CMPS 6630: Introduction to Computational Biology and Bioinformatics

Experimental Structure Determination Methods

Fold Recognition - Threading

Differences Between Fold Recognition Algorithms

- Protein Model and Interaction Description

  The full three-dimensional structure is often simplified

- Energy Parameterization

  Energy functions not as sophisticated as we'll see in molecular simulation

- Alignment Algorithms

  Dynamic Programming with Frozen Approximation
  Double Dynamic Programming
  Monte Carlo Minimization
  Branch-and-Bound

Limitations

- Fold Recognition algorithms will return the fold that minimizes the energy function or maximizes the alignment score - but that doesn't mean the identified model is correct.
- Identified model structure is often not as good as in homology modeling

Experimental Structure Determination

Methods

X-Ray Diffraction - *X-Ray Crystallography*

Nuclear Magnetic Resonance Spect.
 - *NMR Spectroscopy*

Produce atomic coordinates for most atoms

Objective end-products

- XRC produces an electron density map
- NMR produces a set of geometric constraints

Objective end-products are interpreted

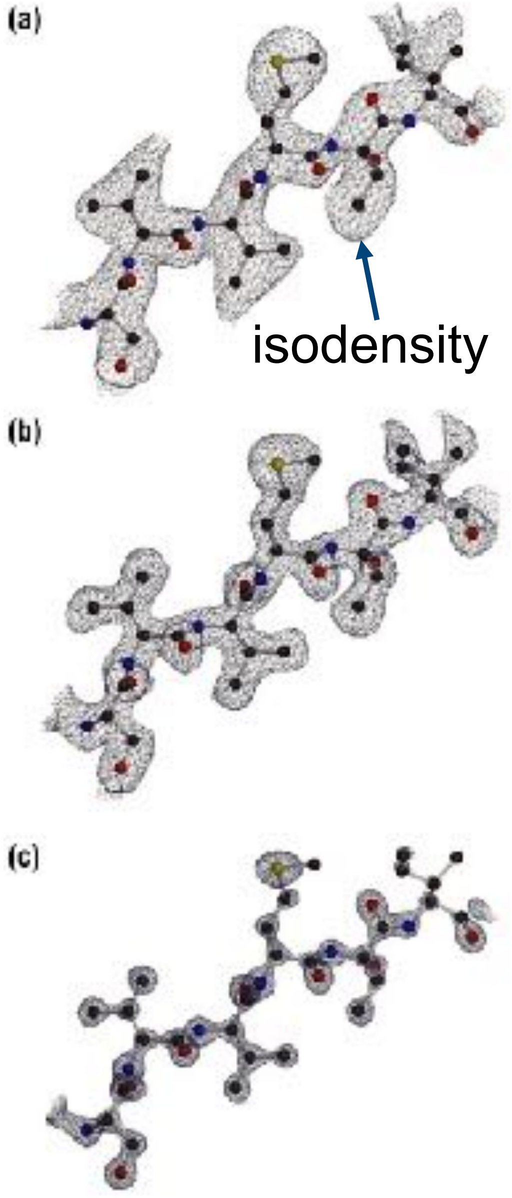
- Structures can have errors (usually small)

For larger proteins (>50-100kDa) XRC is best

Smaller proteins or complexes either ok

Study of dynamics best with NMR

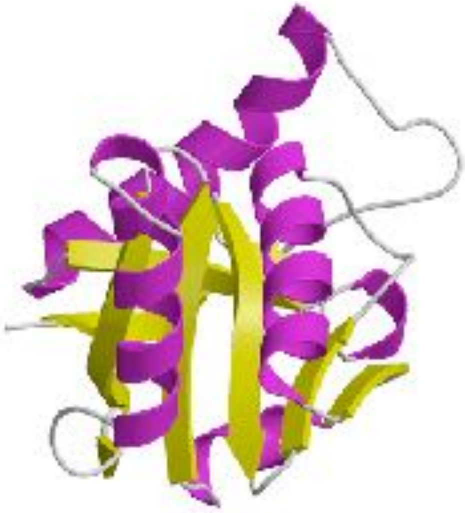
But constraints on what will crystallize
or dissolve at high concentrations



Experimental Structure Studies

Use a Microscope?

Take a picture?



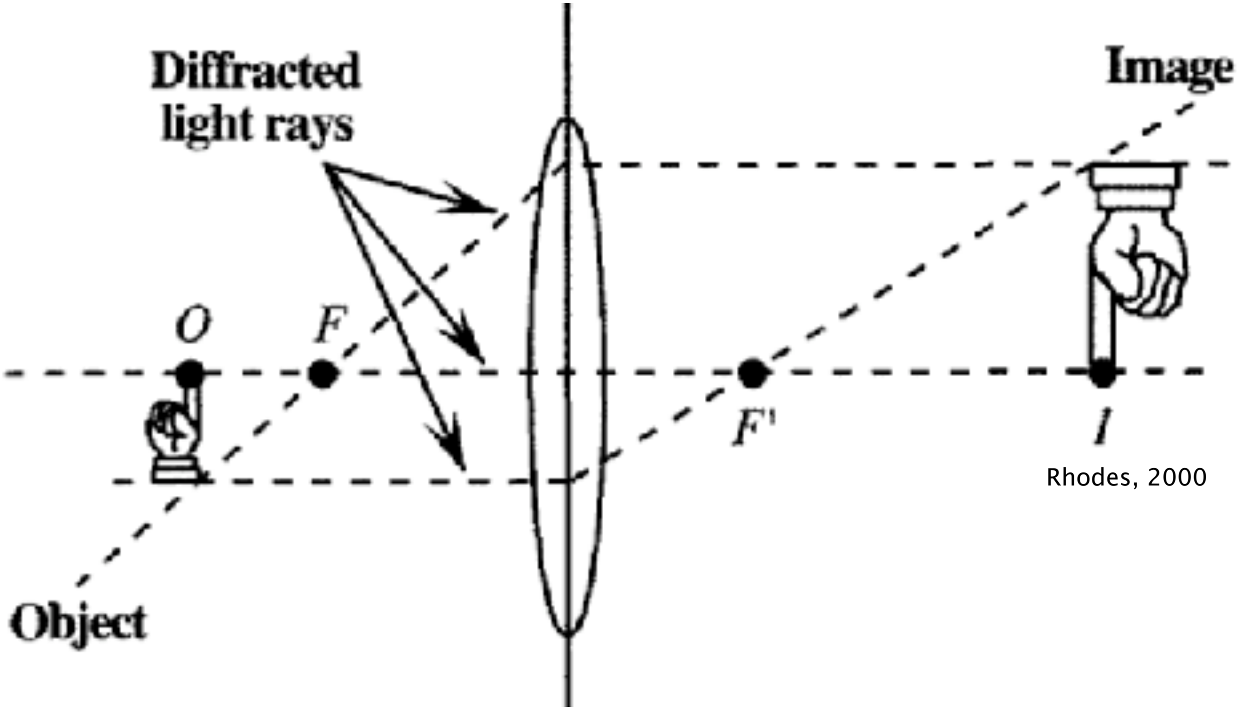
To diffract light, wavelength of light must be no larger than the object (or object features)

Visible Light
400-700nm (4000-7000A)

Atomic Spacing
0.15nm (1.5A)

X-rays

X-Ray Crystallography



Rhodes, 2000

Problems:

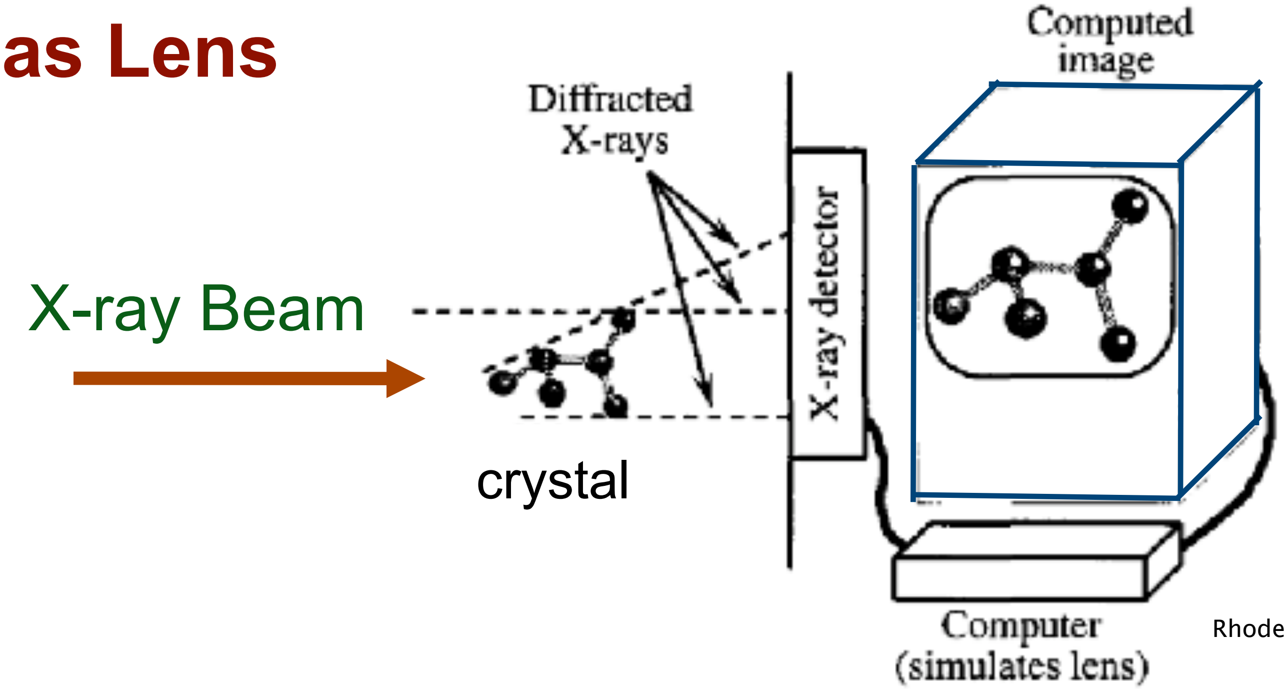
Single molecule is very weak diffractor
We don’t know how to build X-ray lenses

Solutions:

Use multiple molecules
Observe scattered diffractions - use the computer as a lens

