7.91 Amy Keating

How do we use computational methods to **analyze**, **predict**, or **design** protein sequences and structures?

Theme:
*Methods based on **physics** vs. methods based on our **accumulated empirical knowledge** of protein properties*

For a molecular simulation or model you need:

1. A representation of the protein

2. An energy function

3. A search algorithm or optimizer

Covalent Potential Energy Terms

$$U_{\text{Covalent}} = U_{\text{bond}} + U_{\text{bond angle}} + U_{\text{improper dihedral}} + U_{\text{torsion}}$$

$$U_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b (b - b_0)^2$$

$$U_{\text{bond angle}} = \sum_{\text{bond angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

$$U_{\text{improper dihedral}} = \sum_{\text{improper dihedrals}} \frac{1}{2} k_\Phi (\Phi - \Phi_0)^2$$

$$U_{\text{torsion}} = \sum_{\text{torsions}} \frac{1}{2} k_\phi [1 + \cos(n\phi - \delta)]$$

Brooks et al., *J. Comput. Chem*. 4: 187-217 (1983)

Non-Covalent Potential Energy Terms

$$U_{\text{Non-covalent}} = U_{\text{vdW}} + U_{\text{elec}}$$

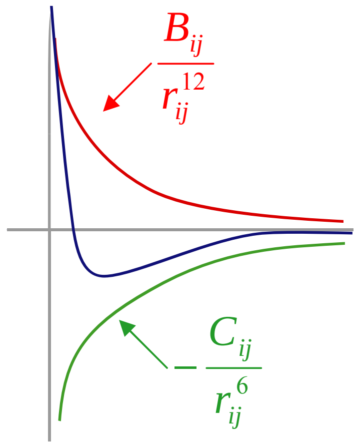
$$U_{\text{vdW}} = \sum_{i \succ j} \left(\frac{B_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}} \right)$$

“accurate”

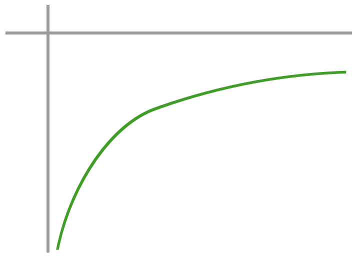
approximate

$$U_{\text{elec}} = \sum_{i \succ j} \frac{q_i q_j}{\varepsilon r_{ij}}$$

Lennard-Jones potential



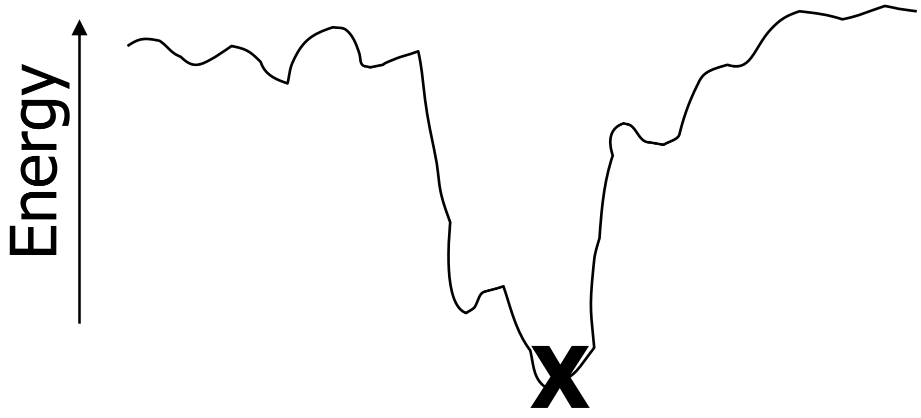
Coulomb's law



The potential energy surface is a 3N-6 dimensional space.

For a protein, we assume a single native-structure minimum.

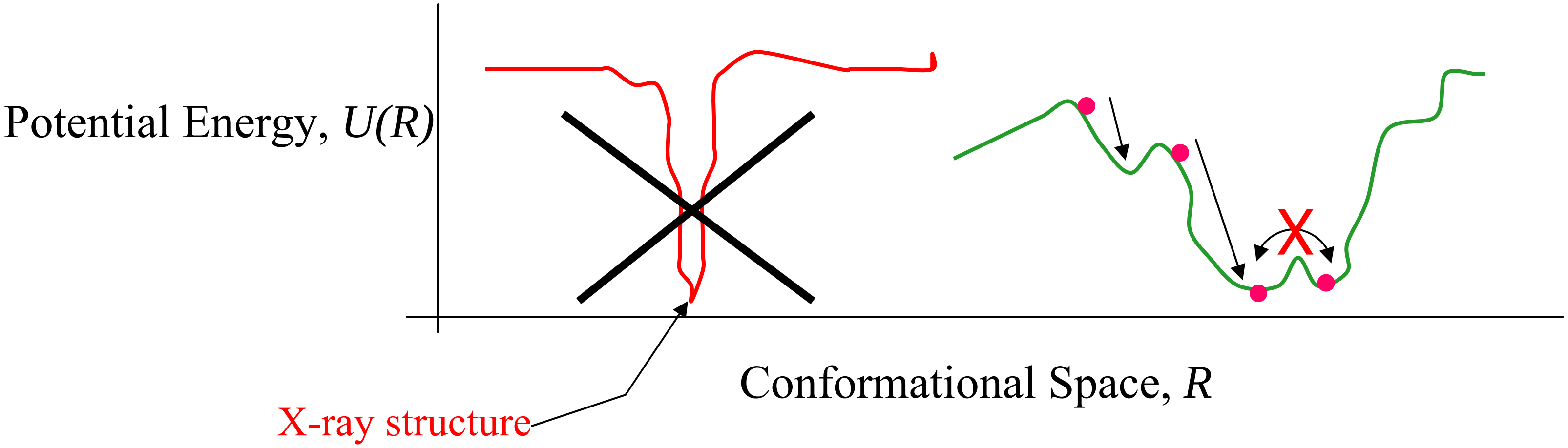
There are many local minima, and some *may* be close in energy to the global minimum.



Sampling the Potential Energy Surface

- **Energy minimization**
  - "downhill" search, generally to nearest local minimum
  - can be used to relax structures
  - might be useful to define local changes due to mutation
- **Normal mode analysis**
  - defines "characteristic motions", which are distortions about a local minimum structure
  - orders motions "easy" (low frequency) to "hard" (high)
- **Molecular dynamics**
  - movie of motion at given temperature (300 K)
  - equivalent to statistical mechanical ensemble
- **Monte Carlo/Simulated Annealing**
  - Describe properties of the landscape and thermodyanmic parameters without simulating how the molecular actually moves

Energy Minimization



$$\overrightarrow{F_i} = -\nabla U(\overrightarrow{r_i})$$

$$\overrightarrow{r_{i+1}} = \overrightarrow{r_i} + \delta \overrightarrow{F_i}$$

- Iterative procedures; terminate when reach tolerance, such as small gradient
- Poor initial structure leads to poor local minimum
- Multiple minimum problem

ONLY FINDS *LOCAL* MINIMA!

