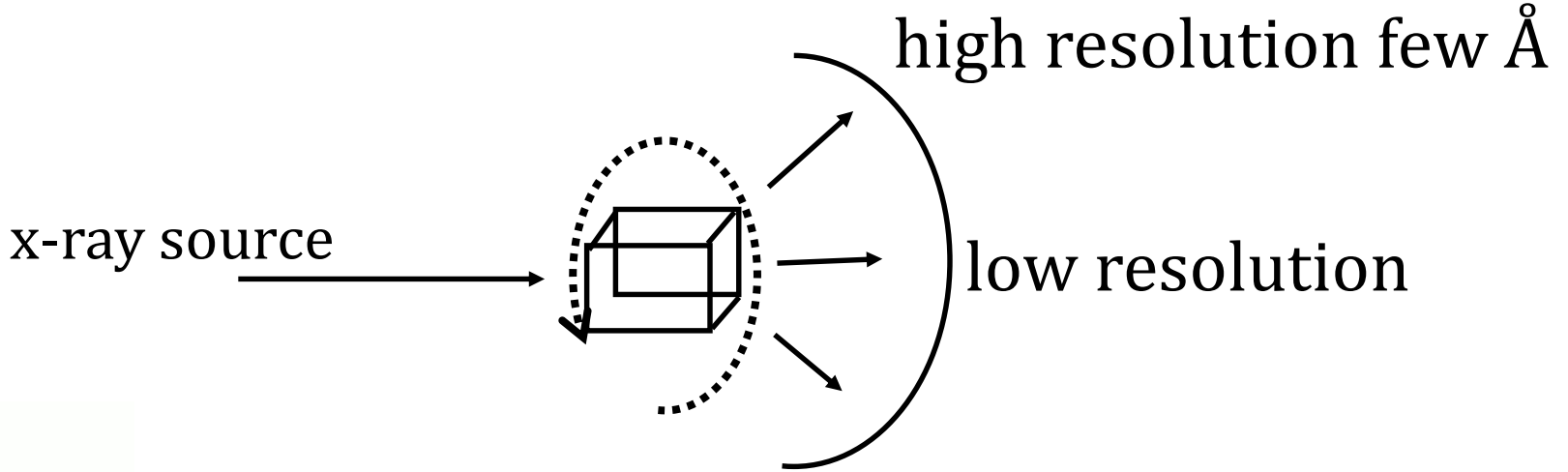
$$n\lambda = \overline{\mathrm{ABC}} = 2(d \sin\theta)$$

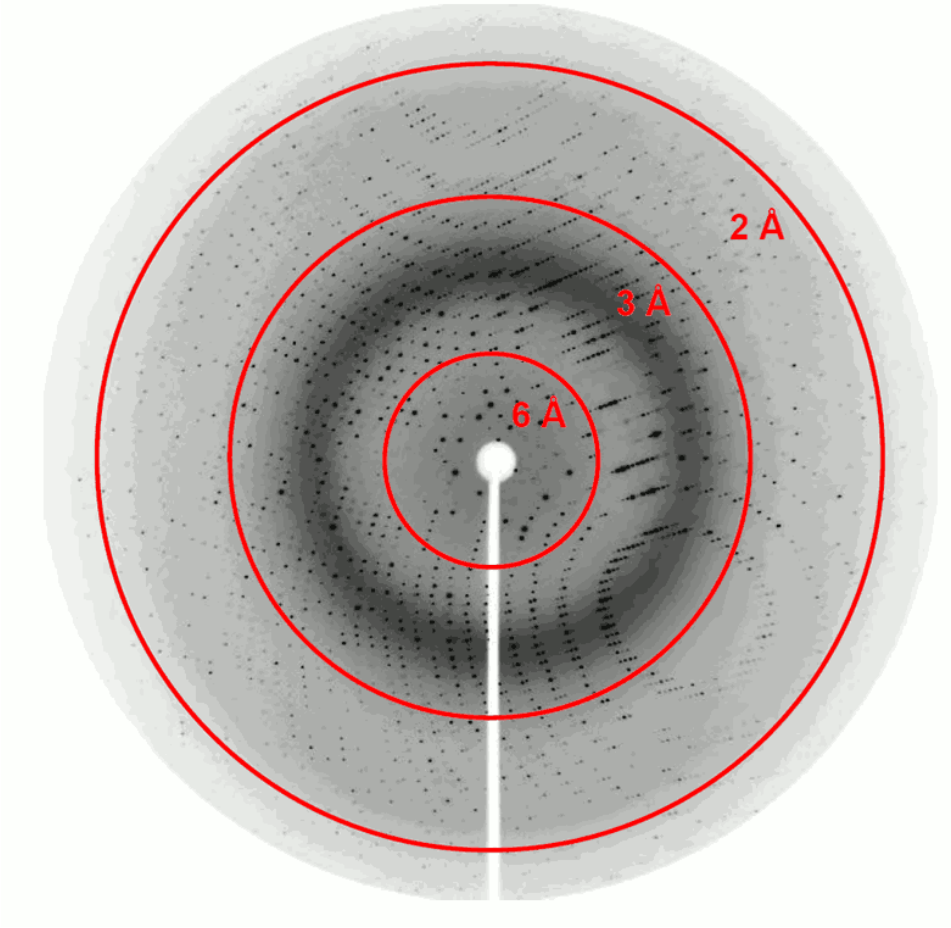
- then $d = \frac{n\lambda}{2 \sin\theta}$



Resolution consequences

- smaller wavelength λ, the better
- angle θ – you cannot do much but..





In the centre .. reflections

- did not diffract much
- due to low resolution information
- strong

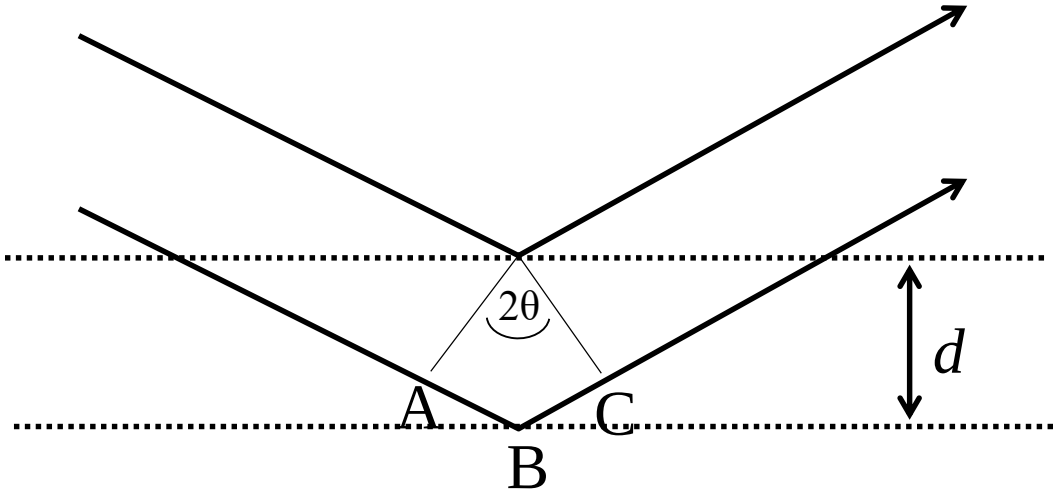
Further out

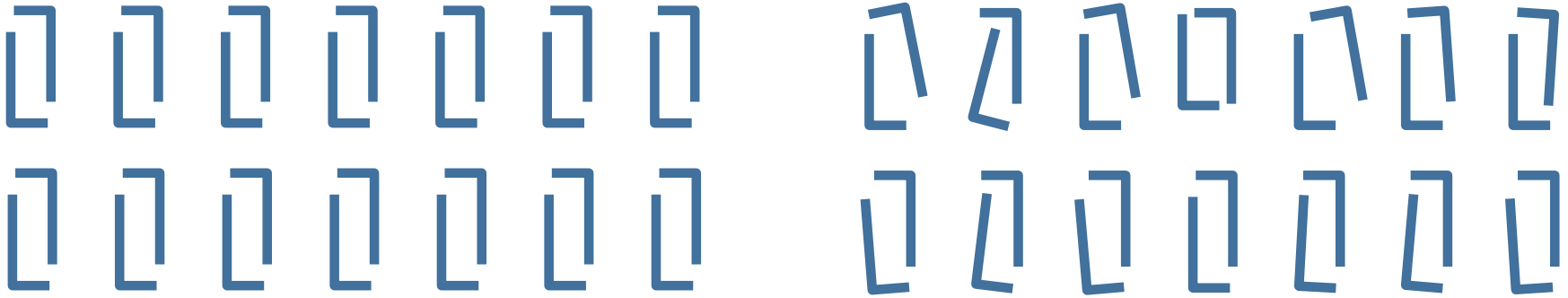
- high resolution information
- signal weak

resolution practical

Mostly a function of crystal

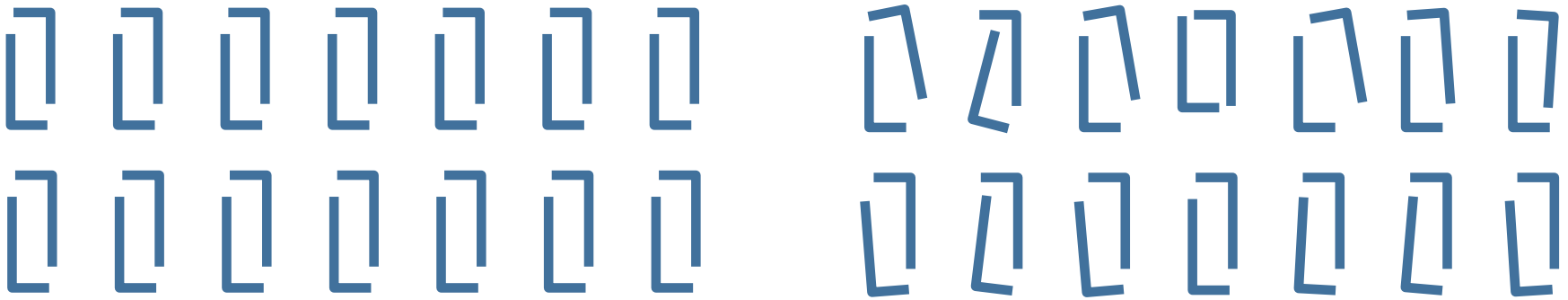
- crystal not so regular and / or
- atoms mobile
  - you are seeing an average
  - there is no high resolution information