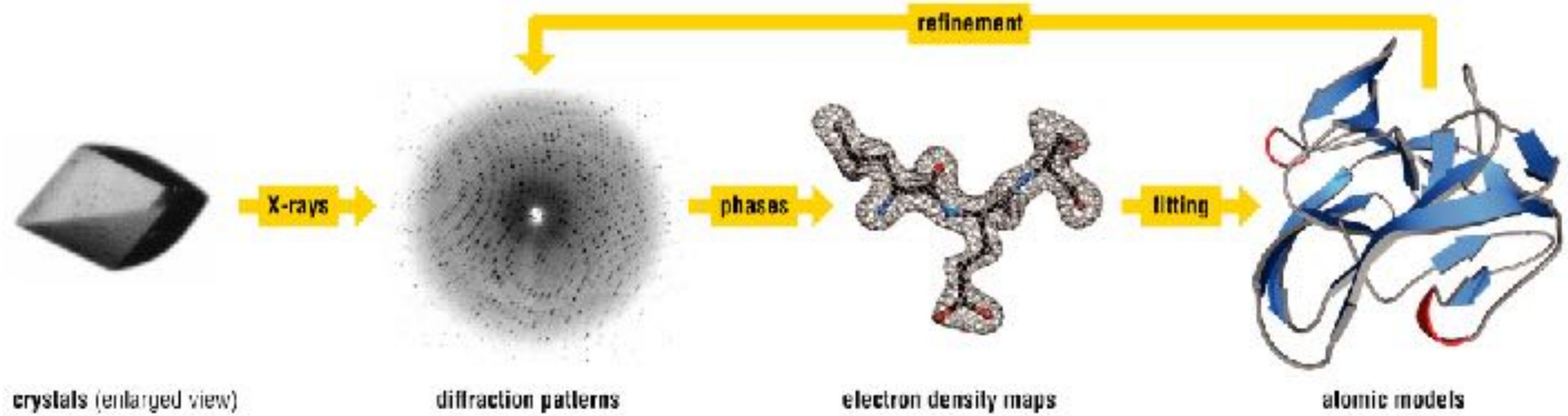
X-Ray Crystallography

Computer as Lens


X-ray Beam

crystal

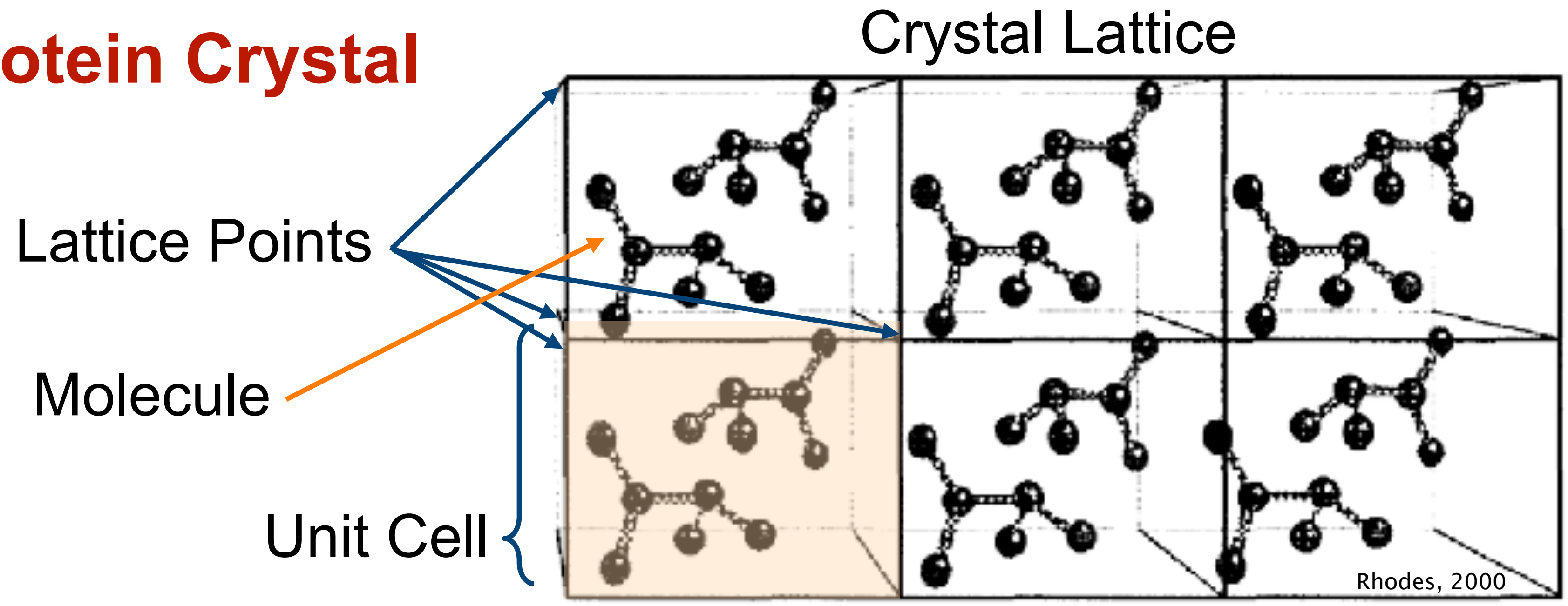
Rhodes, 2000



Petsko, Ringe, Prot Struct and Function, 2004

X-Ray Crystallography

Protein Crystal

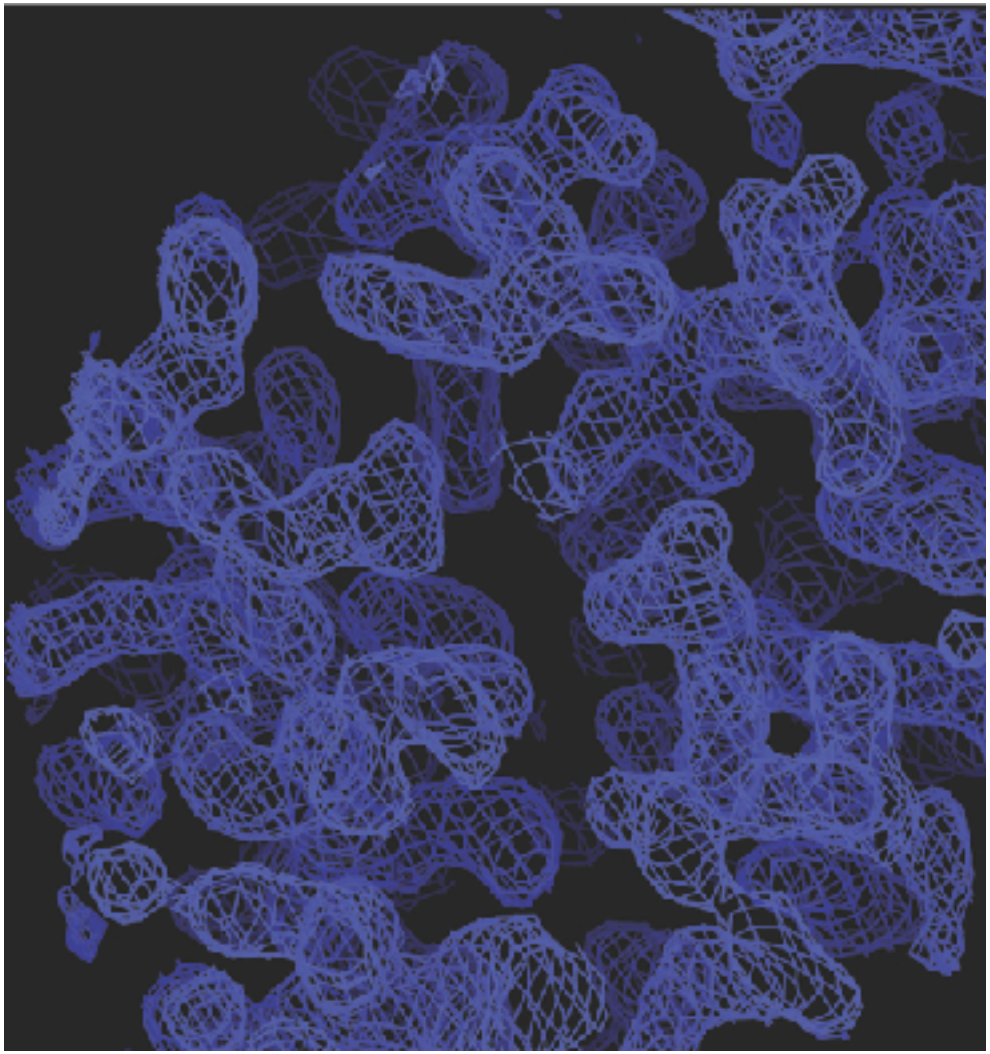


Unit cell - smallest volume element that can fully reproduce the crystal structure via translation only

Goal - determine electron density of the average unit cell

X-Ray Crystallography

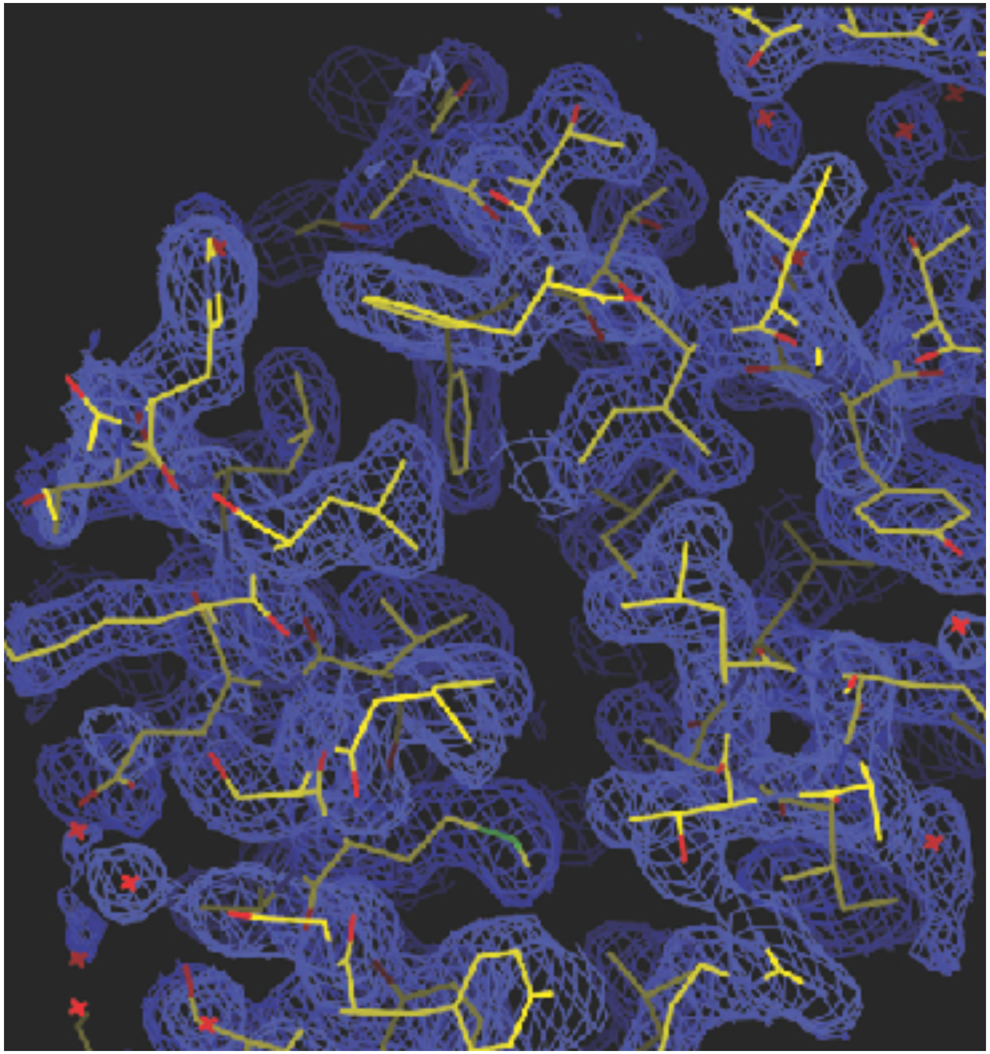
Computed electron density ...



X-Ray Crystallography

Computed electron density ...

From which we infer atomic positions



X-Ray Crystallography

1) Overview

➡ 2) Diffraction Theory

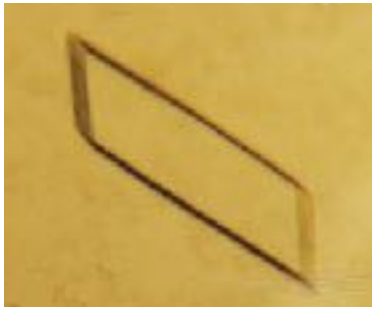
3) Protein Crystals

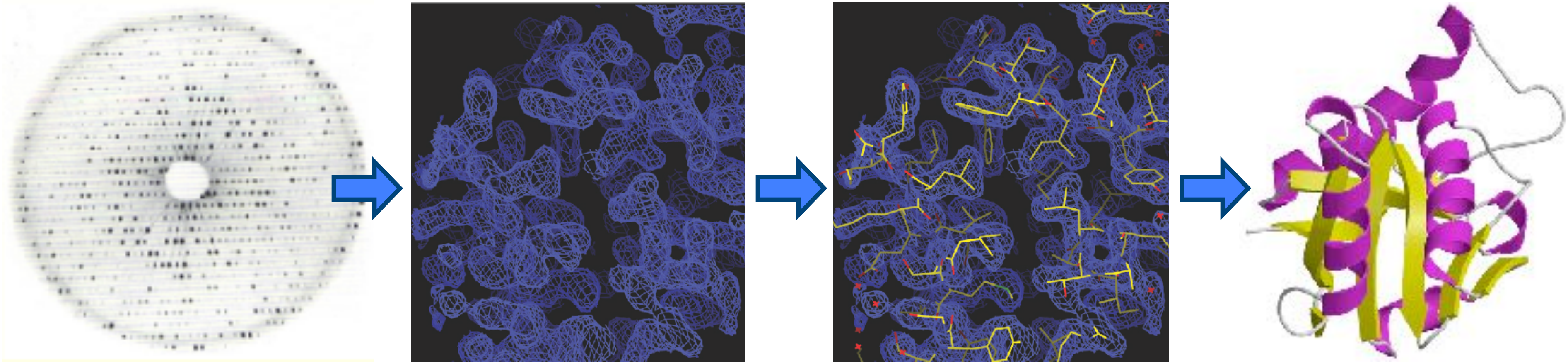
4) Collecting Diffraction Data

5) ‘Solving’ Diffraction Data - Phasing

6) Electron Density Map

7) Fitting the Map - Generating the Molecular Structure





Diffraction Data → Elect. Density Map → Fit Elect. Density Map → Structure

Diffraction Theory

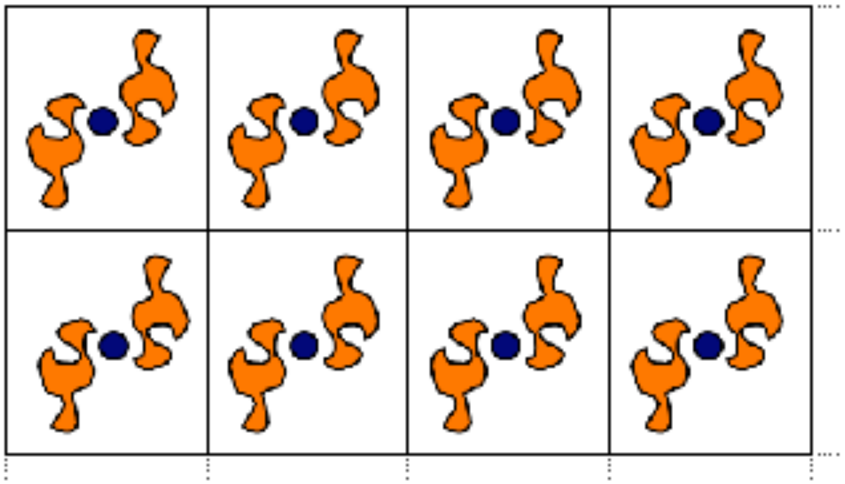
Periodic Functions / Wave Equations

$$f(x) = A\cos 2\pi(hx+\phi) \qquad f(x) = A\sin 2\pi(hx+\phi)$$

Fourier theory:

Any ***periodic function*** can be expressed as a sum of basis periodic functions (infinite sum of *sin* and *cos* terms)

In the Fourier Transform, basis functions consist of *sin* and *cos* with all possible frequencies.



We have a periodic function!

X-Ray Crystallography

If we sum over all atoms in the crystallographic unit cell:
The diffraction point observed at **S** is

Structure Factor

n: num atoms

$$\underline{\mathbf{F}(\mathbf{S})} = \sum_{j=1}^{n} f_j \exp[2\pi i \mathbf{r}_j \cdot \mathbf{S}]$$

resulting mag and phase of the wave incident on detector plate

r: position (xyz)

S: spatial frequency (hkl)
resolution

atomic scattering factor

$$f = \int_{\mathbf{r}} \rho(\mathbf{r}) \exp[2\pi i \mathbf{r} \cdot \mathbf{S}] dr$$

Although the x-rays are a single frequency, each diffraction point corresponds to a different spatial frequency.

Diffraction follows the FT of the electron density of the crystal.