Uses of simple minimization

1. The “minimum perturbation approach” to modeling a mutation
   – Assume structure of single-site mutant is close to known wild-type structure
     - Find stable conformations for mutant side chain in context of wild-type protein
     - Use energy minimization to relax candidate structures (all degrees of freedom)

Shih, Brady, and Karplus, *Proc. Natl. Acad. Sci. USA* **82**: 1697–1700 (1985); hemagglutinin Gly to Asp mutation modeled accurately

2. Relieving strain before analyzing the energy of an experimental or predicted structure
3. Structure building and refinement when solving structures using X-ray crystallography or NMR

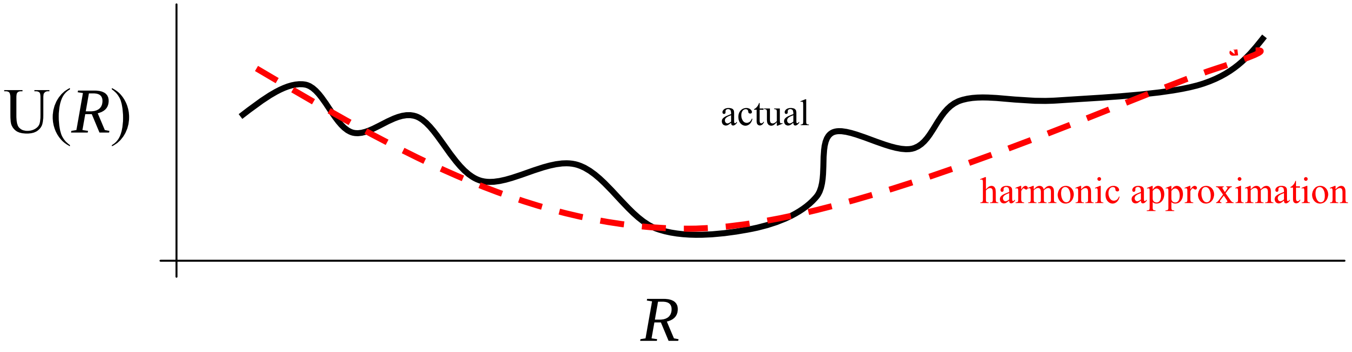
Normal Mode Analysis

- Characteristic Motions and their Relative Ease
- Thermodynamic Properties

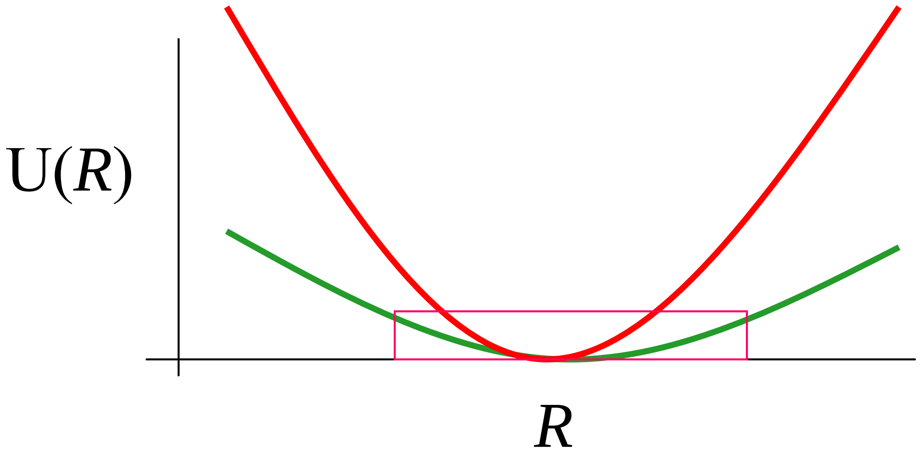
Mathematical Approximation: Series of Independent Harmonic Oscillators

$$U(\vec{R}) = U(\vec{R}_0) + \cancelto{0}{\nabla U(\vec{R}_0)}(\vec{R} - \vec{R}_0) + \frac{1}{2}\sum_{j>i}\sum_{i}\frac{\partial^2 U}{\partial r_i \partial r_j}(r_i - r_{i,0})(r_j - r_{j,0}) + \cdots$$

Corresponds to Local Expansion of Potential Surface as Parabolic



Normal Modes Locate "Easy" Deformations



- Low-frequency, energetically easy motions will dominate the dynamical behavior of macromolecules

Normal modes of proteins sometimes correspond to biologically relevant motions

- “shearing” or “hinging” conformational changes are common in enzymes
- Can compare structure changes computed from NMA with alternate conformations observed experimentally. Frequently a low-frequency mode describes the change (but it may not be the lowest energy mode)
- NMA is a way to get an idea about motion from a static structure

Molecular Dynamics Simulations

- Simulate motions of molecules as a function of time
  - Collect short “movie” of protein “swimming” in solution
- Can use the simulation to compute:
  - Average structure (at given temperature, *T*)
  - Atomic fluctuations (at *T*)
  - Thermodynamic quantities (temperature dependent)

$$\vec{F}_i = m_i \vec{a}_i \qquad \vec{a}_i = \frac{\vec{F}_i}{m_i} \qquad \frac{\partial^2 x_i}{\partial t^2} = -\frac{\nabla_i U}{m_i}$$

$$\vec{r}(t+\delta t) = \vec{r}(t) + \delta t \vec{v}(t+\delta t/2)$$

$$\vec{v}(t+\delta t/2) = \vec{v}(t-\delta t/2) + \delta t \vec{a}(t)$$

Verlet “leap frog” algorithm

Running a Molecular Dynamics Simulation

- Minimize initial coordinates
- Assign random starting velocities from Maxwell-Boltzmann distribution

$$p(v_i) = \left(\frac{m_i}{2\pi kT} \right)^{1/2} \exp\left(\frac{-m_i v_i^2}{2kT} \right)$$

- Use **small time step** to integrate Newton’s equations of motion (1 fs =1 x 10^{-15} sec is typical)
- Equilibration (running dynamics until system equilibrates)
- Production dynamics (useful dynamics at a desired temp)

Sampling conformational space using Monte Carlo

Monte Carlo randomly samples configurations, rather than simulating how the molecule would actually move in time.

```
1. Begin with a random (energy-minimized) conformation
2. Evaluate the energy = $E^0$
3. Make a random conformational change
   (e.g. rotation around a bond)
4. Evaluate the new energy = $E'$
5. DECISION:
   if $E' < E^0$ ACCEPT the move, set $E^0 = E'$, go to 3
   if $E' > E^0$ AND if $\exp[-(E'-E^0)/kT] >$ random #
       then ACCEPT the move, set $E^0 = E'$, go to 3
   else reject the move, go to 3
6. Proceed for an arbitrary amount of time
```

The acceptance criterion is such that the conformations generated sample the Boltzman distribution.