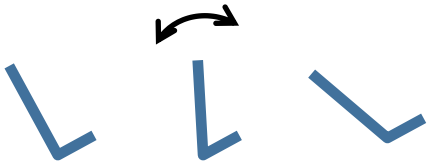


disorder – static and dynamic

static



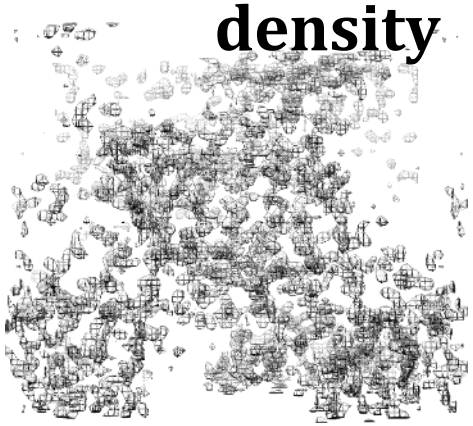
dynamic



model for this - soon

model fitting - refinement

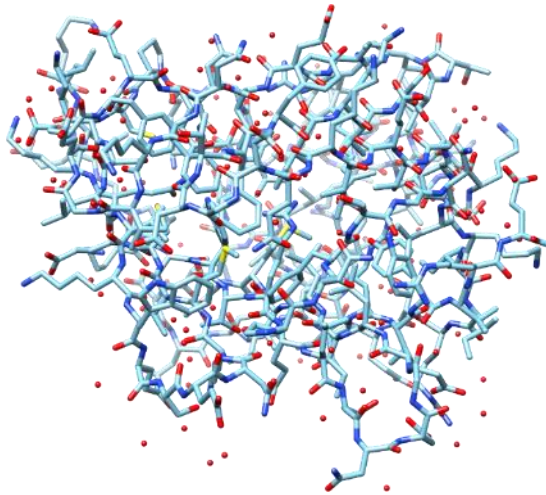
electron density



- you have initial electron density and initial phases
- have to fit atoms (build a model)

- what are the variables
  - x, y, z for atoms
  - B-factors (mobility) … next slide

- Given atoms, you can calculate density
- given density, you can calculate reflections (structure factors F)

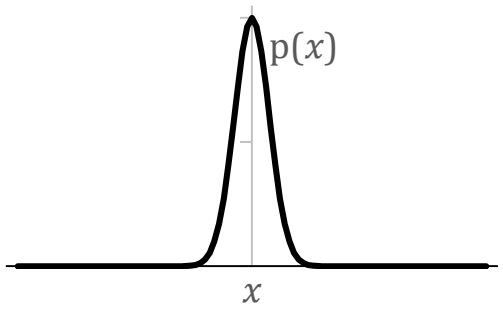


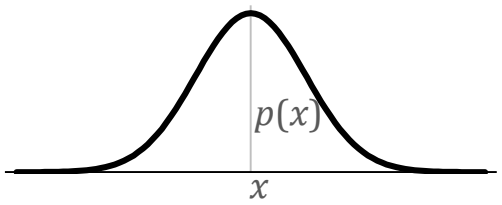
Refinement
- how well do the structure factors from your model agree with measurements ?

B-factors

A model for atom location

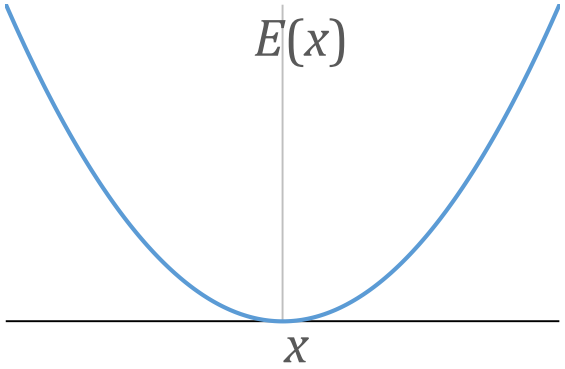
- Gaussian (normal) probability



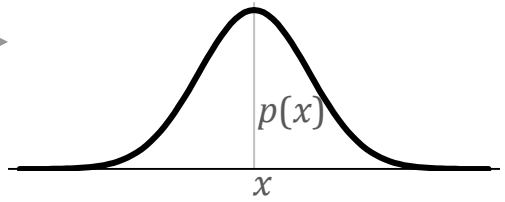


Why Gaussian ?

- pretend a particle moves in a harmonic well $E(x) = x^2$
- from Boltzmann relation, $p(x) \propto \exp\left(\frac{-x^2}{kT}\right)$ (formal in Sommersemester)
  - k Boltzmann constant, T temperature



energy probability



B-factors interpretation

Probability distribution

- how likely are you see an atom at a position or
- what is the typical movement at room temperature ?

$B = 8\,\pi^2 u^2$ where u are fluctuations in Å

$$u = \left(\frac{B}{8\pi^2}\right)^{\frac{1}{2}}$$

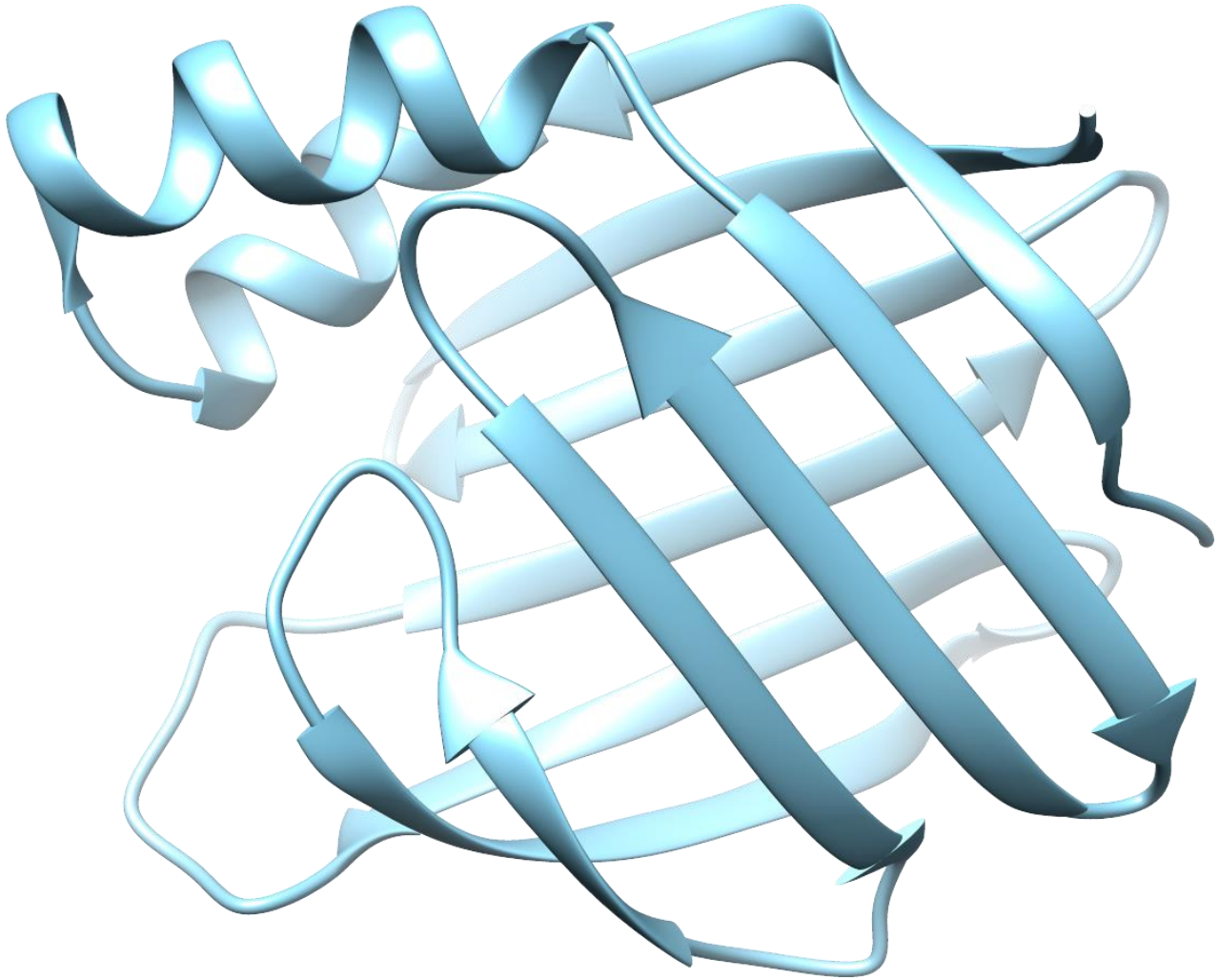
if $B = 50$ Å^2, typical displacement ≈ 0.8 Å