X-Ray Crystallography

If we sum over all atoms in the crystallographic unit cell:
The diffraction point observed at **S** is

Structure Factor

n: num atoms

$$\underline{\mathbf{F}(\mathbf{S})} = \sum_{j=1}^{n} f_j \exp[2\pi i \mathbf{r}_j \cdot \mathbf{S}]$$

resulting mag and phase of the wave incident on detector plate

r: position (xyz)

S: spatial frequency (hkl)
resolution

atomic scattering factor

$$f = \int_{\mathbf{r}} \rho(\mathbf{r}) \exp[2\pi i \mathbf{r} \cdot \mathbf{S}] dr$$

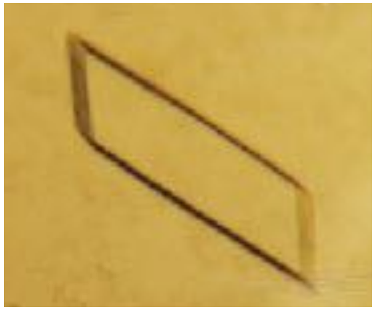
To reconstruct density:

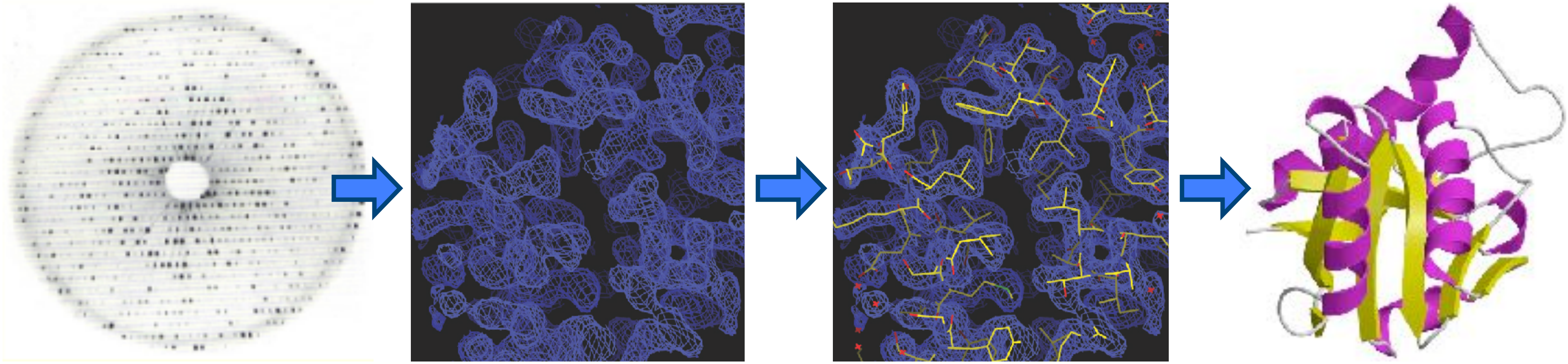
$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| e^{\alpha'_{hkl}} e^{-2\pi i(hx+ky+lz)}$$

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}| \, e^{-2\pi i(hx+ky+lz-\alpha'_{hkl})}$$

X-Ray Crystallography

1) Overview
2) Diffraction Theory
3) Protein Crystals
4) Collecting Diffraction Data
5) ‘Solving’ Diffraction Data - Phasing
6) Electron Density Map
7) Fitting the Map - Generating the Molecular Structure

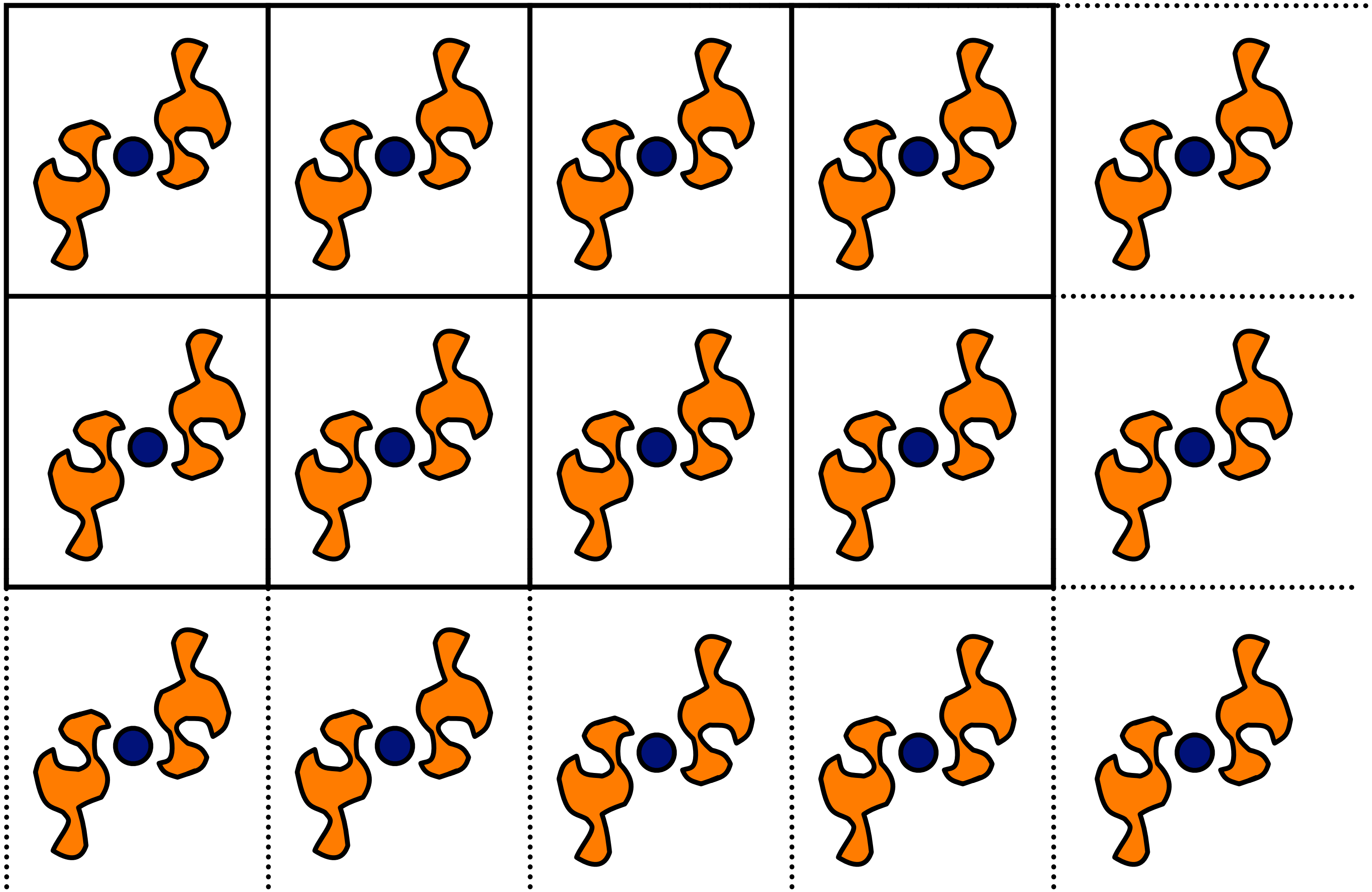




Diffraction Data Elect. Density Map Fit Elect. Density Map Structure

Crystal Growth

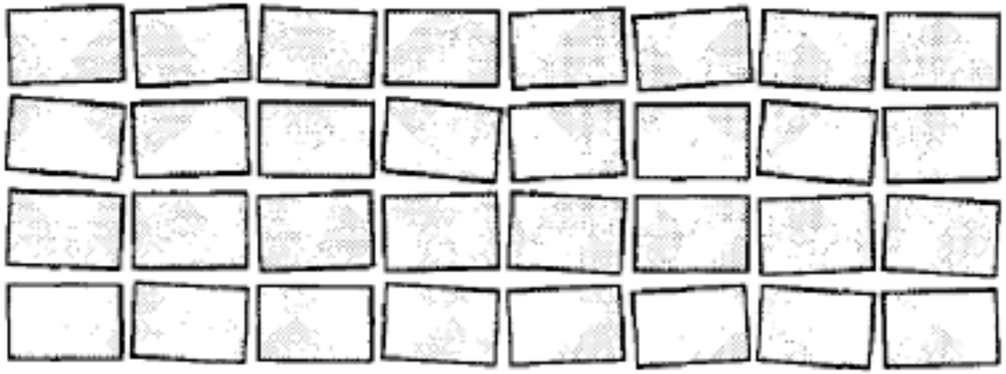
2D ‘Crystal’



Crystal Growth

Inorganic crystals (ie. NaCl) are very strong
Protein crystals held together with weaker forces
- are ***weak, fragile,*** and ***hard to grow***

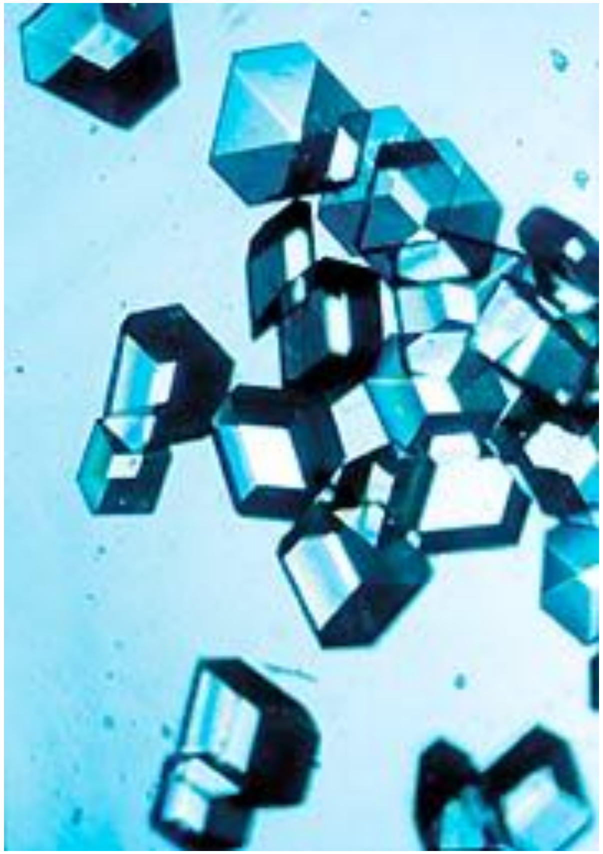
Not perfect in arrangement

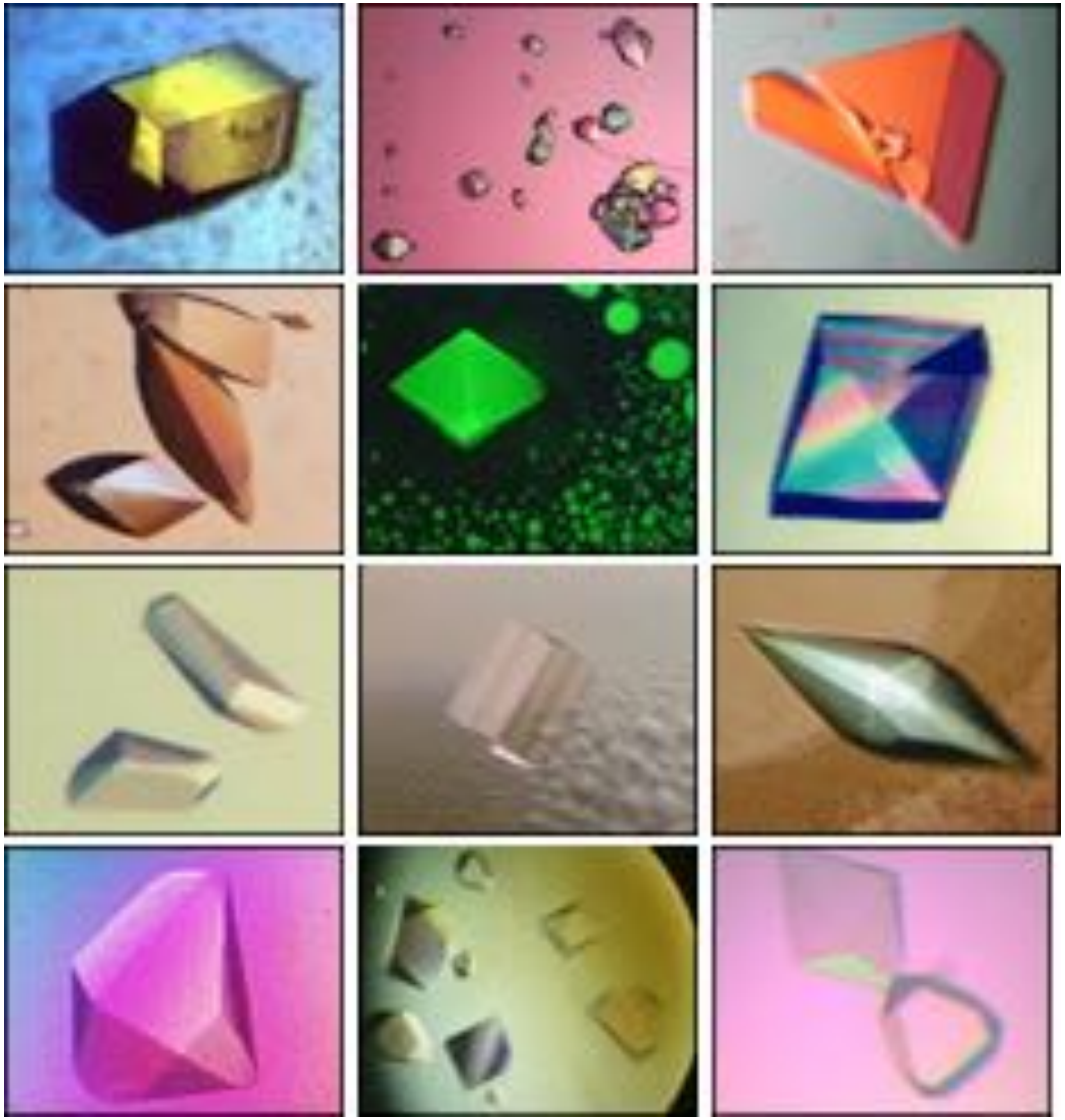


Drenth, 1994

Multiple crystals are needed / consumed in data collection
Not all crystals ‘behave’ (diffract)
May want *derivative* crystals - with ligand, cofactors, ...