What are Monte Carlo and Molecular Dynamics Useful For?

1. Exploring the energy landscape

   What is the global minimum?
   What are other minima?
   What are the barriers between minima?

2. Computing thermodynamic quantities

3. “Observing” pathways connecting different states

What are the differences between MD and MC?

Molecular dynamics and Monte Carlo can be used for many of the same purposes.

If you want to know how a system evolves in time, you must use MD.

MC can be better at sampling broad conformational spaces because it makes random moves that can get you out of local minima.

A Dynamic Model for the Allosteric Mechanism of GroEL

MD simulation used to probe the mechanism of conformational change in the chaperonin GroEL that occurs upon binding ATP and GroES.

Please see figure 1a of

Ma, J, PB Sigler, Z Xu, and M Karplus. "A Dynamic Model for The Allosteric Mechanism of GroEL." *J Mol Biol.* 302, no. 2 (15 September 2000): 303-13.

Shown are conformational changes that occur upon binding ATP.

The endpoints of the calculation were determined from experimental X-ray structures, but the path in between isn’t accessible experimentally.

The conformation change is too slow to simulate normally, but “targeted MD” can identify such trajectories.

The yellow intermediate domain moves downwards first, triggering the upwards movement of the green apical domain.

Simulation shows details of interactions responsible for both positive and negative cooperativity.

Please see figures 2b and 5b of

Ma, J, PB Sigler, Z Xu, and M Karplus. "A Dynamic Model for The Allosteric Mechanism of GroEL." *J Mol Biol.* 302, no. 2 (15 September 2000): 303-13.

MD simulations are now run on ENORMOUS systems!

Simulation of the aquaporin channel: permeable to water but not protons. The waters and the lipid bilayer are both present in the calculation - 101,000 atoms total.

Please see figure 1 of

de Groot, Bert L., and Helmut Grubmüller. "Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF." *Science* 294 (14 December 2001): 2353-2357.

MD simulation of water permeation through aquaporin

A 10 ns simulation was sufficient to see water move in “real time” because the permeation rate is 3x109 s-1.
Can look at the mechanism of selectivity and what determines the rate.

Please see figure 2 of

de Groot, Bert L., and Helmut Grubmüller. "Water Permeation Across Biological Membranes: Mechanism and Dynamics of Aquaporin-1 and GlpF." *Science* 294 (14 December 2001): 2353-2357.

Can you fold a protein using Molecular Dynamics?

We don’t really know yet!

Simulate the folding of villin headpiece:
36 residues, with explicit solvent.
1 us simulation with 2 fs timestep, took, 512 Cray processor months!

Results:
The NMR structure is stable in the simulation.
After 60 ns the structure showed 50% helicity, in good agreement with experimental data for other proteins
There was an initial “burst” phase giving helix formation and native contacts
A “metastable intermediate” was found with secondary and tertiary contacts similar to the native state.

Duan & Kollman Science (1998) 282, 740-744

Villin headpiece

A. unfolded
B. 980 ns partly folded
C. native
E. structure from stable cluster

Time evolution of:

A. Native helical content
B. Native contacts
C. Radius of gyration
D. Solvation free energy

Please see figures 1 and 2 of

Duan, Yong, and Peter A. Kollman. "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution." *Science* 282 (23 October 1998): 740-744.

Molecular Modeling - a few final words

- Potential functions used for modeling are approximate (though quite good for some purposes); calculations are not replacement for experiment

- Ability to “dissect” computational results leads to insights unavailable by other approaches

- The need to specify the problem precisely in order to build a model is itself valuable

Solving structures using

X-ray crystallography

&

NMR spectroscopy

How are X-ray crystal structures determined?

1. **Grow crystals** - structure determination by X-ray crystallography relies on the repeating structure of a crystalline lattice.

2. **Collect a diffraction pattern** - periodically spaced atoms in the crystal give specific “spots” where X-rays interfere constructively.

3. Carry out a **Fourier transform** to get from “reciprocal space” to a real space description of the electron density.

4. THIS STEP REQUIRES KNOWLEDGE OF THE PHASES OF THE INTERFERING WAVES, WHICH CAN’T BE DIRECTLY MEASURED “**THE PHASE PROBLEM**”

4. **Build a preliminary model** of the protein into the envelope of electron density that results from the experiment.

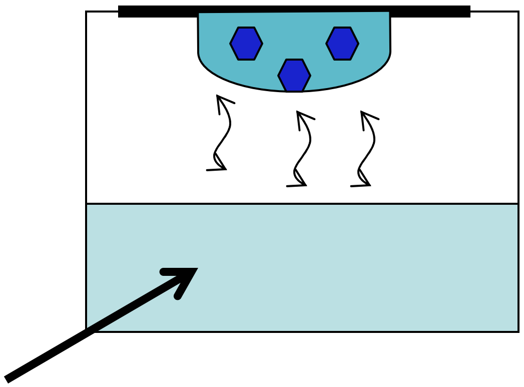
5. **Refine the structure** through an iterative process of changing the model and comparing how it fits the data.

Growing Protein Crystals

This is often the rate-limiting step in solving a structure!

1. choose a sequence that you think folds into a well-defined structure
2. prepare highly-purified protein
3. concentrate to > 10 mg/ml