if $B = 20$Å^2, typical displacement ≈ 0.5 Å

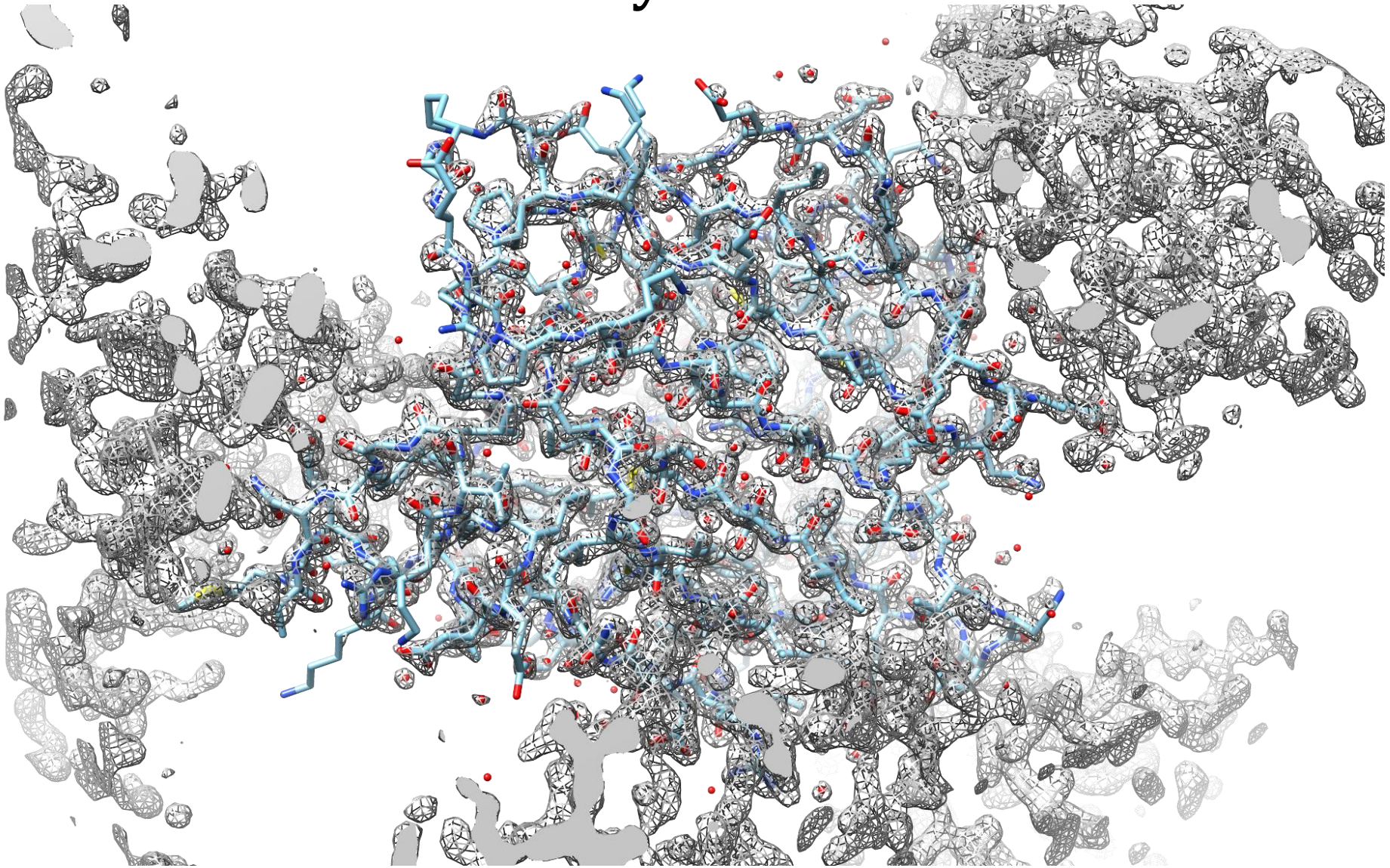
The connection to fitting…

fitting to density

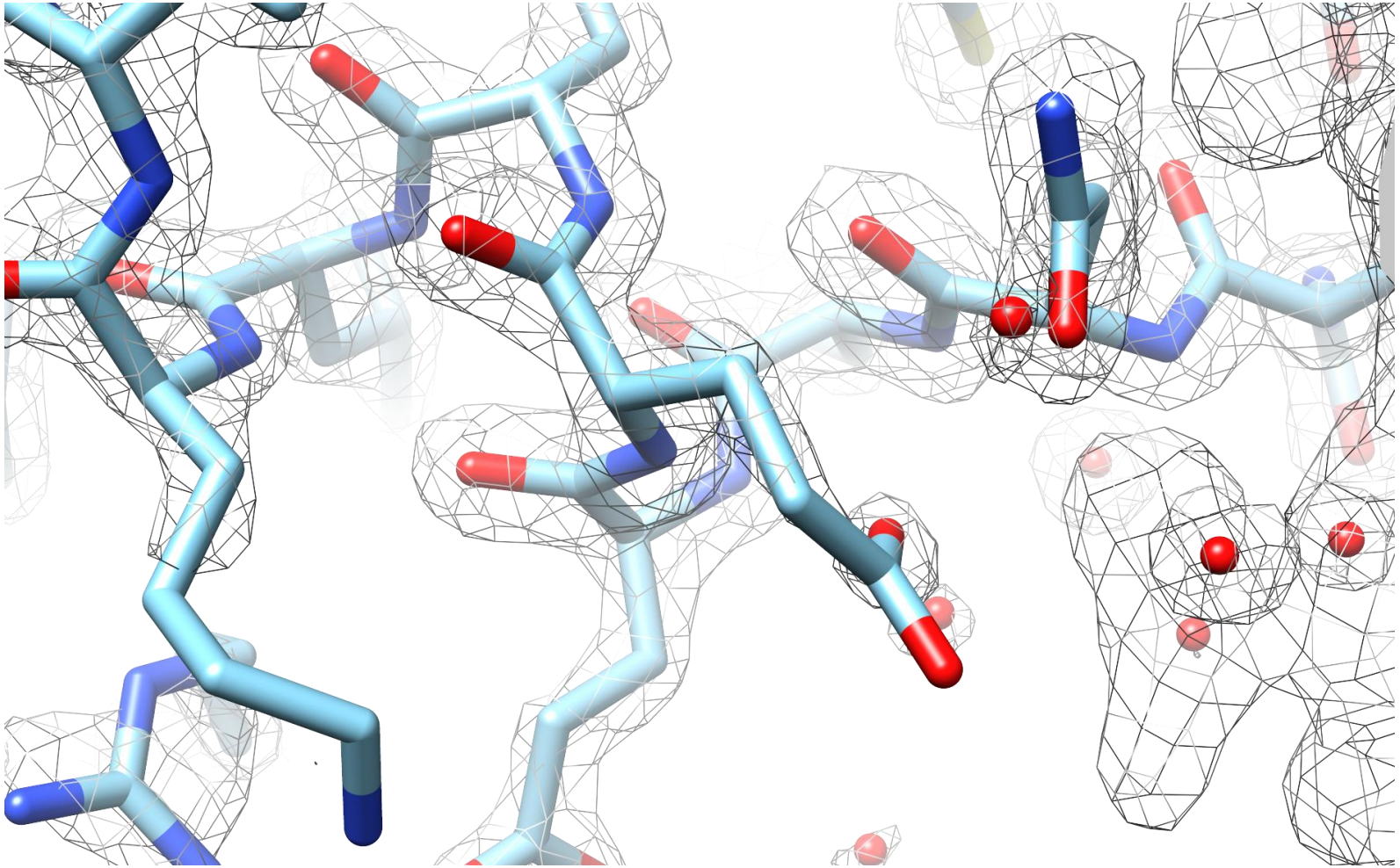
The path to this picture…



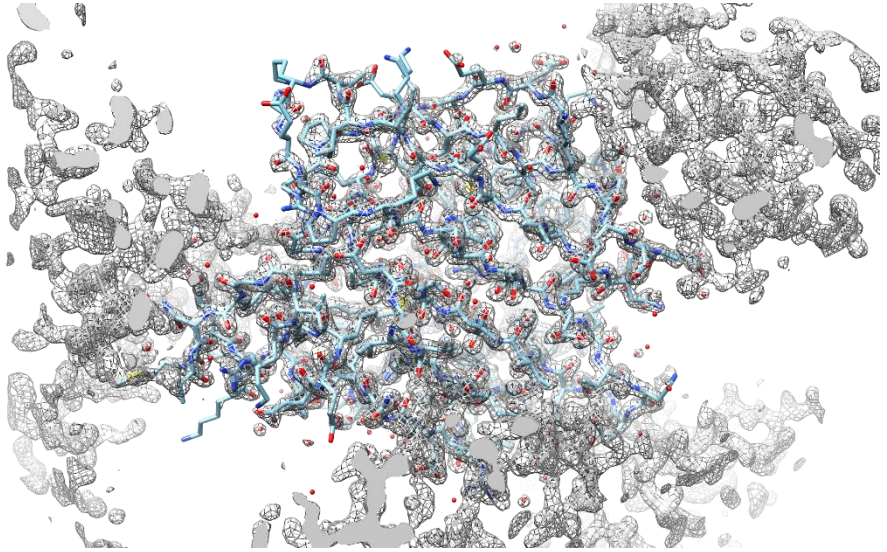
density



- water molecules
- atoms have different
  - sizes
    electron clouds
  - mobilities