Crystal Growth





Crystal Growth

Crystallization Condition Search

Essentially infinite combination of:

salts, pH-buffers, polymers, organic molecules, temperature

Trial and Error

Use of ‘Crystal Screens’ (commonly successful conditions)

Use of previous knowledge

Coarse Search followed by Fine Search

Sometimes hit is ***never*** found

First (Coarse) Screen

0.2M Calcium Chloride dihydrate, HEPES pH 7.5, 28% PEG 4000

0.2M tri-Sodium Citrate dihydrate, Tris Hydrochloride pH 8.5, 30% PEG 4000

Second (Fine) Screen

0.1M Calcium Chloride dihydrate, HEPES pH 7.5, 28% PEG 4000

0.1M Calcium Chloride dihydrate, HEPES pH 7.5, 30% PEG 4000

0.2M Calcium Chloride dihydrate, HEPES pH 7.5, 30% PEG 4000

0.3M Calcium Chloride dihydrate, HEPES pH 7.5, 28% PEG 4000

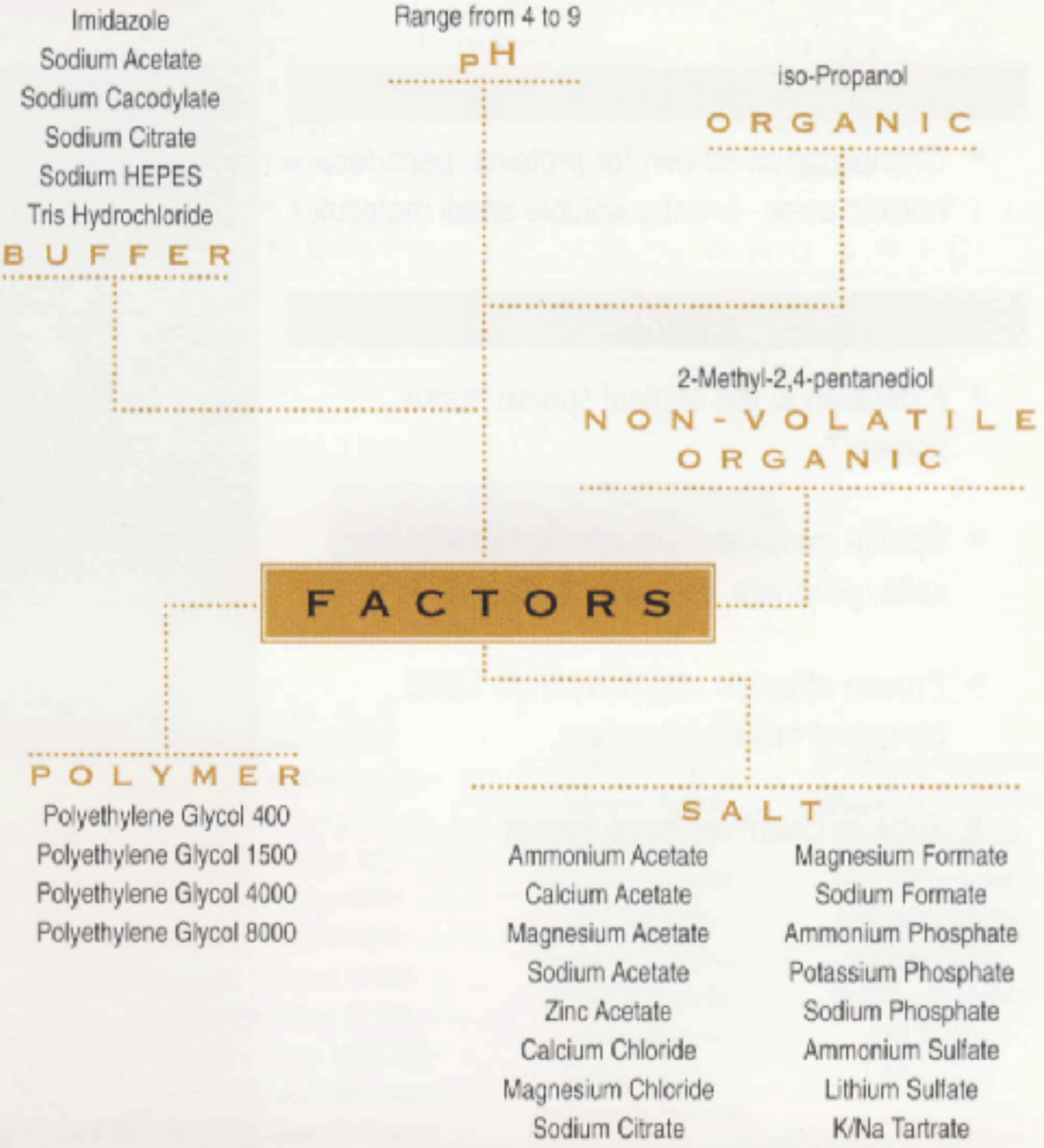
0.3M Calcium Chloride dihydrate, HEPES pH 7.5, 30% PEG 4000 → ...

‘Crystal Screen’
from Hampton Research

CRYSTAL SCREEN™ FORMULATION

Salt | Buffer | Precipitant

[illegible]



Imidazole
Sodium Acetate
Sodium Cacodylate
Sodium Citrate
Sodium HEPES
Tris Hydrochloride

BUFFER

Range from 4 to 9

pH

iso-Propanol

ORGANIC

2-Methyl-2,4-pentanediol

NON-VOLATILE ORGANIC

FACTORS

POLYMER

Polyethylene Glycol 400
Polyethylene Glycol 1500
Polyethylene Glycol 4000
Polyethylene Glycol 8000

SALT

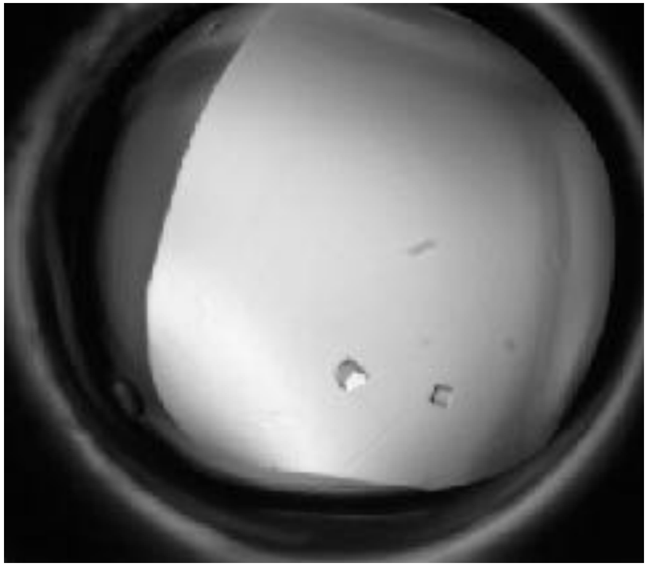
Ammonium Acetate
Calcium Acetate
Magnesium Acetate
Sodium Acetate
Zinc Acetate
Calcium Chloride
Magnesium Chloride
Sodium Citrate

Magnesium Formate
Sodium Formate
Ammonium Phosphate
Potassium Phosphate
Sodium Phosphate
Ammonium Sulfate
Lithium Sulfate
K/Na Tartrate

Crystal Growth

Robots and Automation

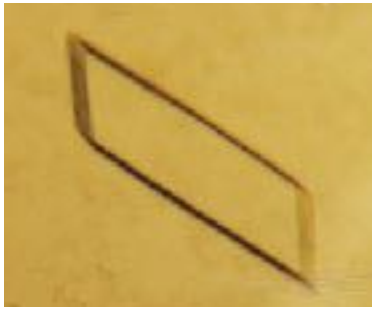
Robots for **Cloning** (ie. getting your gene into a bacteria)
Robots for **Bacterial growth** and **Protein Expression**
Robots for **Protein Purification**
Robots for **Crystallization**
Robots for **Imaging** (crystal detection)

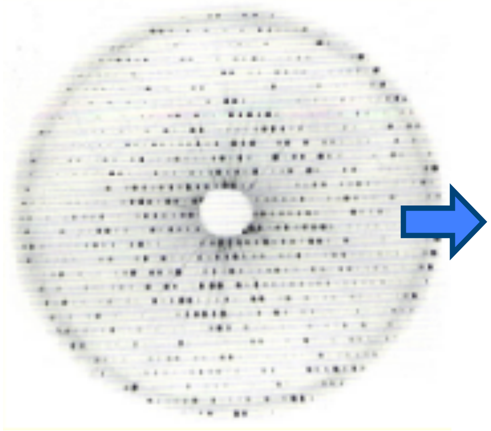


porter.llnl.gov

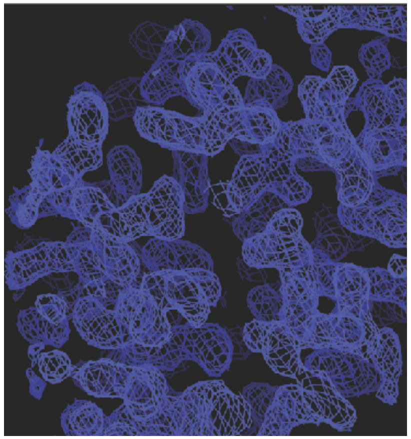
X-Ray Crystallography

1) Overview
2) Diffraction Theory
3) Protein Crystals
4) ➡ Collecting Diffraction Data
5) ‘Solving’ Diffraction Data - Phasing
6) Electron Density Map
7) Fitting the Map - Generating the Molecular Structure

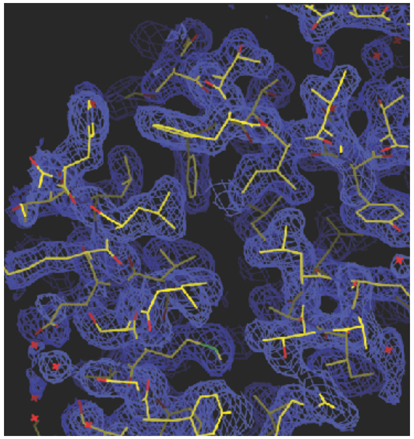




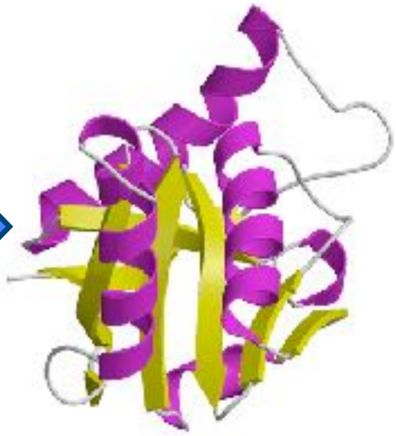
Diffraction Data



Elect. Density Map



Fit Elect. Density Map



Structure

X-Ray Sources

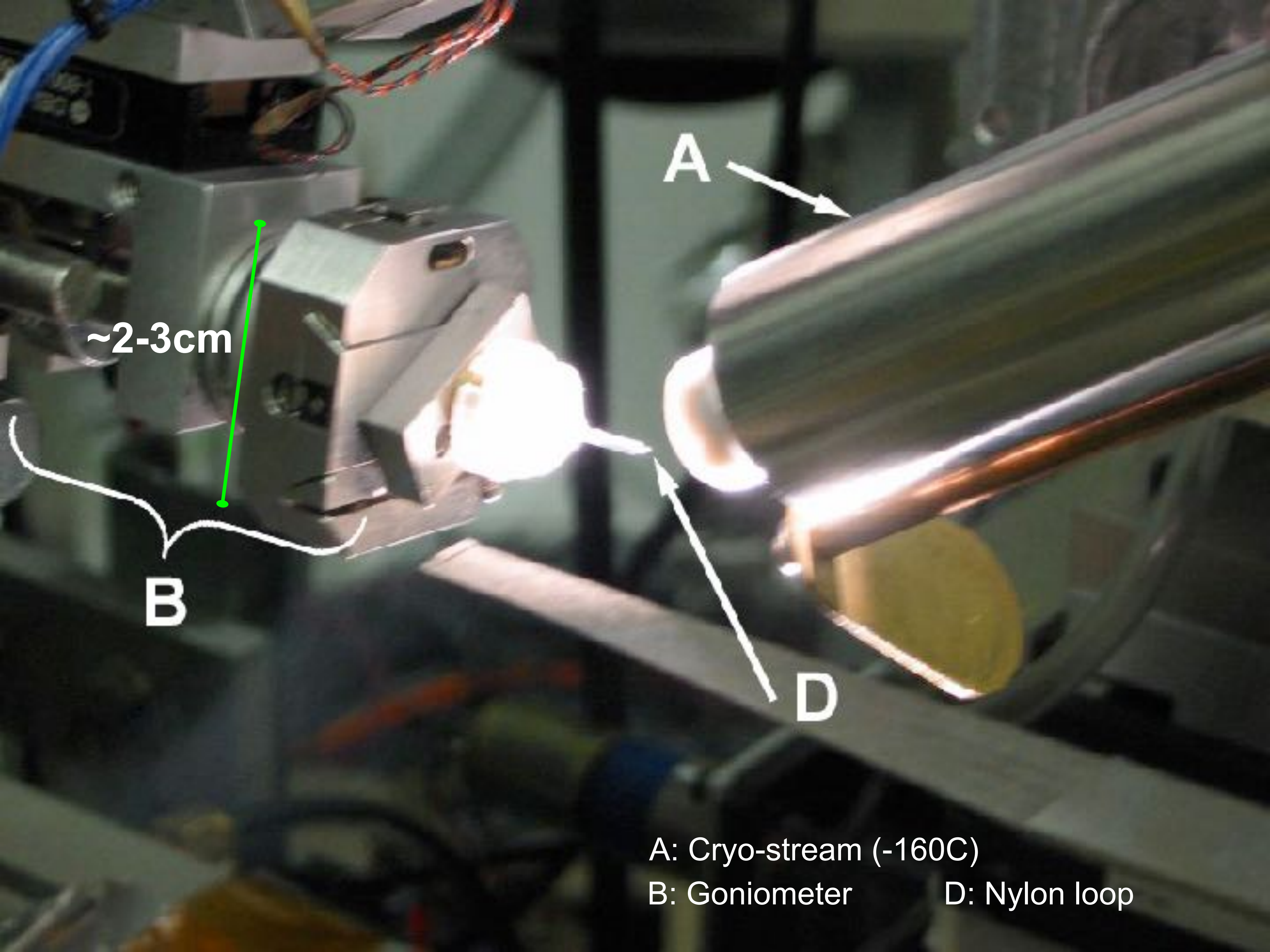
Requires high-energy X-ray source

- -home sources
- -synchrotron (particle accelerators)

Wavelengths: 0.6A - 1.5A

Advanced Photon Source at Argonne (Illinois, USA)