hanging-drop vapor diffusion

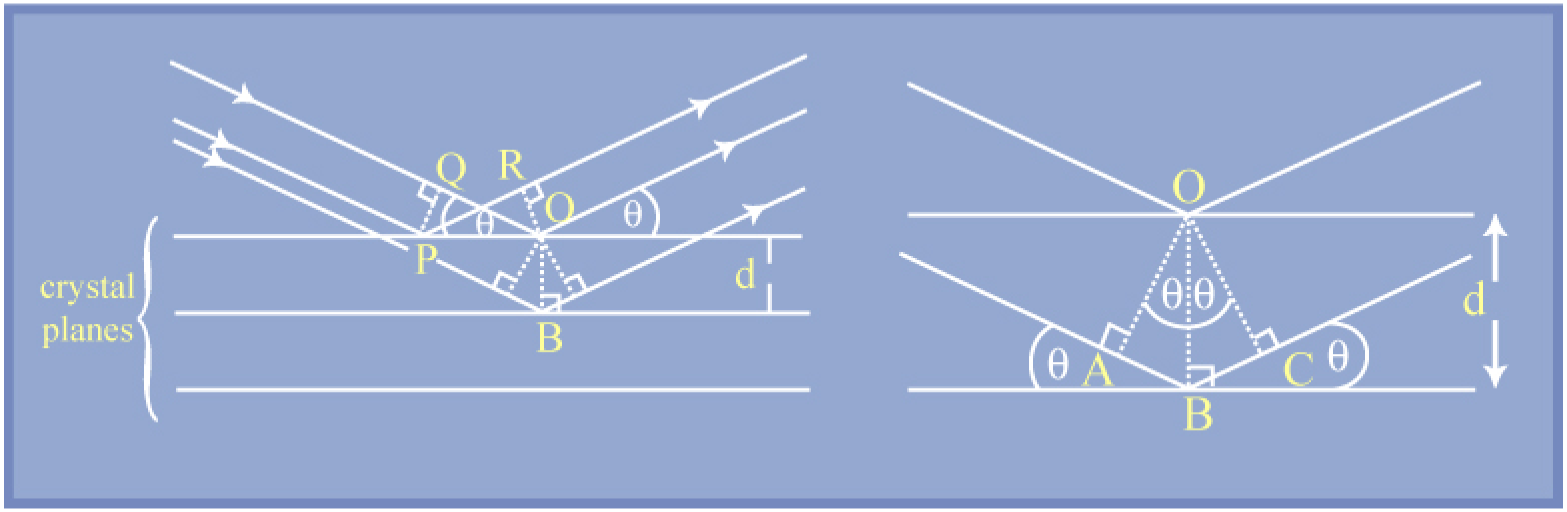


buffer, salts, precipitant

Collecting Diffraction Data

What do the spots in the diffraction pattern mean?

Every spot represents *constructive* interference between diffraction from a set of atoms with spacing satisfying Bragg's Law:



$$n\lambda = 2d\sin\theta$$

So the INTENSITY of a spot at a given *q* tells you something about how much electron density lies in a set of periodic planes with spacing *d*.

NOTE: λ = 0.5 to 1.5 Å

The diffraction pattern represents <u>reciprocal space</u>

large θ -> small d; HIGH resolution

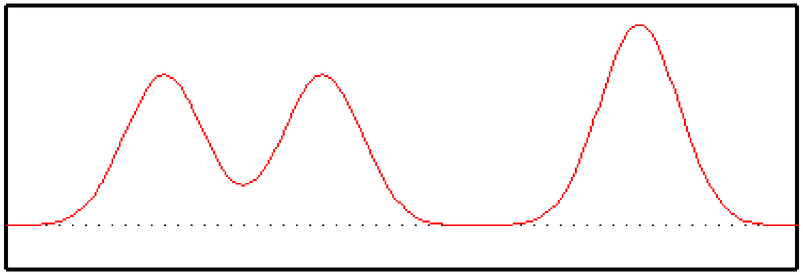
each spot results from a differently-oriented set of planes with the appropriate spacing

small θ -> large d
LOW resolution

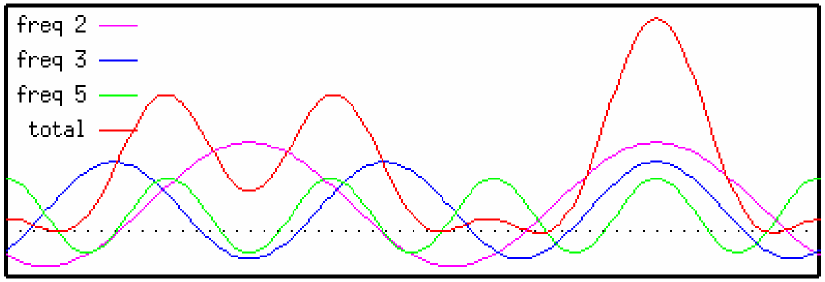
Image of diffraction data removed for reasons of copyright.

The electron density in the protein is a superposition of all of these periodic functions…

Fourier Transform

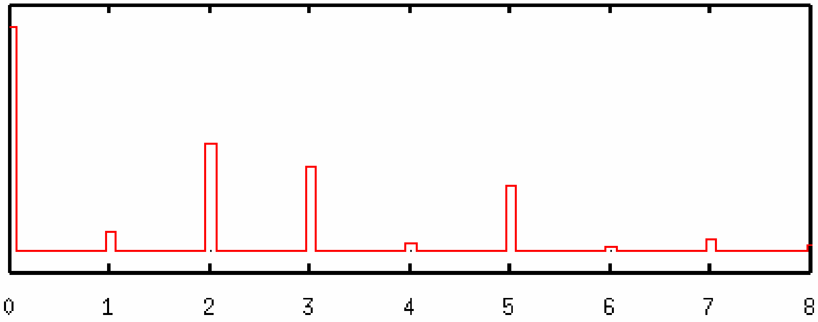


This is the “electron density”



$$f(x) = \int F(s) \exp(-i\, 2\pi xs)\, ds$$

$$F(s) = \int f(x) \exp(i\, 2\pi xs)\, dx$$



This is the “diffraction pattern”

Courtesy of Kevin Cowtan: http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html

The Phase Problem: we don’t know what phases to use to add up all of the contributing waves. BIG PROBLEM.

$$\rho_{(x,y,z)} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l} F_{(h,k,l)} \exp\left[-2\pi \cdot i(hx + ky + lz)\right]$$

$$|F_{hkl}|\exp(i\alpha_{hkl}) = \quad F_{(h,k,l)} = \sum_{j=1}^{atoms} f_{(j)} \exp\left[2\pi \cdot i(hx_{(j)} + ky_{(j)} + lz_{(j)})\right]$$