Refinement – cost function – R



The cost function

- atoms → density → structure factors (F)
  - F_{hkl}^{calc} structure factors calculated
  - F_{hkl}^{obs} structure factors observed

very important… R factor

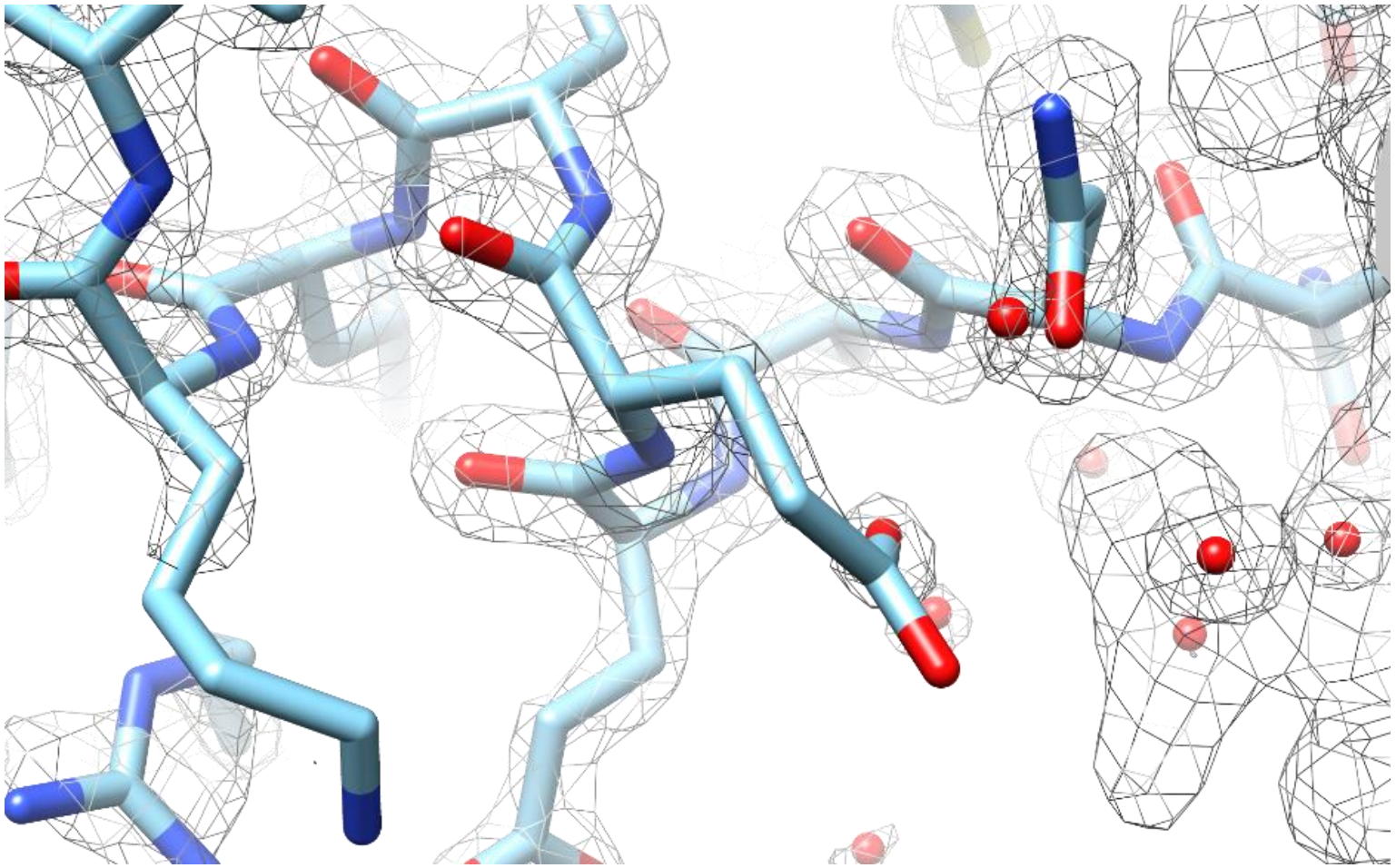
$$R = \frac{\sum_{hkl}\left|F_{hkl}^{obs} - F_{hkl}^{calc}\right|}{\sum_{hkl}\left|F_{hkl}^{obs}\right|}$$

Variables in refinement

What are the variables ?

- x, y, z for every atom +
- B-factors

- adding ions and water to fill density



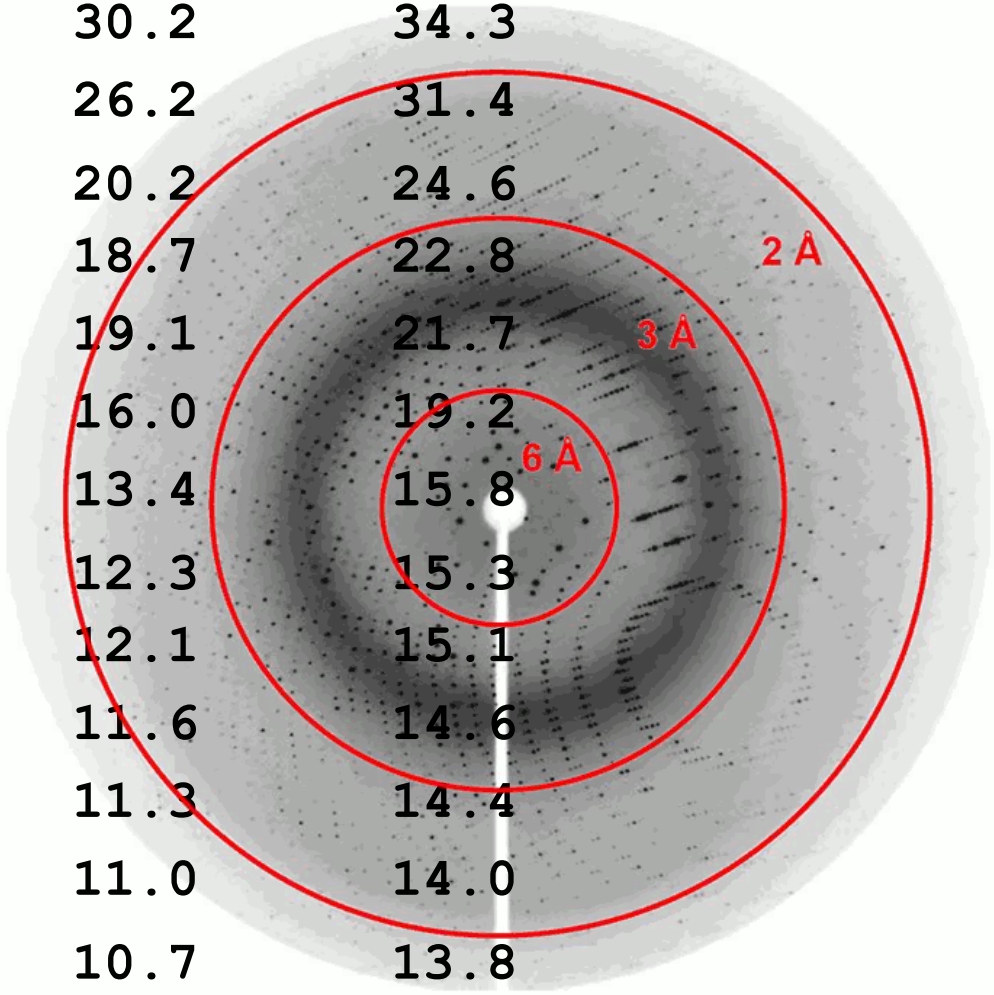
refinement progress

	data used	# par	# obs	Rw	Rf
molecular replacement	15-4.0	3	775	44.5	
rigid body	10-2.5	9	2997	46.4	47.4
first round	10-1.5	3887	13818	30.2	34.3
after first		3735		26.2	31.4
SHELXWAT		4091		20.2	24.6
bld + SHELXWAT		4231		18.7	22.8
include all data	10-1.1	4203	33993	19.1	21.7
ANIS 20		9453		16.0	19.2
Rebuild SHELXWAT		9557		13.4	15.8
rebuild		10481		12.3	15.3
rebuild		10819		12.1	15.1
rebuild		10838		11.6	14.6
rebuild		11494		11.3	14.4
rebuild		11576		11.0	14.0
rebuild		11774		10.7	13.8