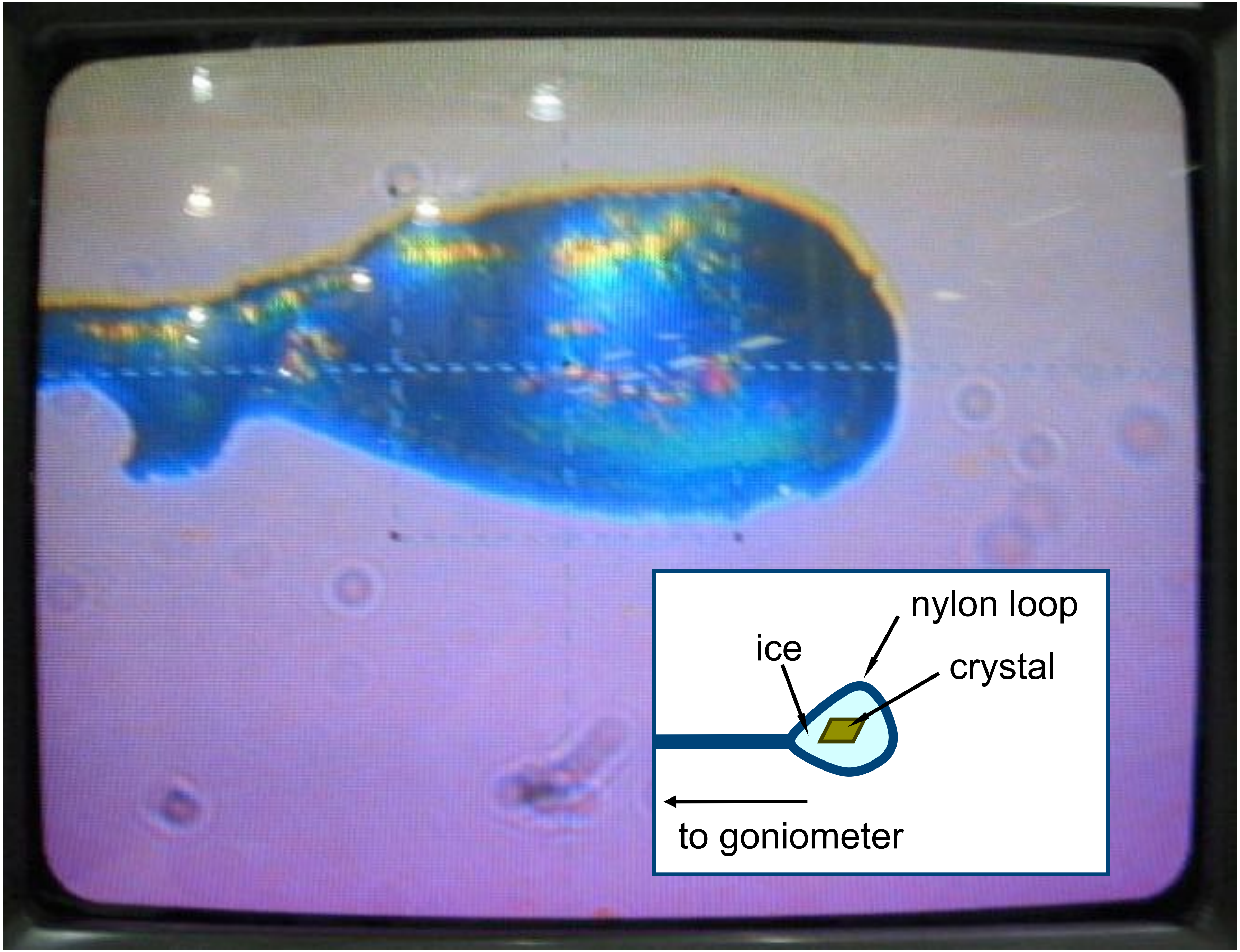
Appx. 3km around



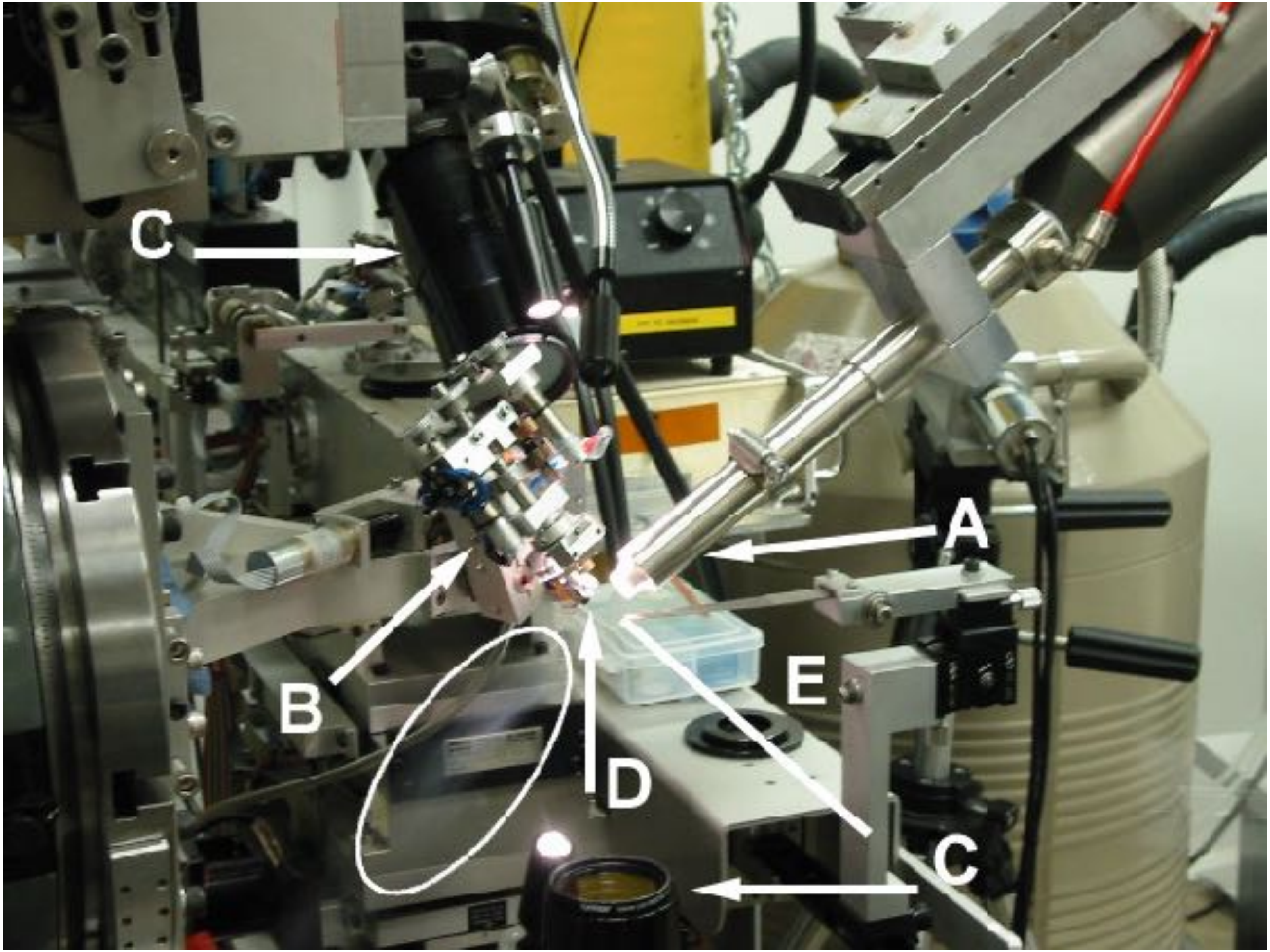
A: Cryo-stream (-160C)