↑ observable amplitude

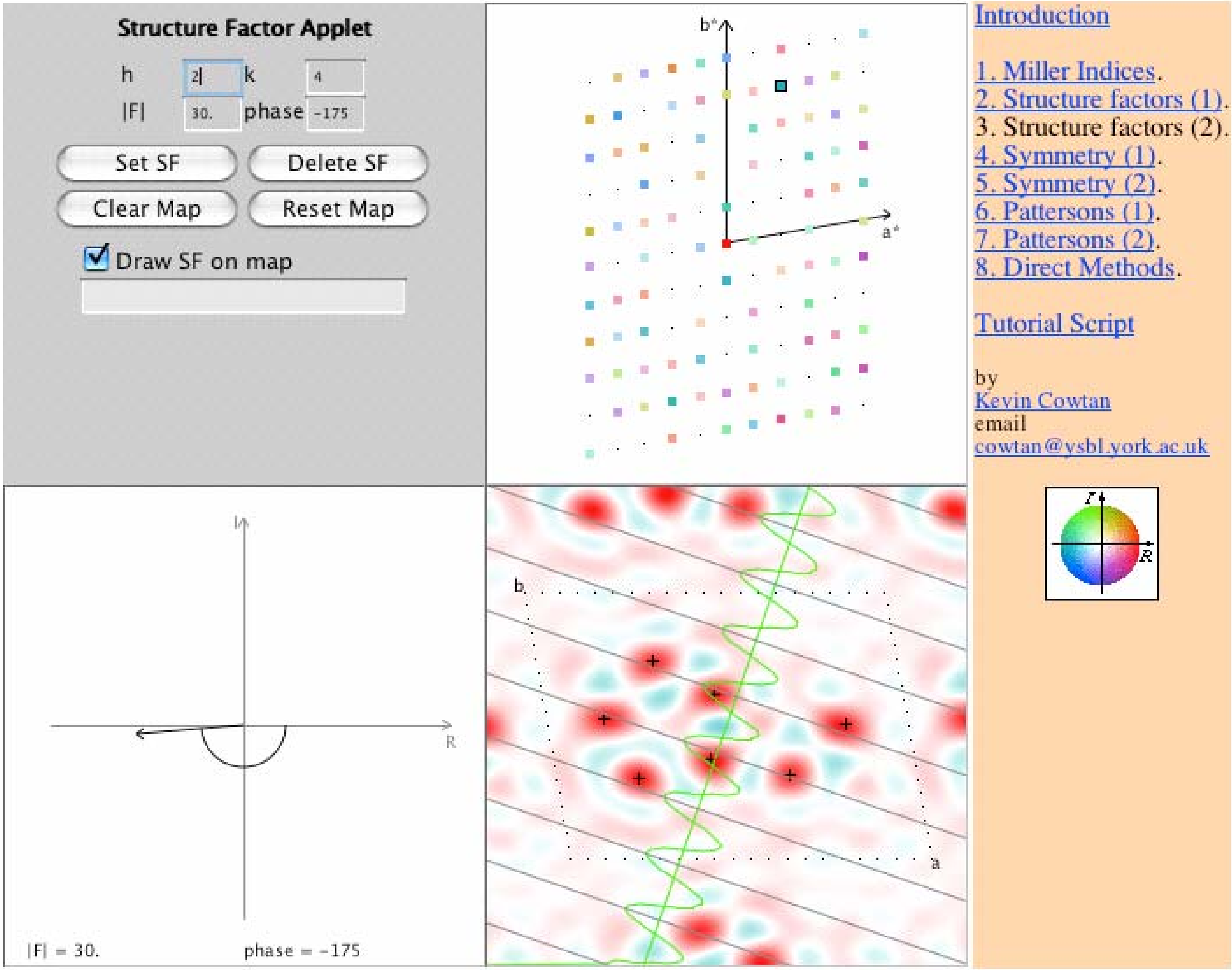
↑ atomic scattering factor - related to electron density around atom j

↑ the *phase* of F is determined by the x, y and z coordinates of the atoms

What we *observe* is $I_{hkl}\ \alpha\ |F_{khl}|^2$

we can’t measure the phases directly

F_{hkl} is called a <u>structure factor</u>



Structure Factor Applet

h 2 k 4

|F| 30. phase -175

☑ Draw SF on map

|F| = 30. phase = -175

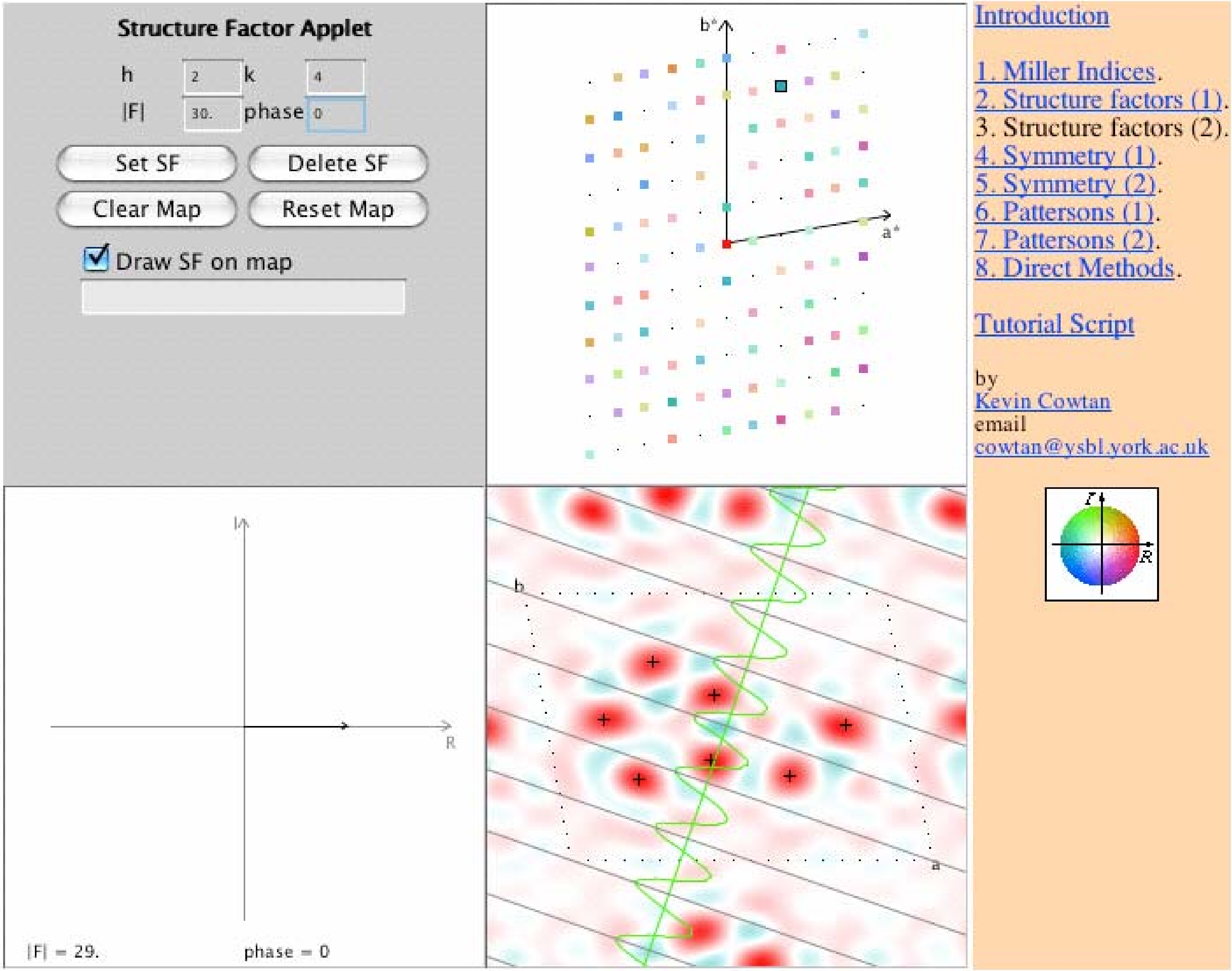
Introduction

1. Miller Indices.
2. Structure factors (1).
3. Structure factors (2).
4. Symmetry (1).
5. Symmetry (2).
6. Pattersons (1).
7. Pattersons (2).
8. Direct Methods.