- data used
- #par number of parameters in model
- #obs number of reflections
- Rw R factor
- Rf R_{free} (soon)

refinement progress

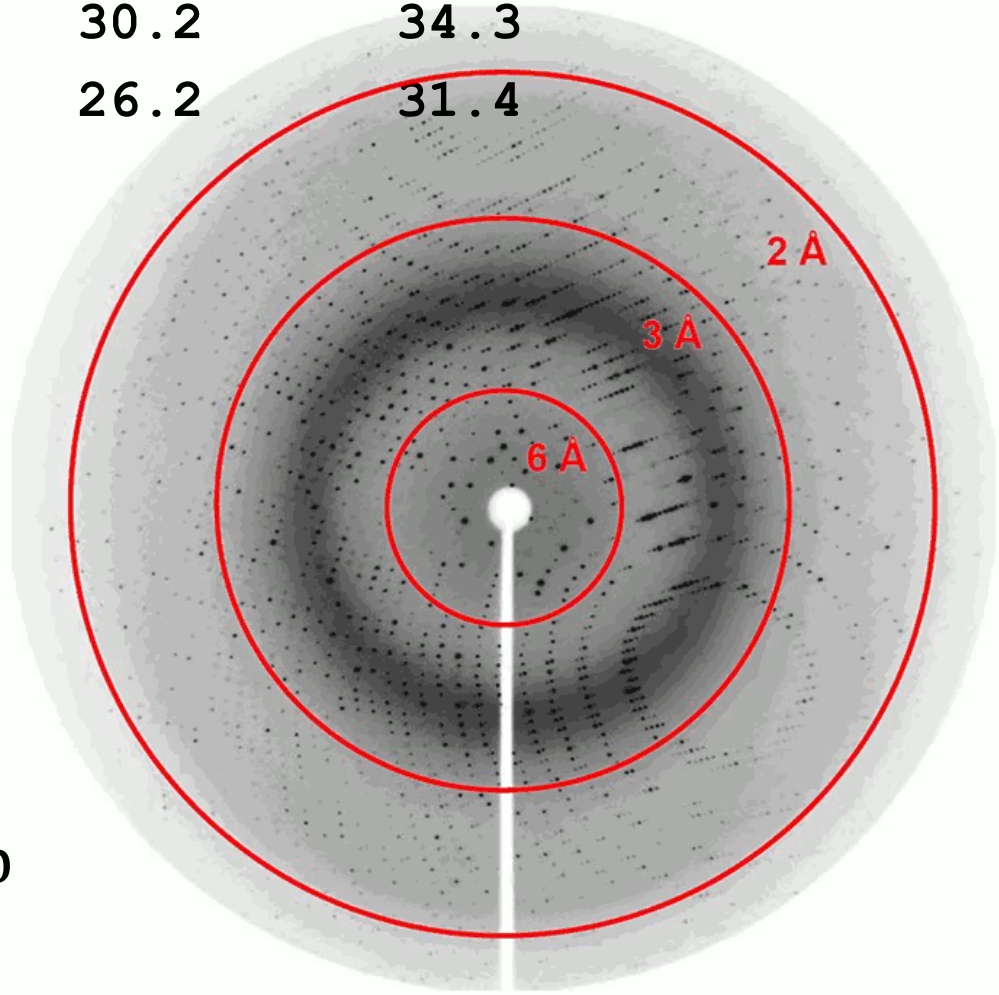
	data used	# par	# obs	Rw	Rf
molecular replacement	15-4.0	3	775	44.5	
rigid body	10-2.5	9	2997	46.4	47.4
first round	10-1.5	3887	13818	30.2	34.3
after first		3735		26.2	31.4
SHELXWAT		4091		20.2	24.6
bld + SHELXWAT		4231		18.7	22.8
include all data	10-1.1	4203	33993	19.1	21.7
ANIS 20		9453		16.0	19.2
Rebuild SHELXWAT		9557		13.4	15.8
rebuild		10481		12.3	15.3
rebuild		10819		12.1	15.1
rebuild		10838		11.6	14.6
rebuild		11494		11.3	14.4
rebuild		11576		11.0	14.0
rebuild		11774		10.7	13.8



refinement progress

	data used	# par	# obs	Rw	Rf
molecular replacement rigid body	15-4.0	3	775	44.5	
	10-2.5	9	2997	46.4	47.4
first round	10-1.5	3887	13818	30.2	34.3
after first		3735		26.2	31.4

- Start: Low resolution data is enough – just a few parameters
- #par and #obs
  - where is the molecule ?
  - where is molecule + phases ?
  - add in first atoms
    - number of parameters grows at each step
    - add in more data (#obs)



refinement progress

	data used	# par	# obs	Rw	Rf
molecular replacement	15-4.0	3	775	44.5	
rigid body	10-2.5	9	2997	46.4	47.4
first round	10-1.5	3887	13818	30.2	34.3
after first		3735		26.2	31.4
SHELXWAT		4091		20.2	24.6
bld + SHELXWAT		4231		18.7	22.8
include all data	10-1.1	4203	33993	19.1	21.7
ANIS 20		9453		16.0	19.2
Rebuild SHELXWAT		9557		13.4	15.8
rebuild		10481		12.3	
rebuild		10819		12.1	
rebuild		10838		11.6	14.6
rebuild		11494		11.3	14.4
rebuild		11576		11.0	14.0
rebuild		11774		10.7	13.8

R of bit more than 20 % is typical

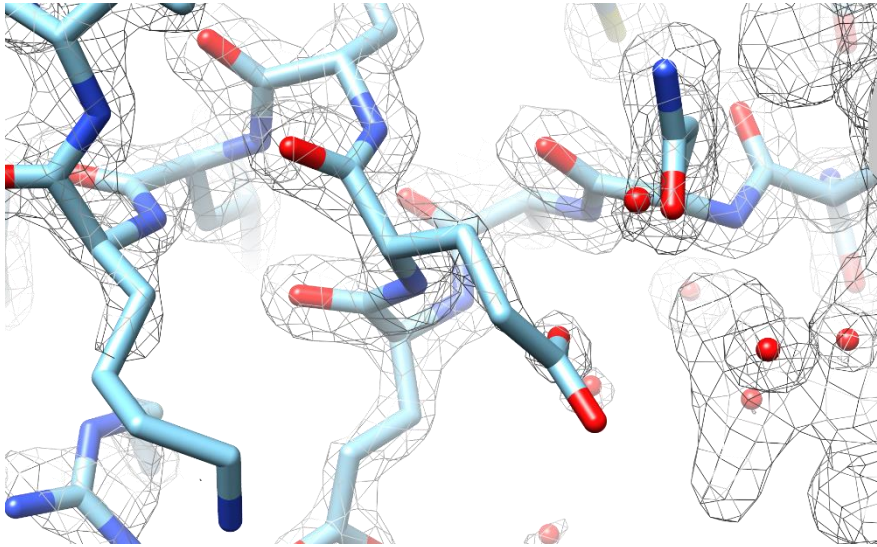
What is happening in refinement

Similar problem to NMR refinement

- use a minimizer
  - move x, y, z and B-factors
    - until agree with experimental data (F_{hkl}^{obs})
    - maintain known chemistry – bond lengths angles

Different to NMR

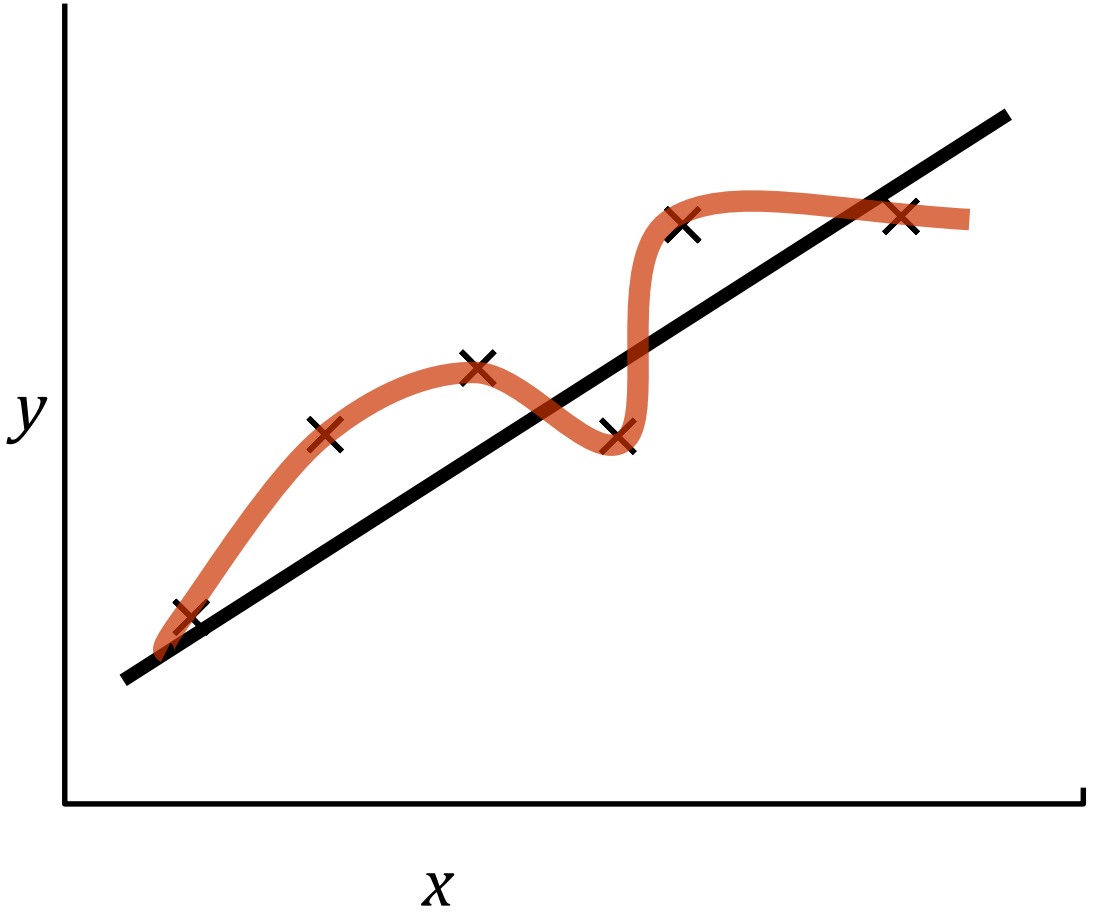
- more variables
  - B-factors
  - you can add water and ions to fill density
- more data
- search for a single solution (not many possibilities)



R free - overfitting

Overfitting ?