B: Goniometer

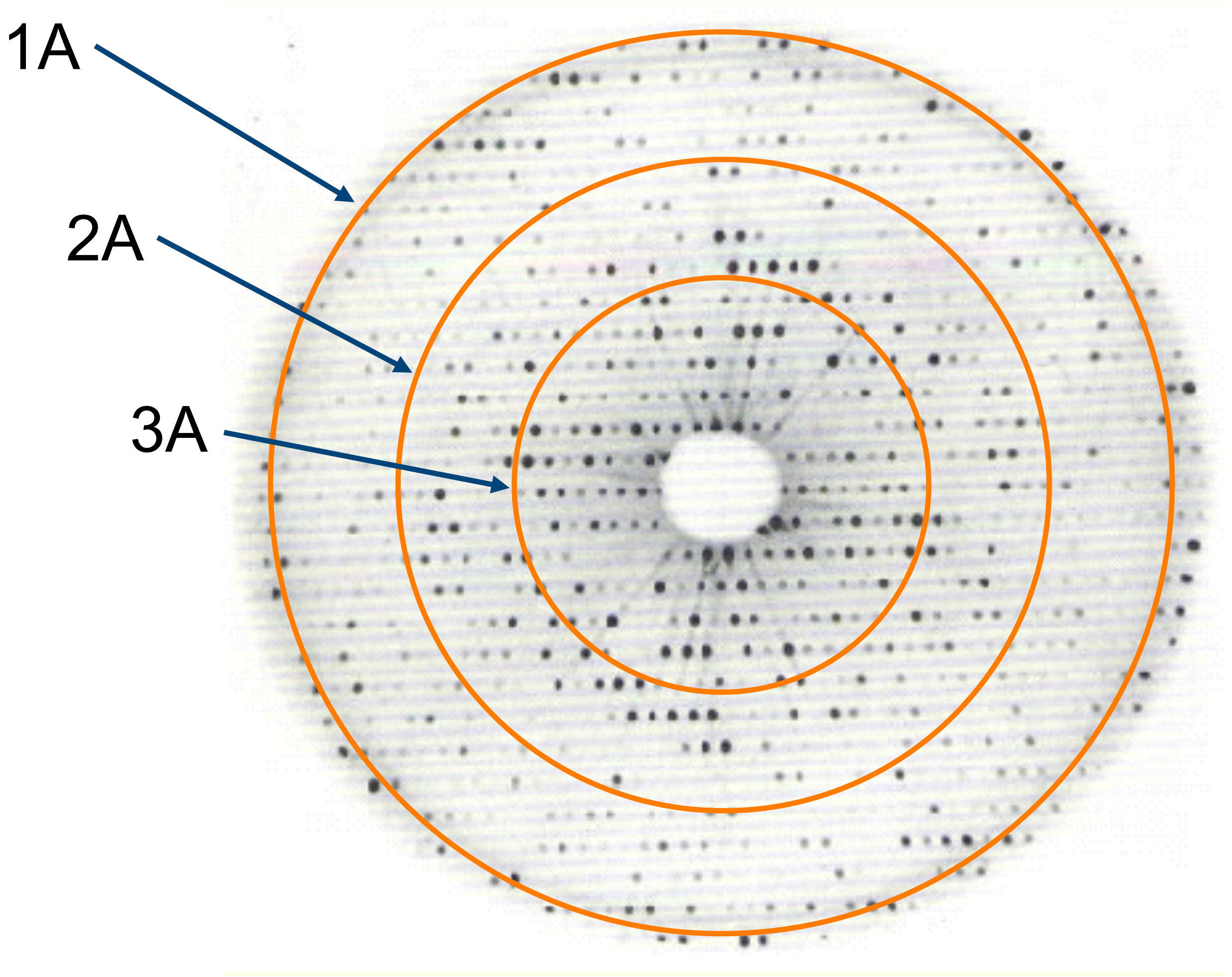
D: Nylon loop


nylon loop
ice
crystal
to goniometer

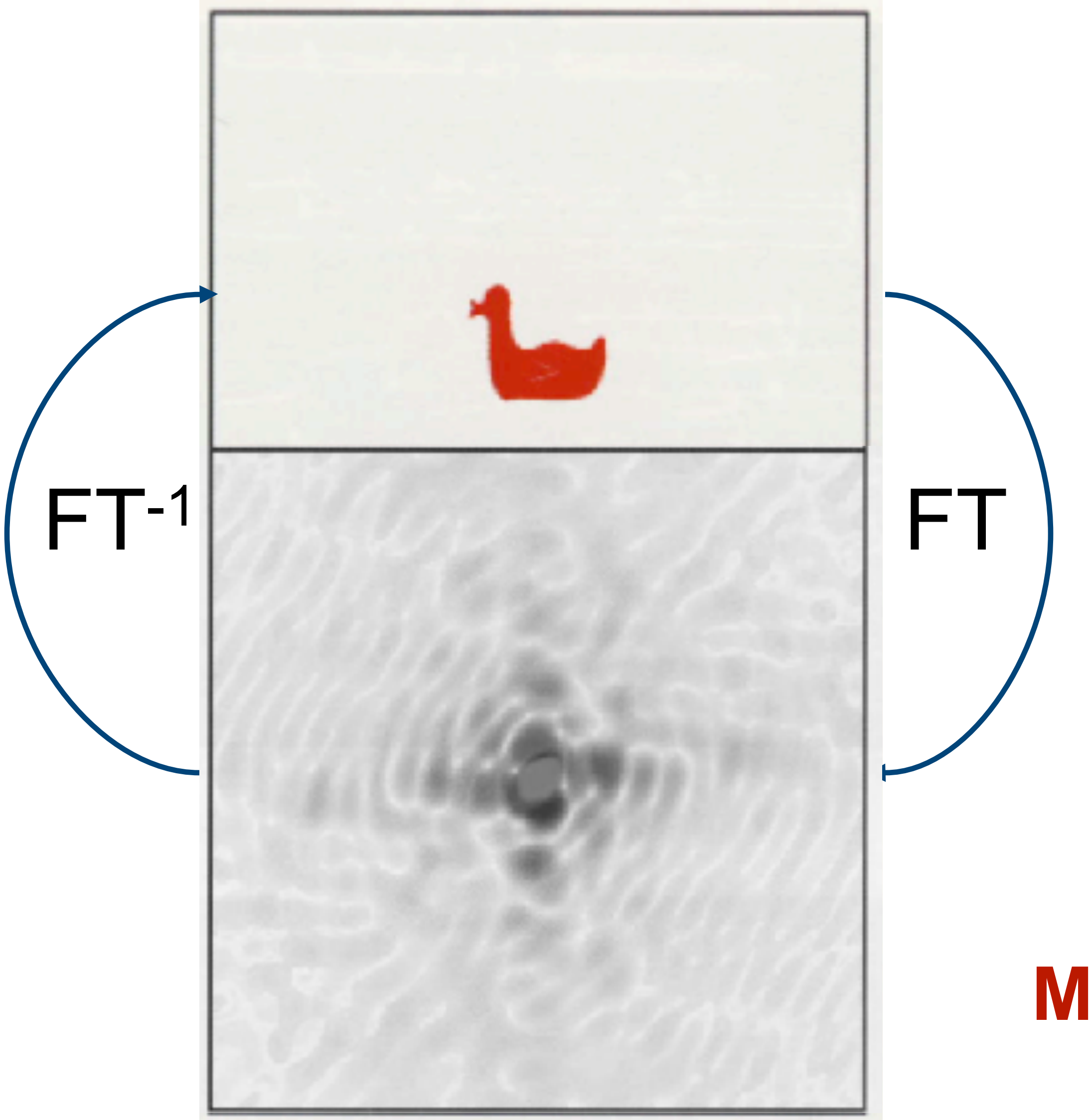
Setup



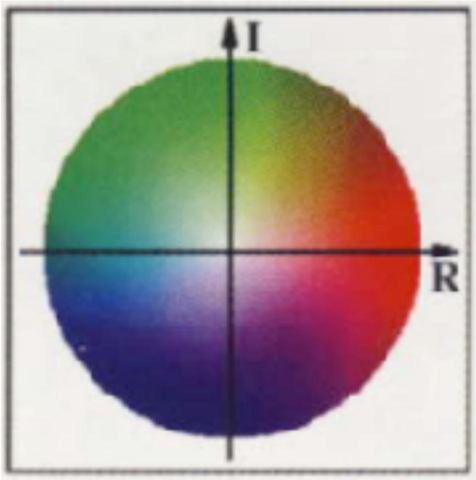
Diffraction



Diffraction



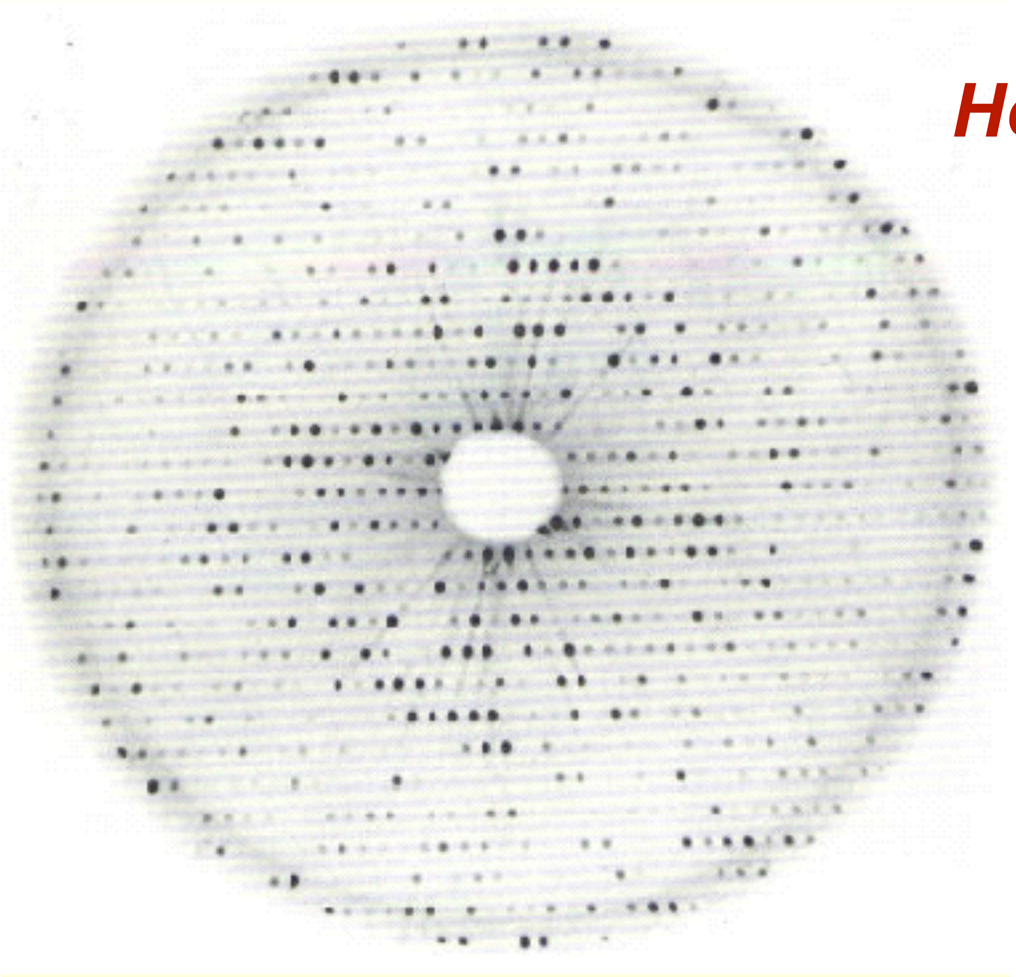
a **Duck and duck FT**



Missing Phases!!!

Diffraction

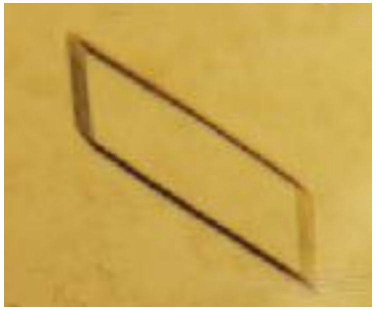
Amplitudes only!

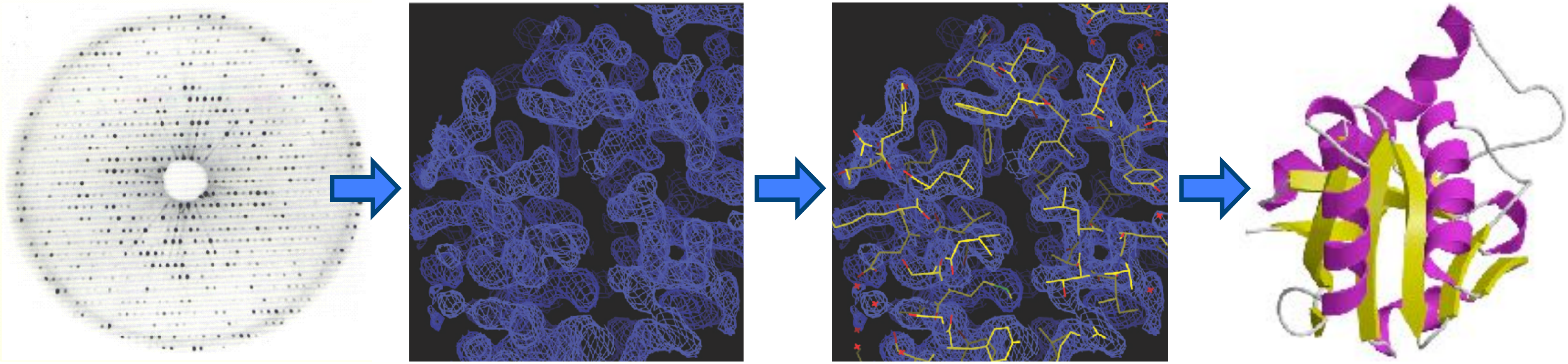


How to determine phases?

X-Ray Crystallography

1) Overview
2) Diffraction Theory
3) Protein Crystals
4) Collecting Diffraction Data
5) ➡ ‘Solving’ Diffraction Data - Phasing
6) Electron Density Map
7) Fitting the Map - Generating the Molecular Structure



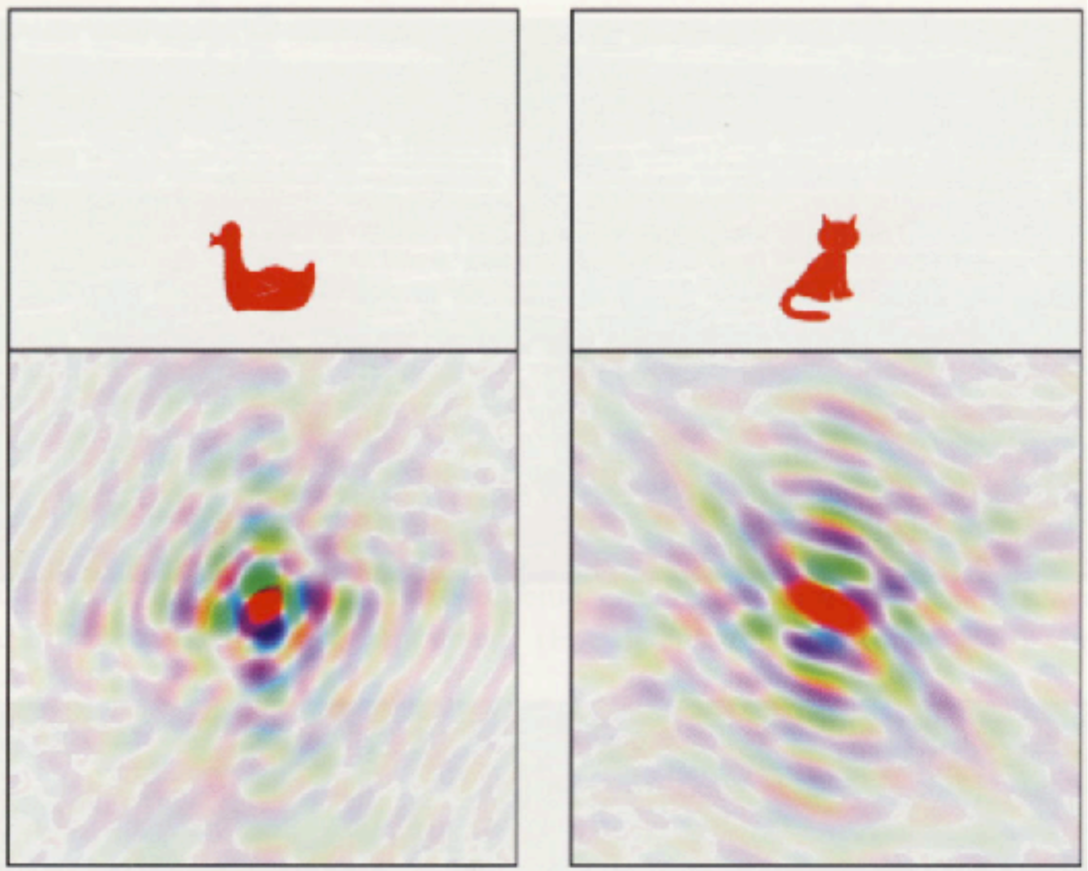


Diffraction Data   Elect. Density Map   Fit Elect. Density Map   Structure

Phasing - MR

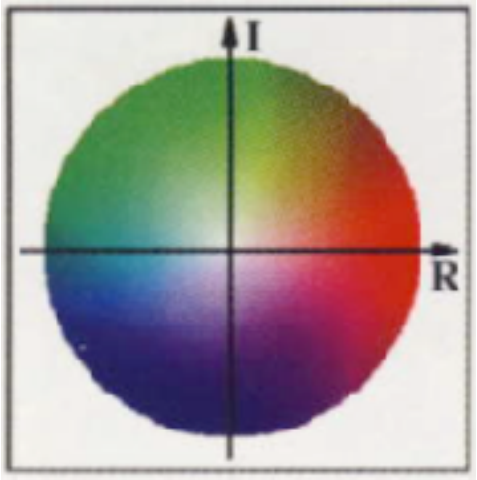
Molecular Replacement (MR)