Tutorial Script

by
Kevin Cowtan
email
cowtan@ysbl.york.ac.uk

Courtesy of Kevin Cowtan: http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html



Structure Factor Applet

h 2 k 4

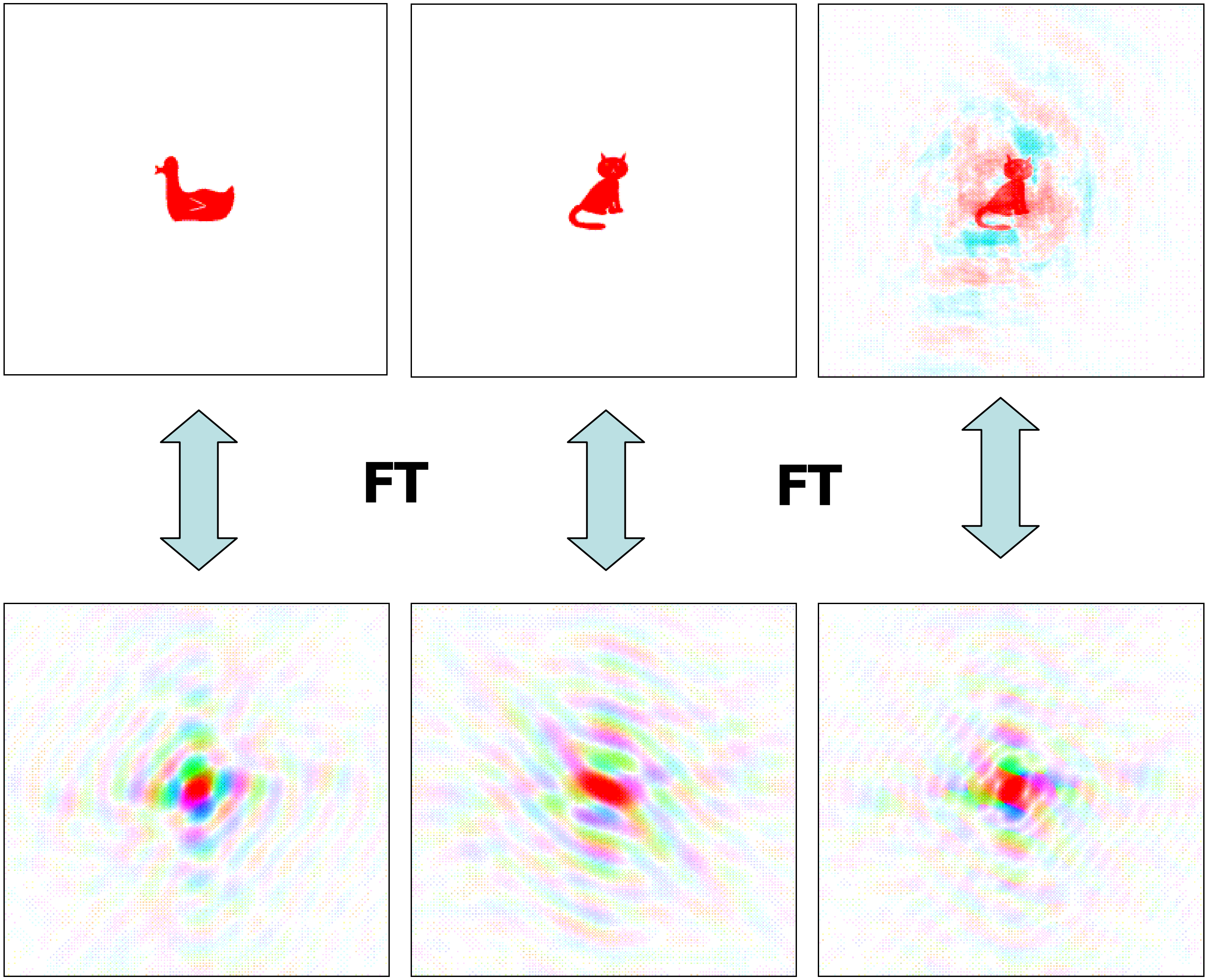
|F| 30. phase 0

☑ Draw SF on map

|F| = 29. phase = 0

Courtesy of Kevin Cowtan: http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html

Contribution of intensities vs. phases to the Fourier Transform



duck intensities
+
cat phases

Courtesy of Kevin Cowtan: http://www.ysbl.york.ac.uk/~cowtan/sfapplet/sfintro.html

So, how do you get the phases?
I. Molecular Replacement

“Borrow” phases from a structure that you think is similar.

e.g. duck intensities plus *goose* phases would give a reasonable estimate of the real structure.

Beware “model bias” - need to run controls to make sure you haven’t forced the data to assume the structure of the model.

How it works:
Take the model structure and translate/rotate it in the unit cell of the crystal. Calculate the expected structure factors.
Monitor the R value.
If you find a significantly non-random R value, this might be a good model. Try to refine it and see if the R_{free} value improves.

So, how do you get the phases?
II. Heavy atom methods

If you can make two or more crystals of the *same geometry* that have heavy atoms incorporated, then $\mathbf{F}_{\text{protein}\bullet\text{heavy}} = \mathbf{F}_{\text{protein}} + \mathbf{F}_{\text{heavy}}$ and taking the difference of a diffraction pattern with and without the heavy atoms will give you the diffraction pattern for *just* the heavy atoms.

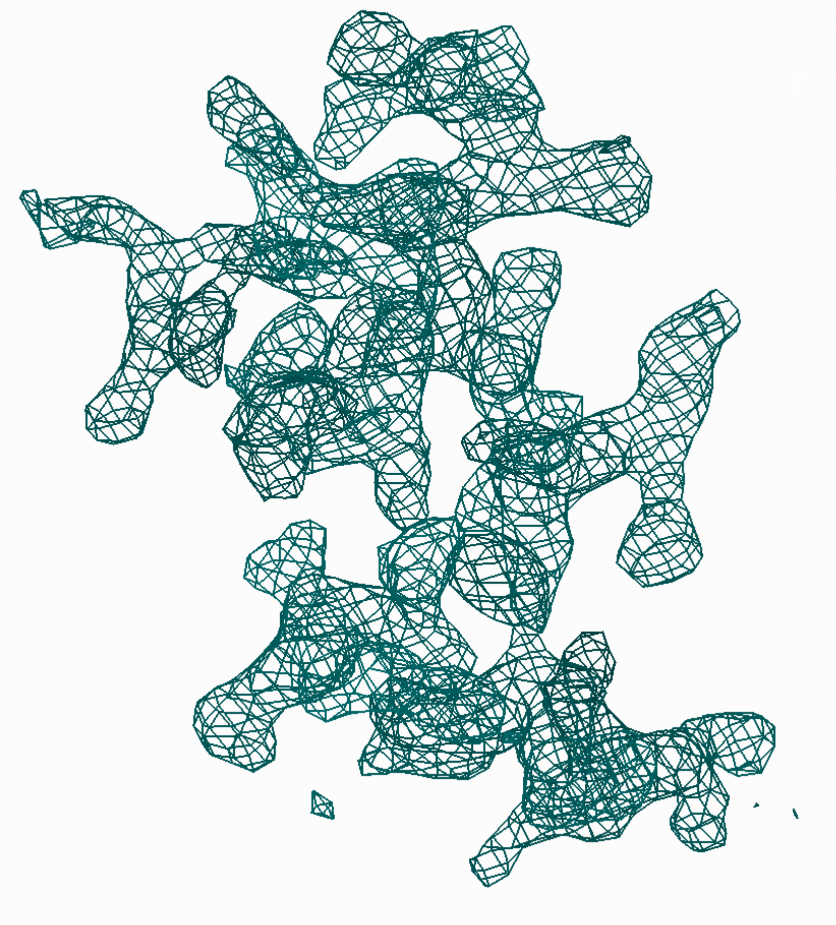
This structure can be solved using a “Patterson map”:

$$P(u,v,w) = \frac{1}{V}\sum_h \sum_k \sum_l |F_{hkl}^2| \exp(-2\pi i(hu+kv+lw))$$