- you give me data
- true model is a line ($y = ax + m$)
  - two parameters
- I fit to a polynomial
  $y = ax^4 + bx^3 + cx^2 + dx + m$
- apparently better fit
  - will not predict correctly



How would you detect this ?

- I have half a dozen points
- do fitting on five – see how good the fit on the sixth point is

In general – fit on 90 % the data and test on remaining 10 %

refinement progress

	data used	# par	# obs	Rw	Rf
molecular replacement	15-4.0	3	775	44.5	
rigid body	10-2.5	9	2997	46.4	47.4
first round	10-1.5	3887	13818	30.2	34.3
after first		3735		26.2	31.4
SHELXWAT		4091		20.2	24.6
bld + SHELXWAT		4231		18.7	22.8
include all data	10-1.1	4203	33993	19.1	21.7
ANIS 20		9453		16.0	19.2
Rebuild SHELXWAT		9557		13.4	15.8
rebuild		10481		12.3	15.3
rebuild		10819		12.1	15.1
rebuild		10838		11.6	14.6
rebuild		11494		11.3	14.4
rebuild		11576		11.0	14.0
rebuild		11774		10.7	13.8

R_{free}

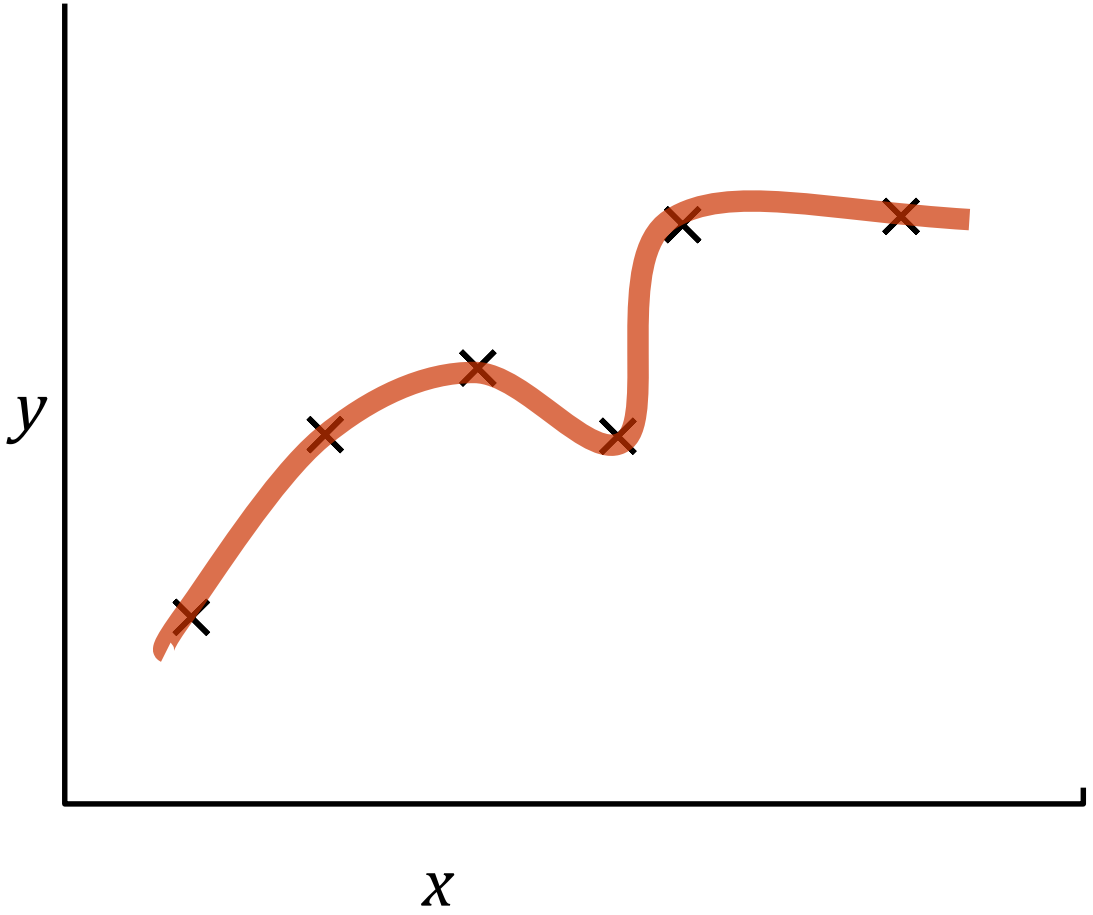
Crystal data – also has noise

- you can always fit to noise
  - add water, ions, move atoms

How to detect ?

- fit on 90 % of data
- calculate R on remaining data

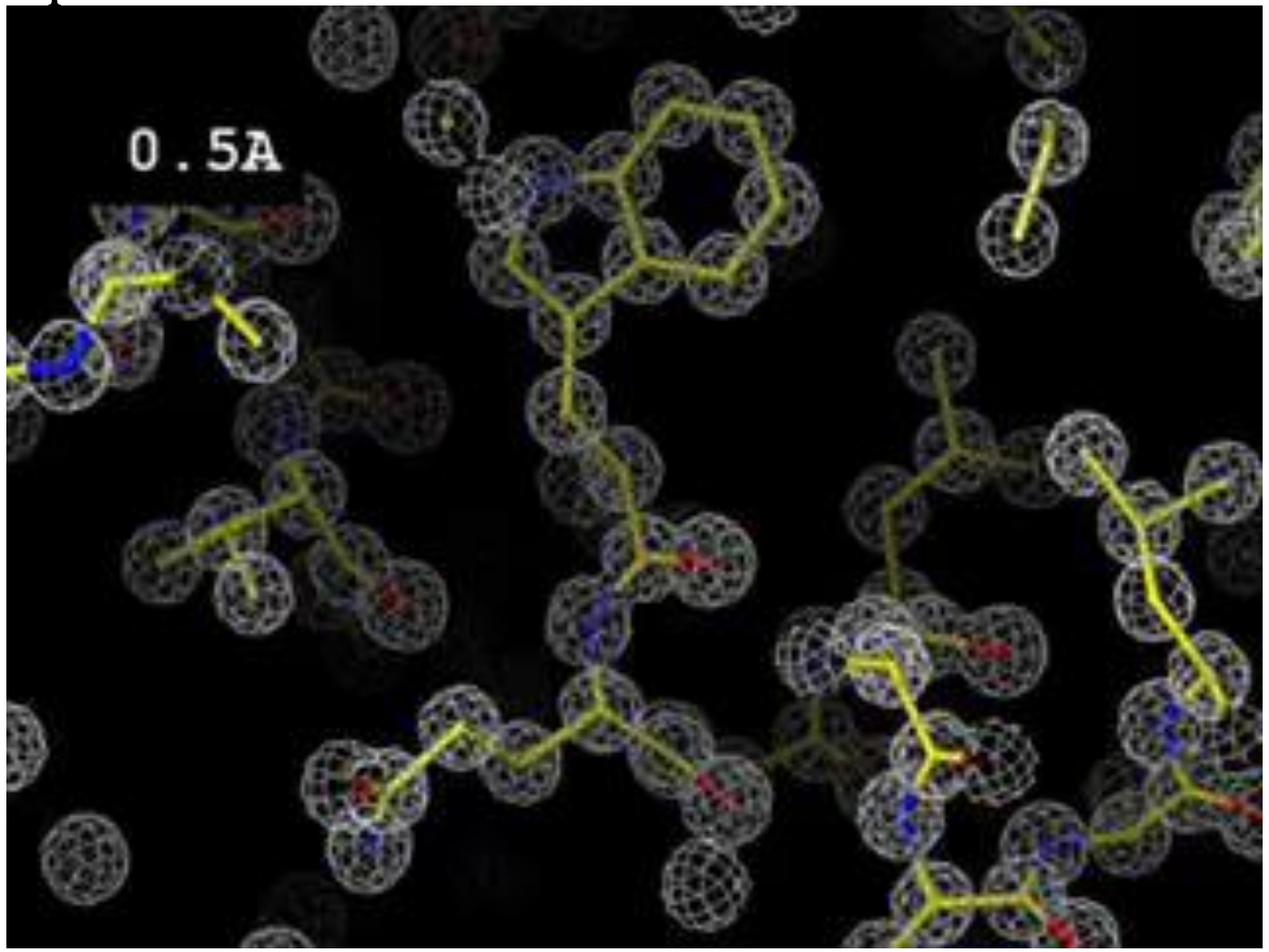
$$R_{free} = \frac{\sum_{hkl} \left| F_{hkl}^{obs} - F_{hkl}^{calc} \right|}{\sum_{hkl} \left| F_{hkl}^{obs} \right|}$$