Bootstrap phase determination using phases from homologous structure

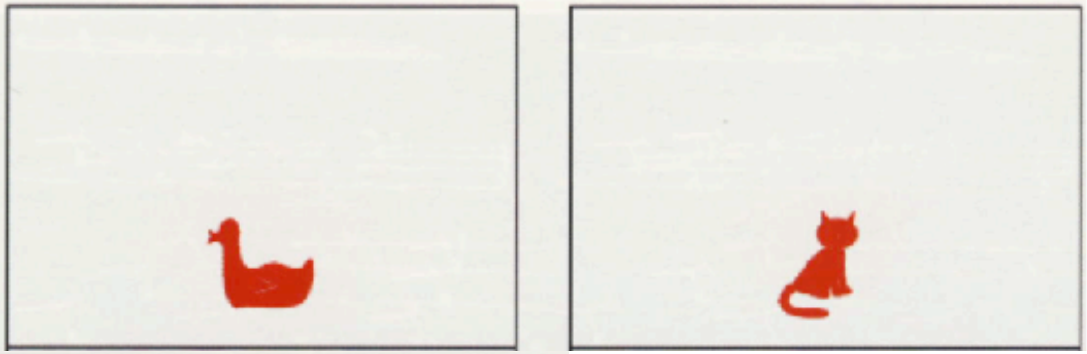


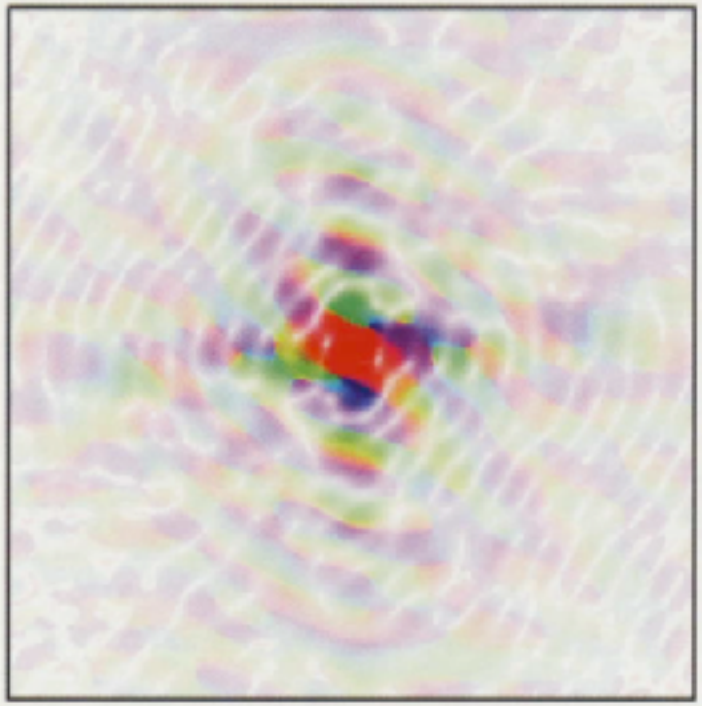
a Duck and duck FT

b Cat and cat FT



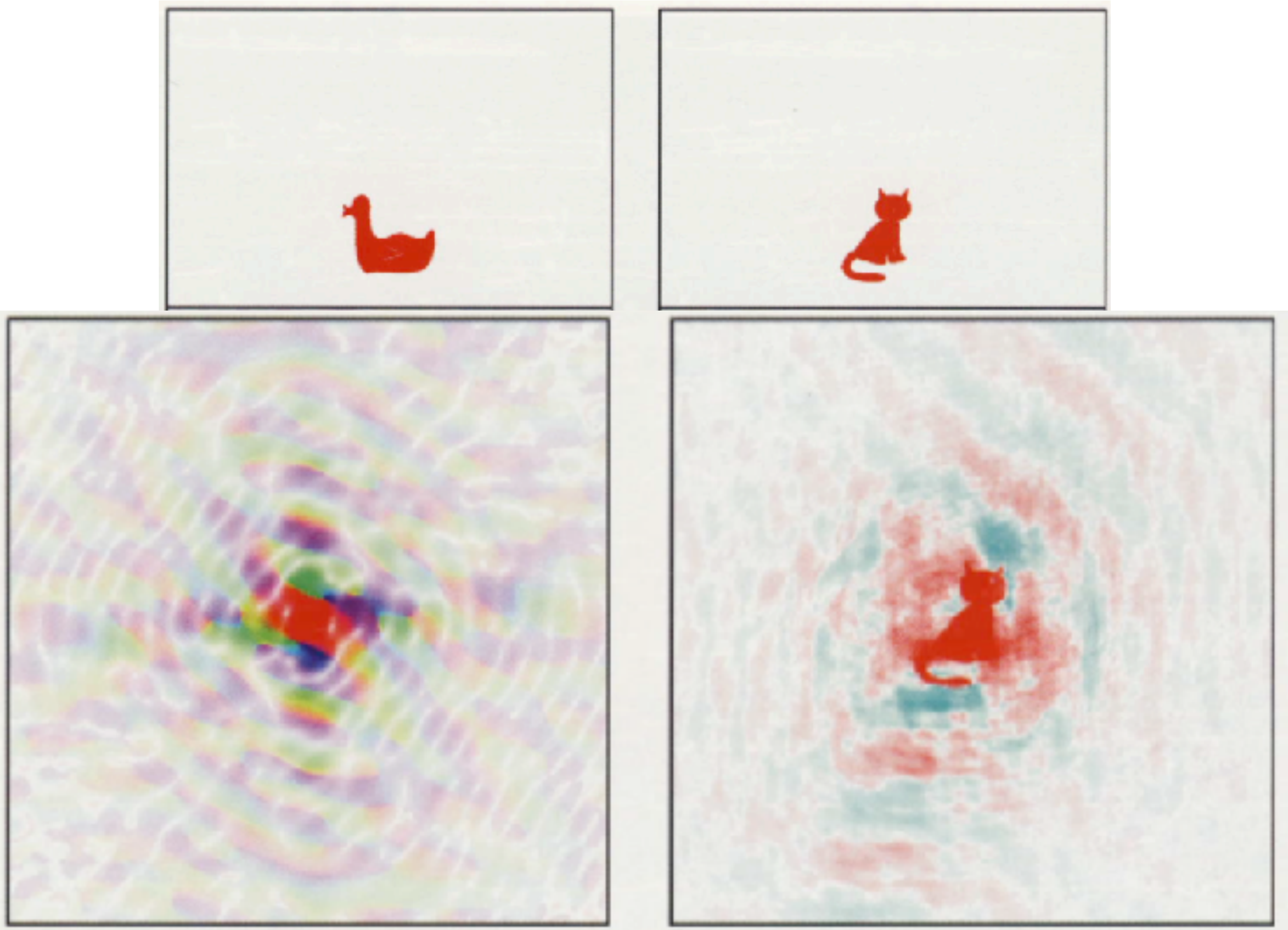
Phasing - MR





c **Duck intensities and cat phases**

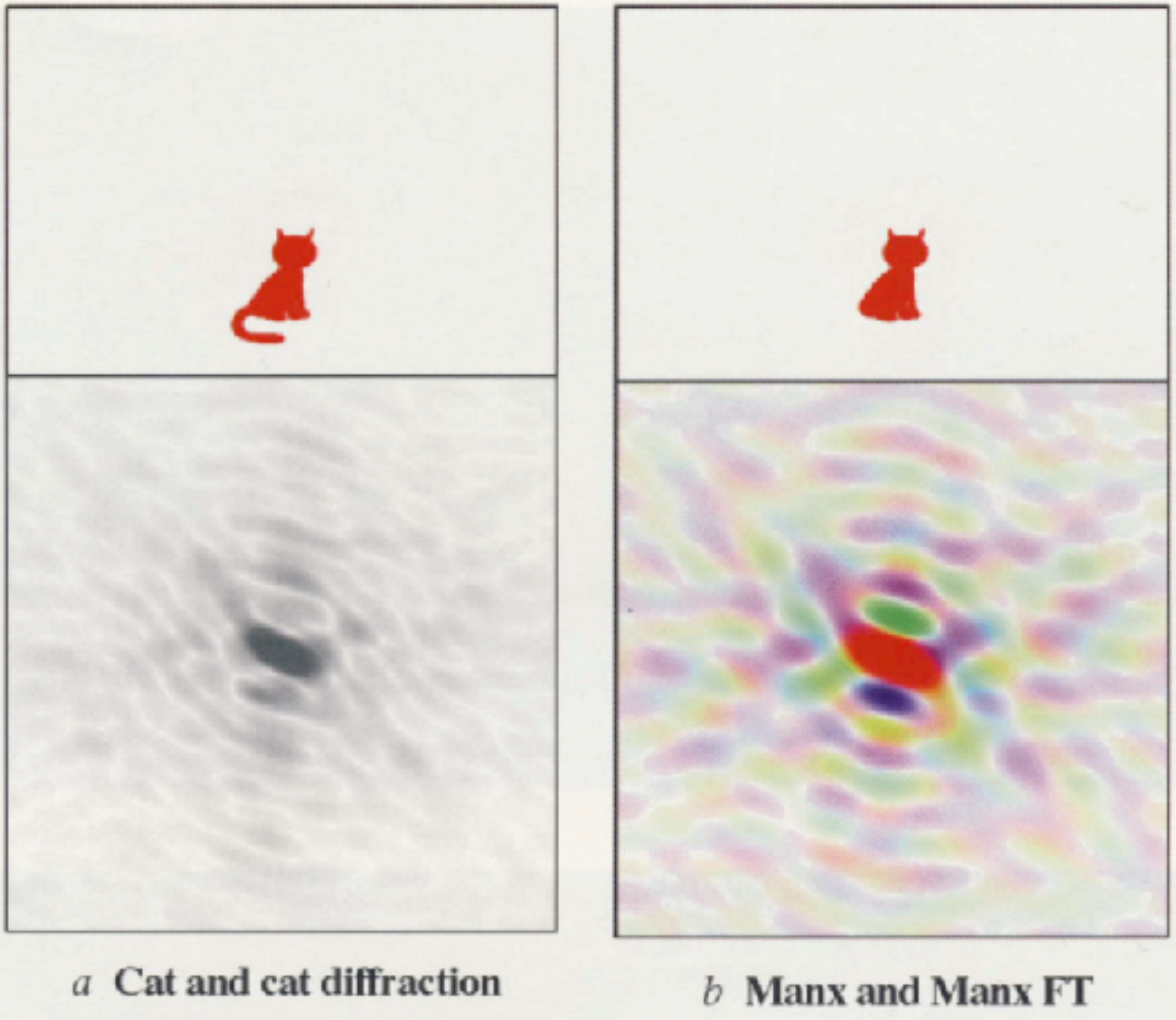
Phasing - MR



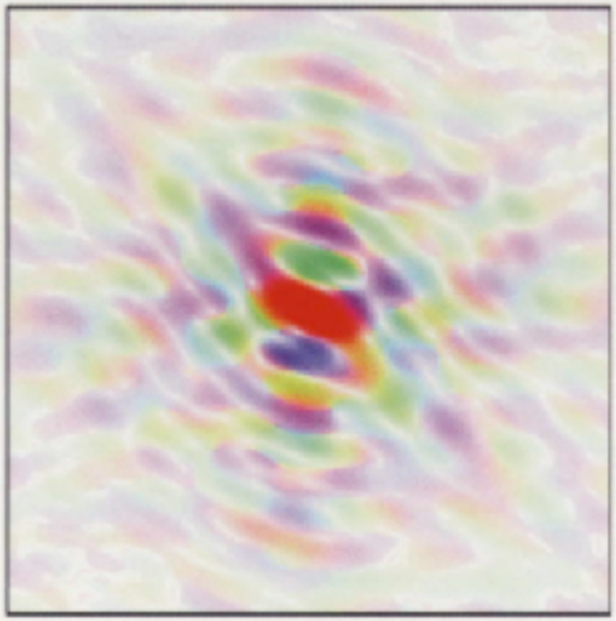
c **Duck intensities and cat phases**

d **Back-transform of *c***

Phasing - MR

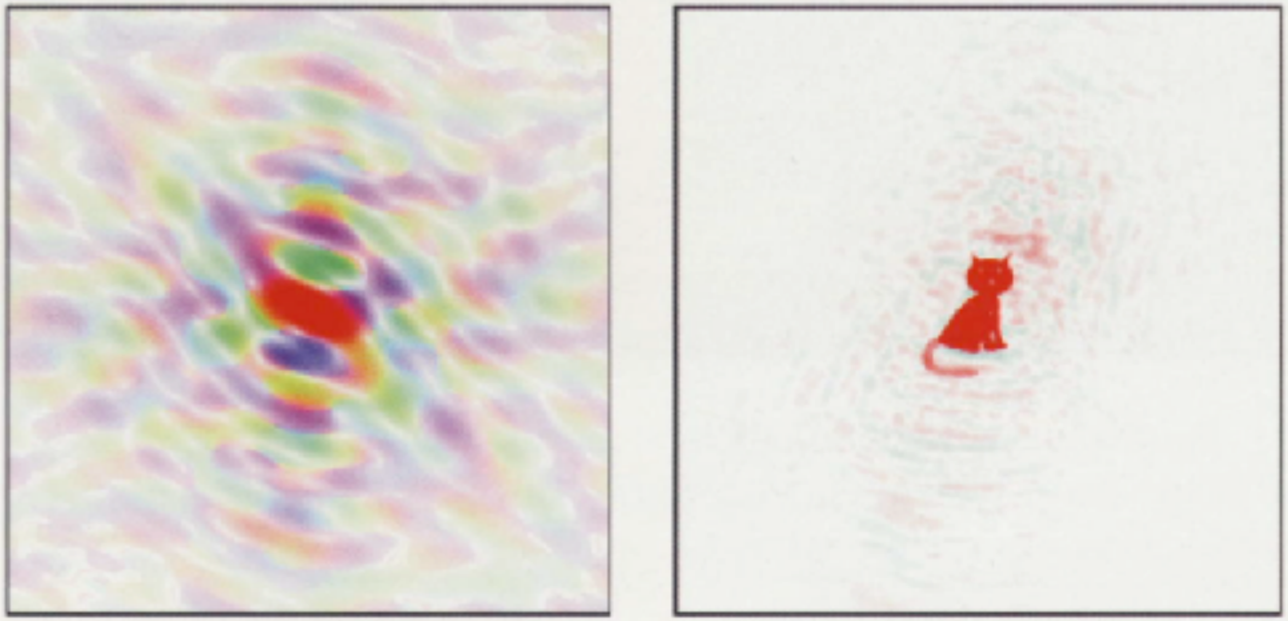


Phasing - MR



c Cat intensities with Manx phases

Phasing - MR

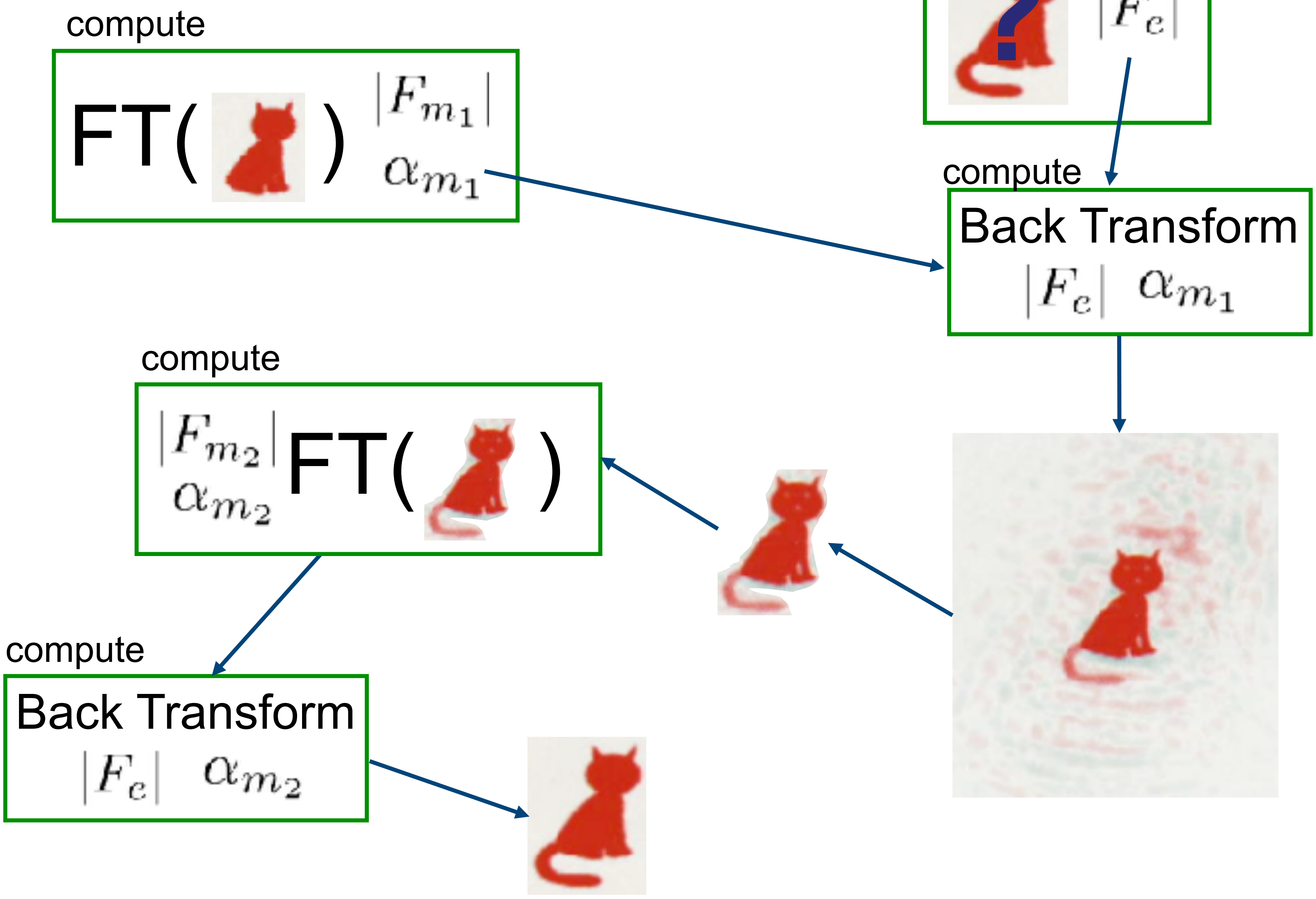


c **Cat intensities with Manx phases**

d **Back-transform of** *c*

Model

Phasing - MR



compute

FT( ) $|F_{m_1}|$ α_{m_1}

Experiment

$|F_c|$

compute

Back Transform

$|F_c|$ α_{m_1}

compute

$|F_{m_2}|$ α_{m_2} FT( )