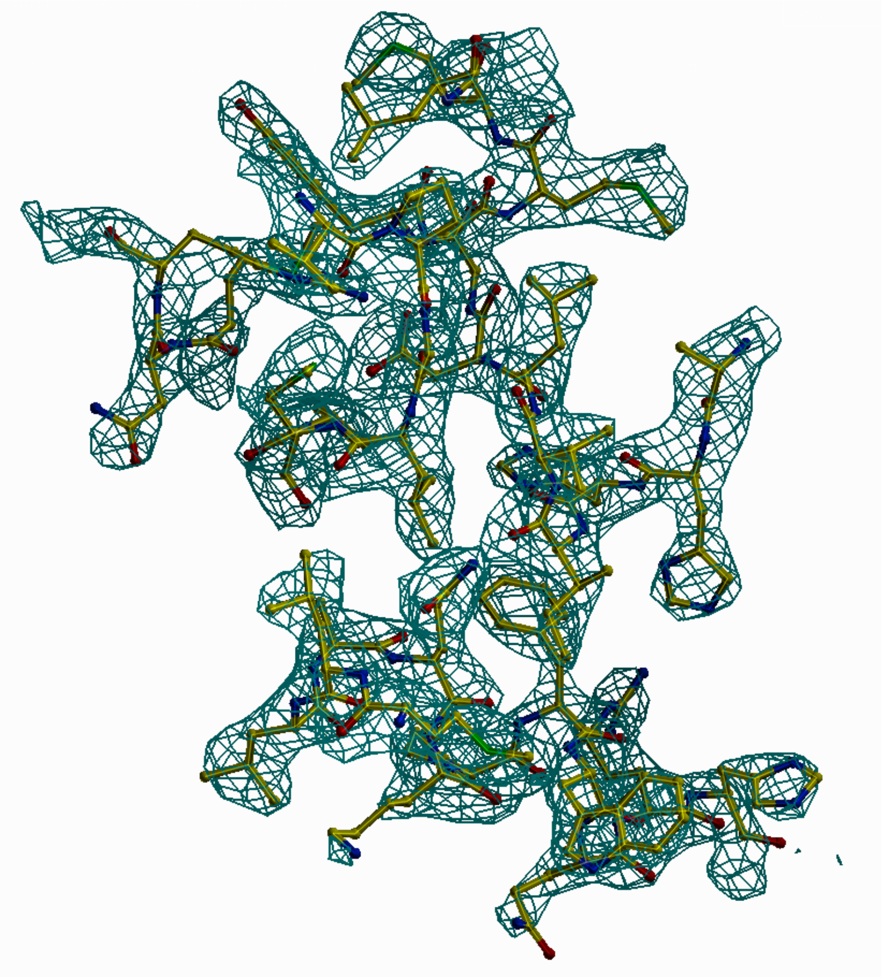
The Patterson map gives you all of the vectors between atoms. This doesn’t help for a whole protein, but for a sufficiently small structure, you can get the entire structure from this data (except the chirality).

Once you have the structures of the heavy atom derivatives, you can use this to derive the phases for the original structure. (Trust me on this - we don’t have time to do the details.)

Map tracing usually requires knowledge of the sequence

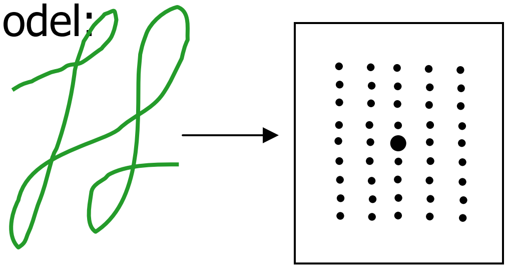


2.55Å map



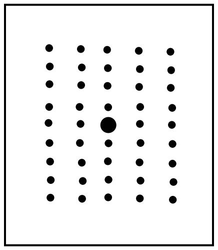
X-Ray Crystal Structure Refinement

The model:



Computed intensities of spots

The data: Actual intensities of spots



$$U_{\text{X-ray expt}} = \sum_{h,k,l} \left[\, \left|F_{\text{obs}}(h,k,l)\right| - \left|F_{\text{calc}}(h,k,l)\right| \, \right]^2$$