- but F_{hkl}^{obs} is from 10% of data not used in fitting
- typical values in protein data bank – 20 to 30 %

Practical meaning of resolution

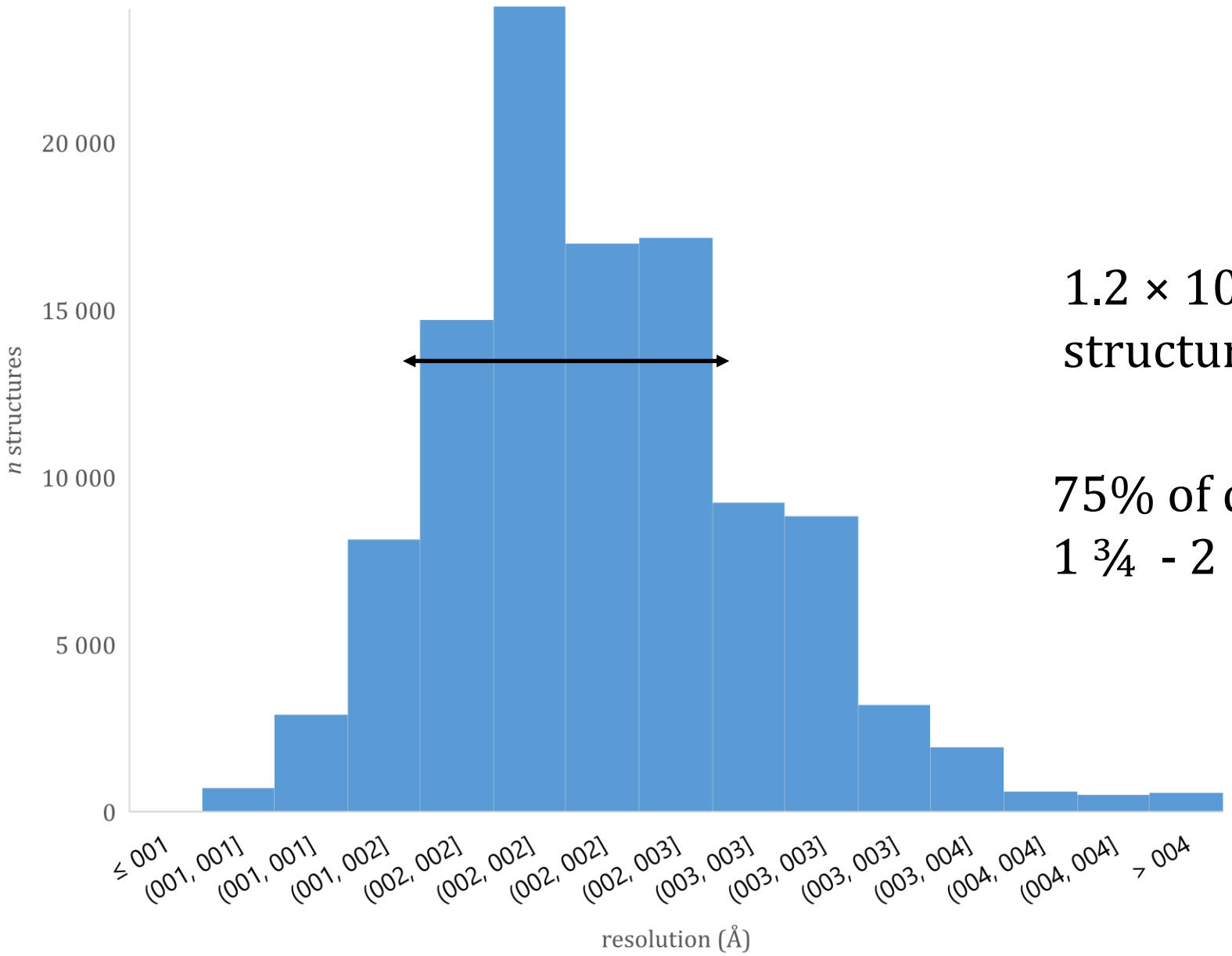
- formally – limit when two points become one point
- practical



best	< 1 Å	see H electrons
	1.2 Å	separate atoms
	2 Å	sidechains
worst	> 4 Å	overall shape

bl831.als.lbl.gov/~jamesh/movies/

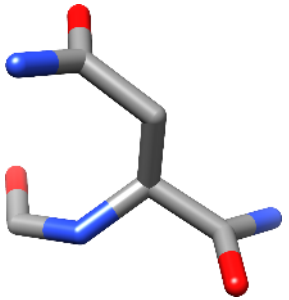
resolution in PDB

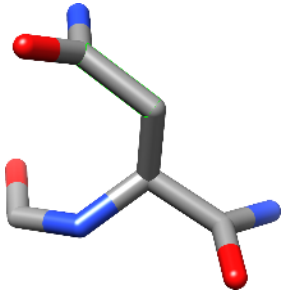


1.2×10^5 structures

75% of data
1 ¾ - 2 ½ Å

Errors, uncertainty





Errors ? – Many

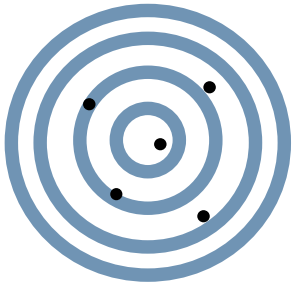
- good data - you cannot tell O from N
- bad data – you may slip by one amino acid
- spectacular errors
  - trace chain backwards, join wrong secondary structure units

Uncertainty

- if I have 2 Å data, what is my uncertainty ?

  Much smaller

  1. averaging over observations
  2. known bonds

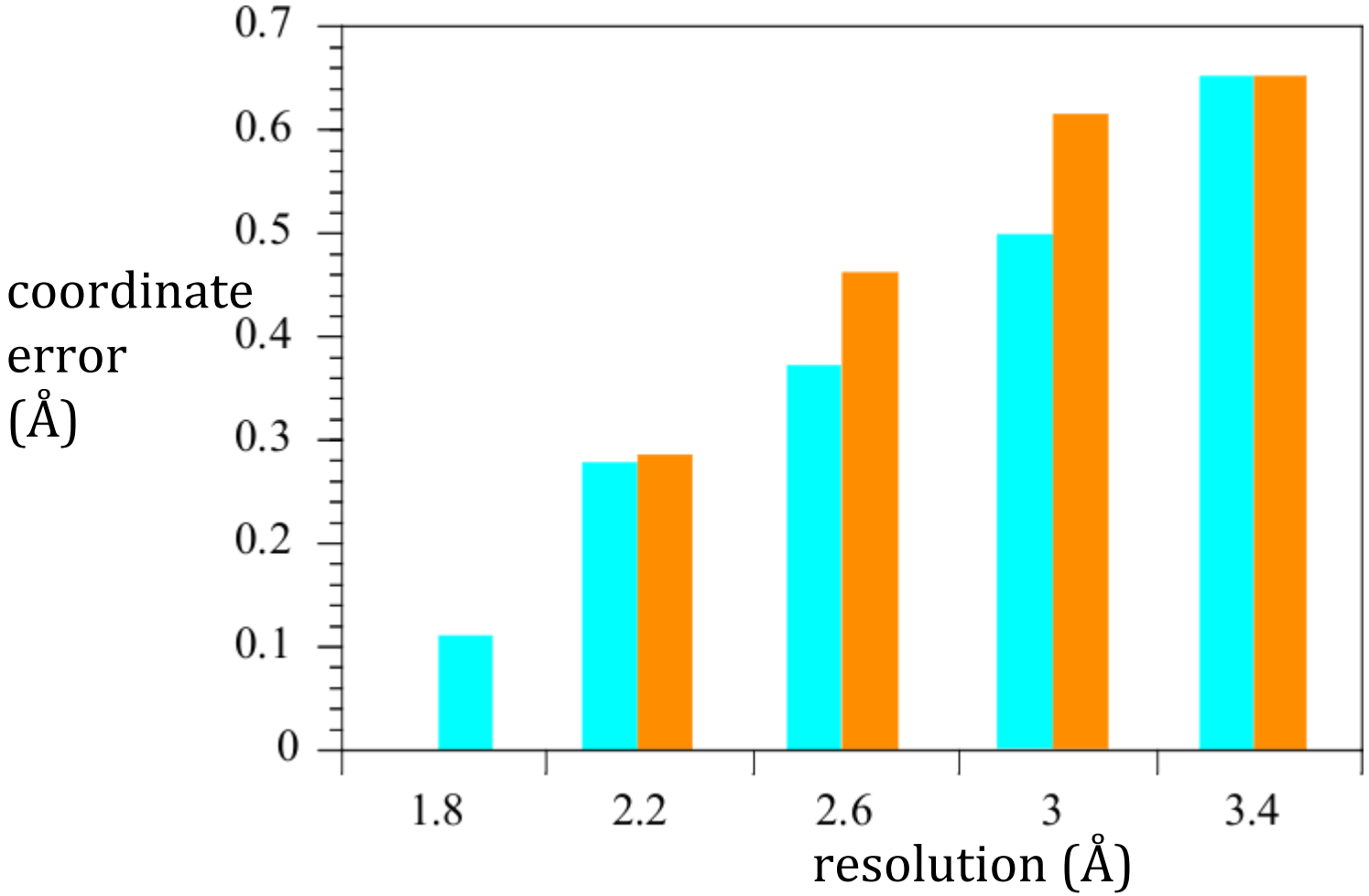


Estimate of uncertainty..