Summation runs over spots

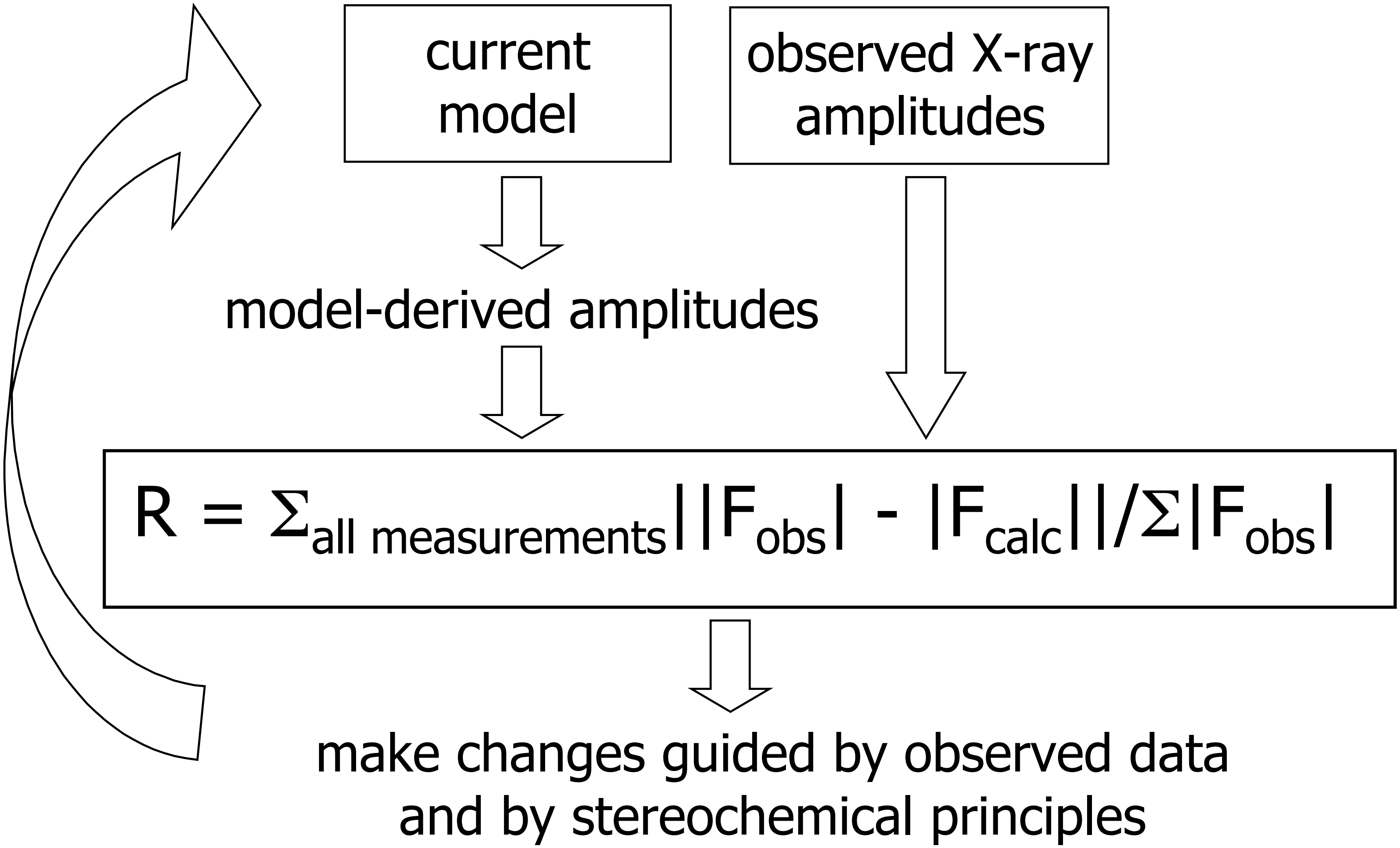
Actual intensity of spot observed in expt

Intensity of spot calculated from trial structure

$$U_{\text{hybrid}} = U_{\text{Molec Model}} + sU_{\text{X-ray expt}}$$

- Simulated annealing on hybrid potential rapidly improves correspondence between structure and X-ray observations while maintaining reasonable chemistry (large radius of convergence)
- Previous method effectively used local minimization which became trapped in local minima (small radius of convergence)

Structure refinement and the R factor



current model

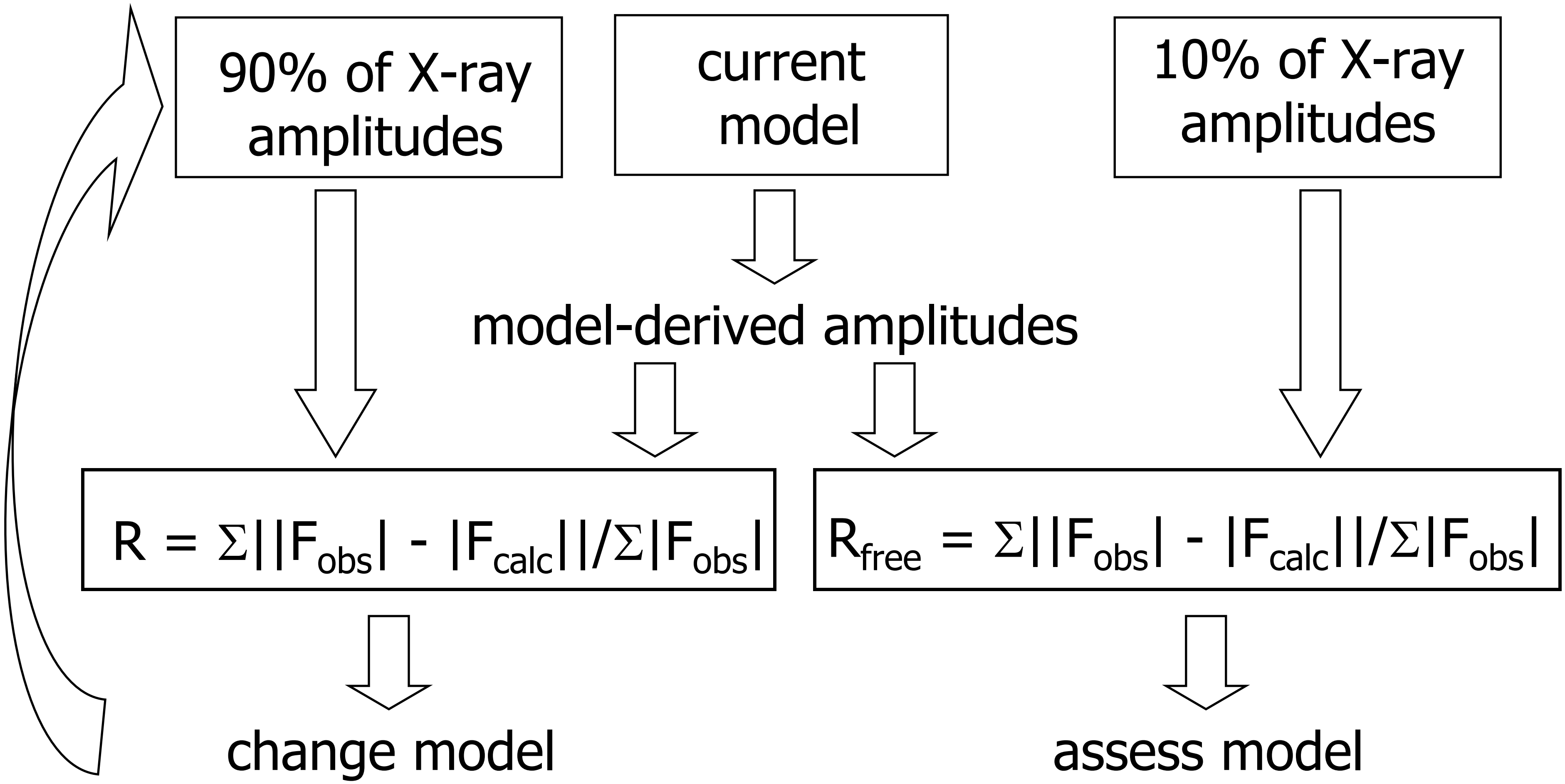
observed X-ray amplitudes

model-derived amplitudes

$$R = \Sigma_{\text{all measurements}}||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$$

make changes guided by observed data and by stereochemical principles

The Free R factor



90% of X-ray amplitudes

current model

10% of X-ray amplitudes

model-derived amplitudes

$R = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$

$R_{free} = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$

change model

assess model