coordinate error - synthetic data

Resolution near 1.5 to 2.0 Å

- errors 0.2 to 0.3 Å



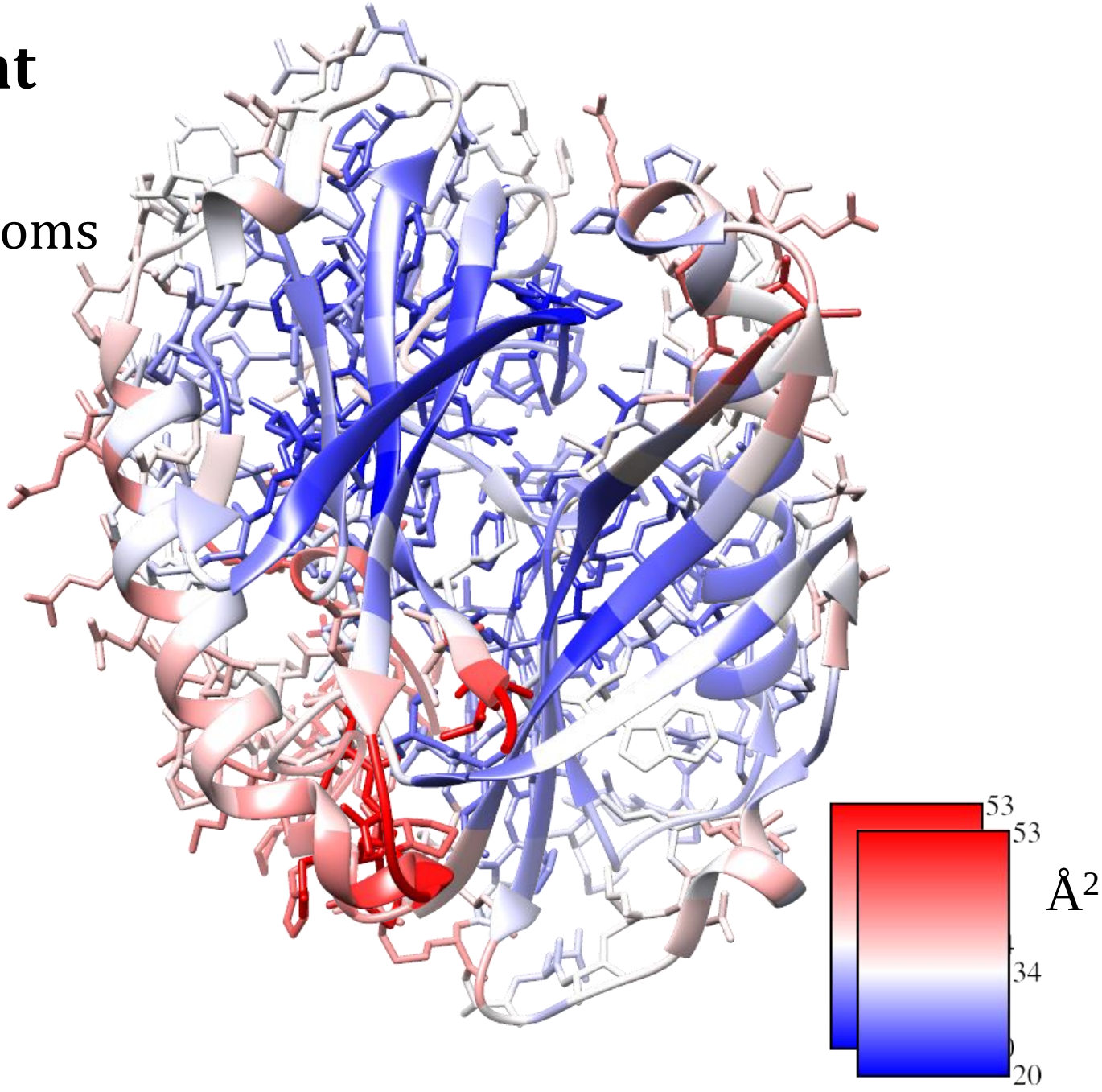
two different estimates of coordinate error – not important for us

Brünger AT, Nat Struct Biol. 1997 4, Suppl:862-865.

mobility – more important

2ei5
all atoms

2ei5
backbone



Real uncertainty

- not experimental error

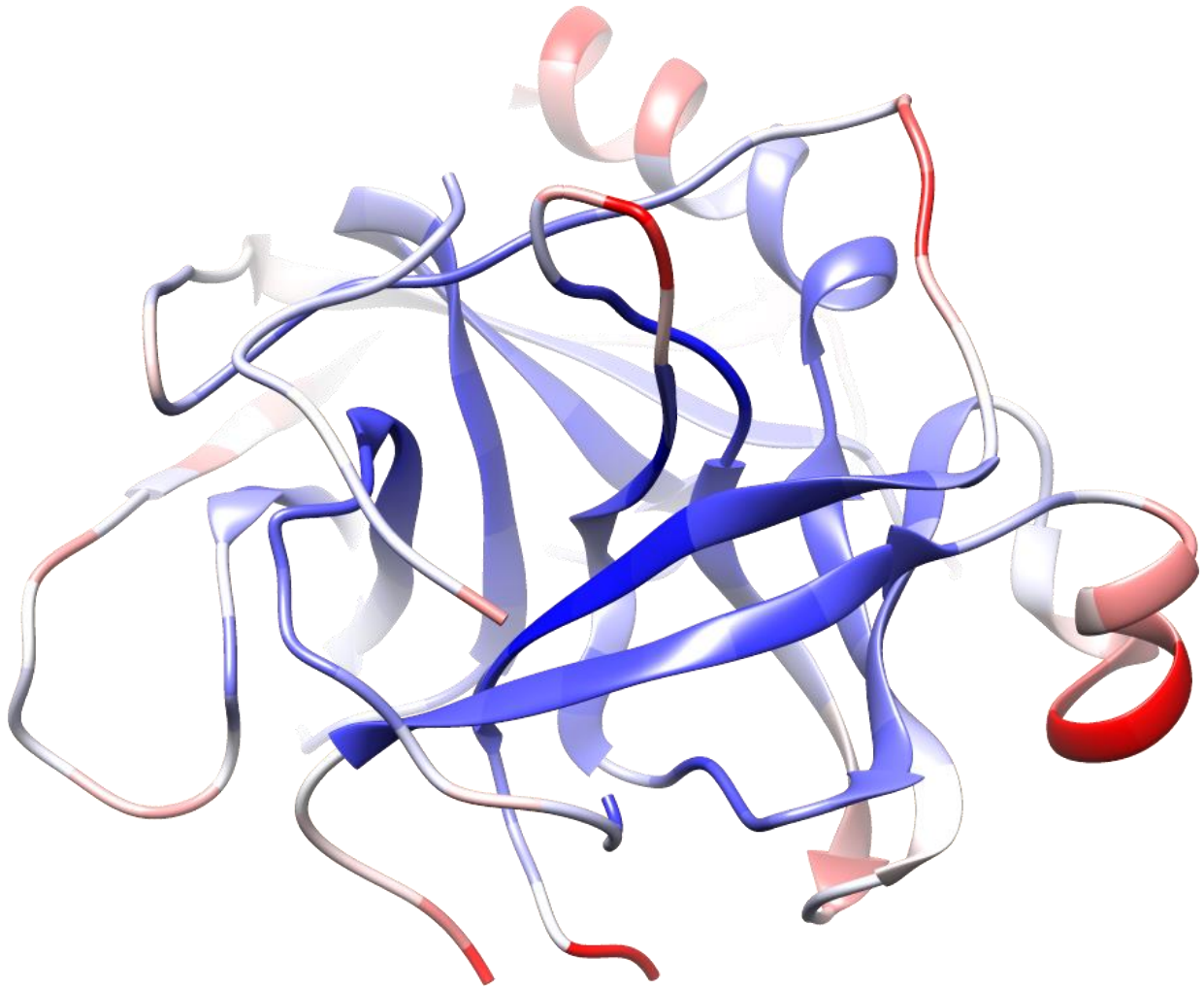
- .. mobility

when mobility is high…

missing atoms

Clear pieces of structure missing

Look at *B*-factors



NMR vs X-ray

	NMR	X-ray
certainty	spread amongst 50 models	*B*-factors / Gaussian model
resolution	no meaning	
size	rarely > 200 residues	big

not for discussion – make sure the ideas are clear

* Forming crystals is a question of energy differences
* Simple refraction, one dimension
* generalising to 3D not discussed
* from reflections to density via Fourier Transform
* phasing methods
* high and low resolution reflections
* static vs dynamic disorder
* fitting, overfitting
* R and R free
* B-factors and missing atoms