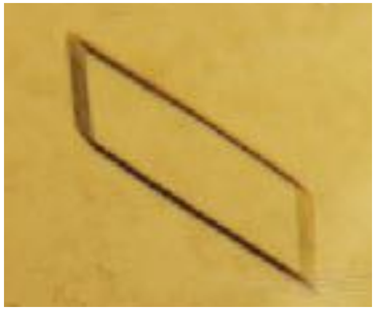
compute

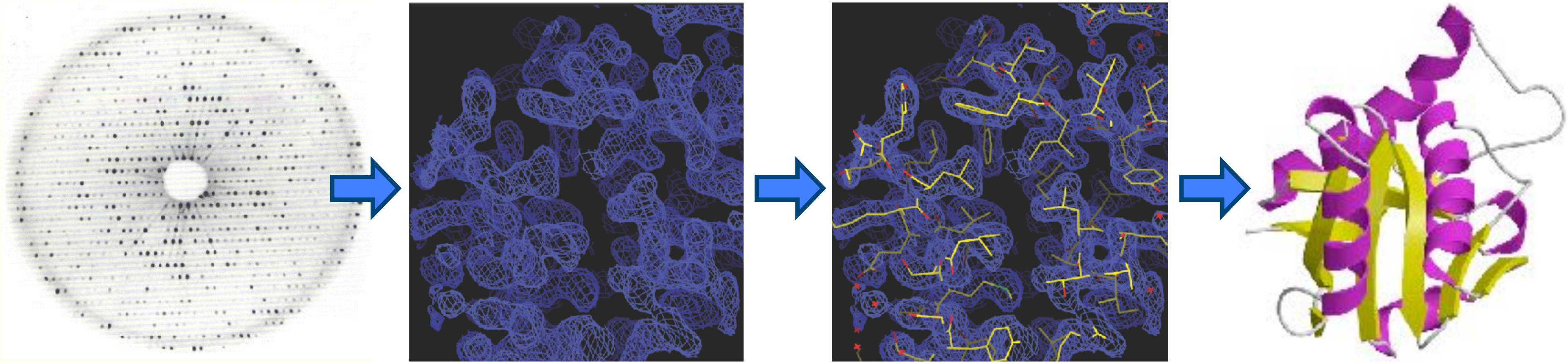
Back Transform

$|F_c|$ α_{m_2}

X-Ray Crystallography

1) Overview
2) Diffraction Theory
3) Protein Crystals
4) Collecting Diffraction Data
5) ‘Solving’ Diffraction Data - Phasing
6) Electron Density Map
7) Fitting the Map - Generating the Molecular Structure





Diffraction Data   Elect. Density Map   Fit Elect. Density Map   Structure

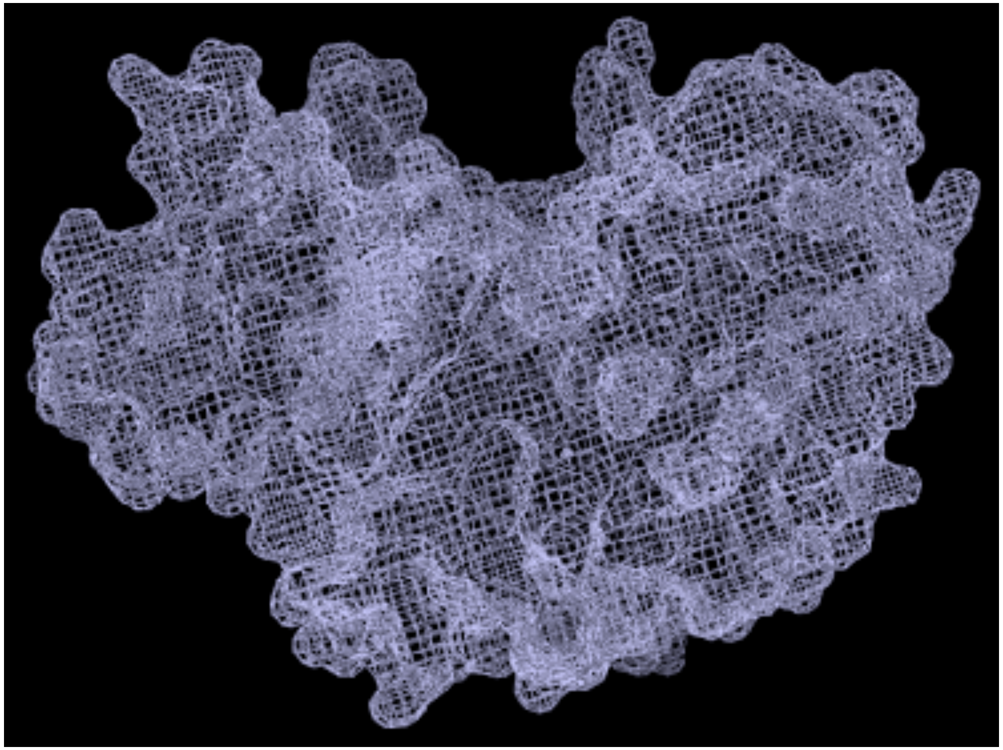
Electron Density Map

We have initial phase estimates, with confidences

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l w_{hkl} \left|F_{\text{obs}}\right| e^{-2\pi i(hx+ky+lz-\alpha'_{\text{calc}})}$$

Initially we will only have confidence in low frequency / resolution terms

molecular envelope



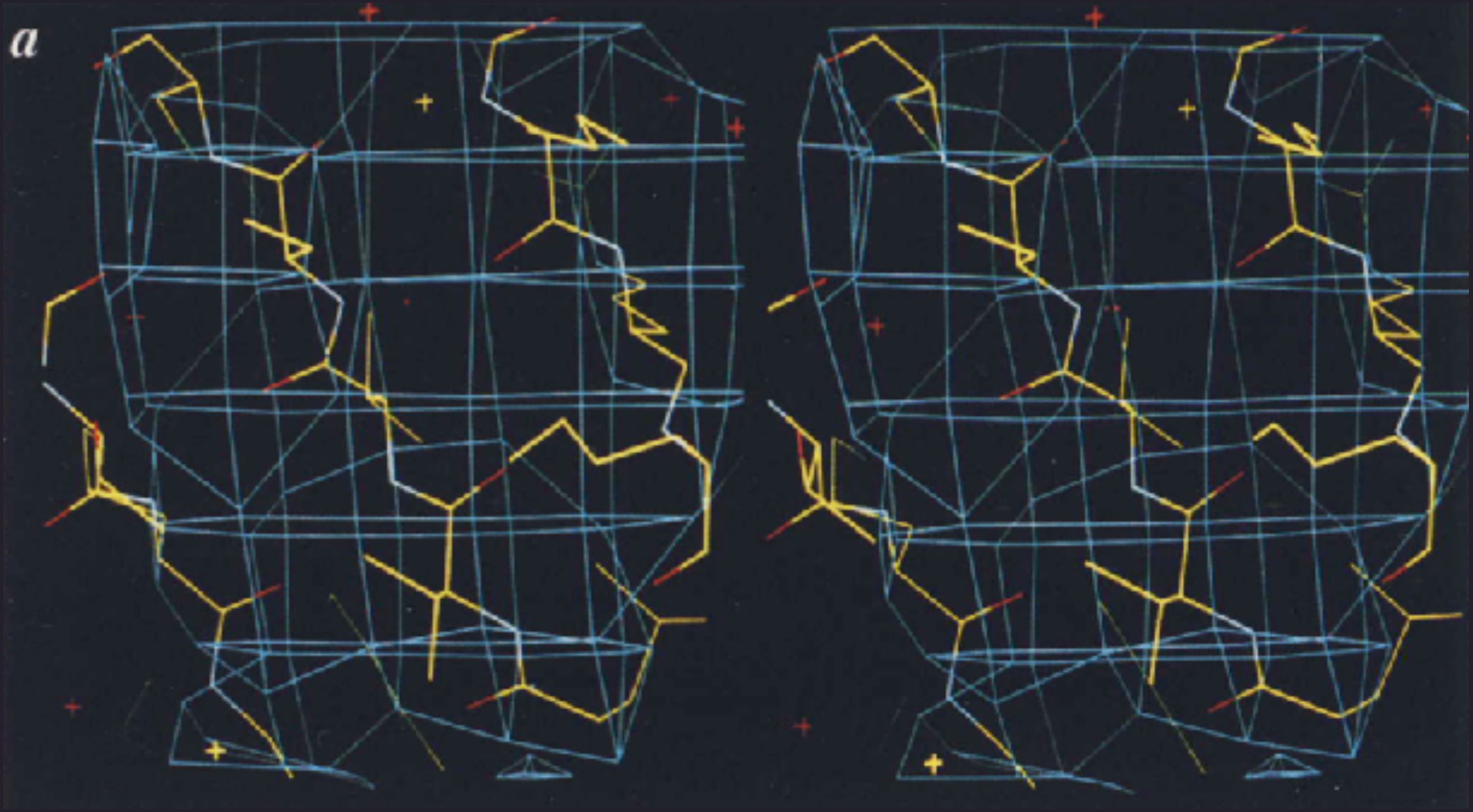
Improve Map

If $\rho(x,y,z) < 0$
then $\rho(x,y,z) = 0$

Increase overall density to expected density

New density to recompute phases

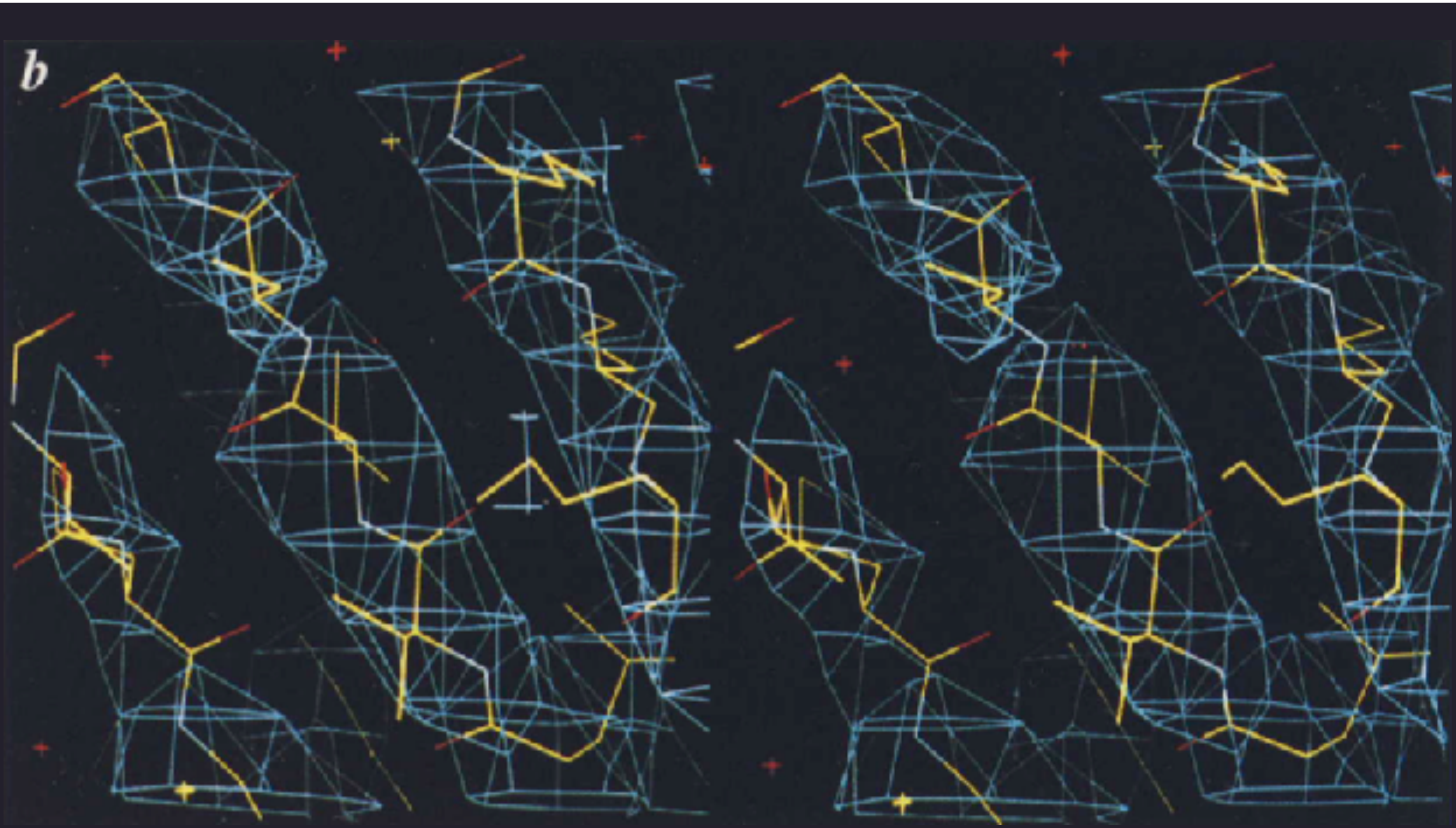
Electron Density Map



Rhodes, 2000

Series truncated at 6.0 A

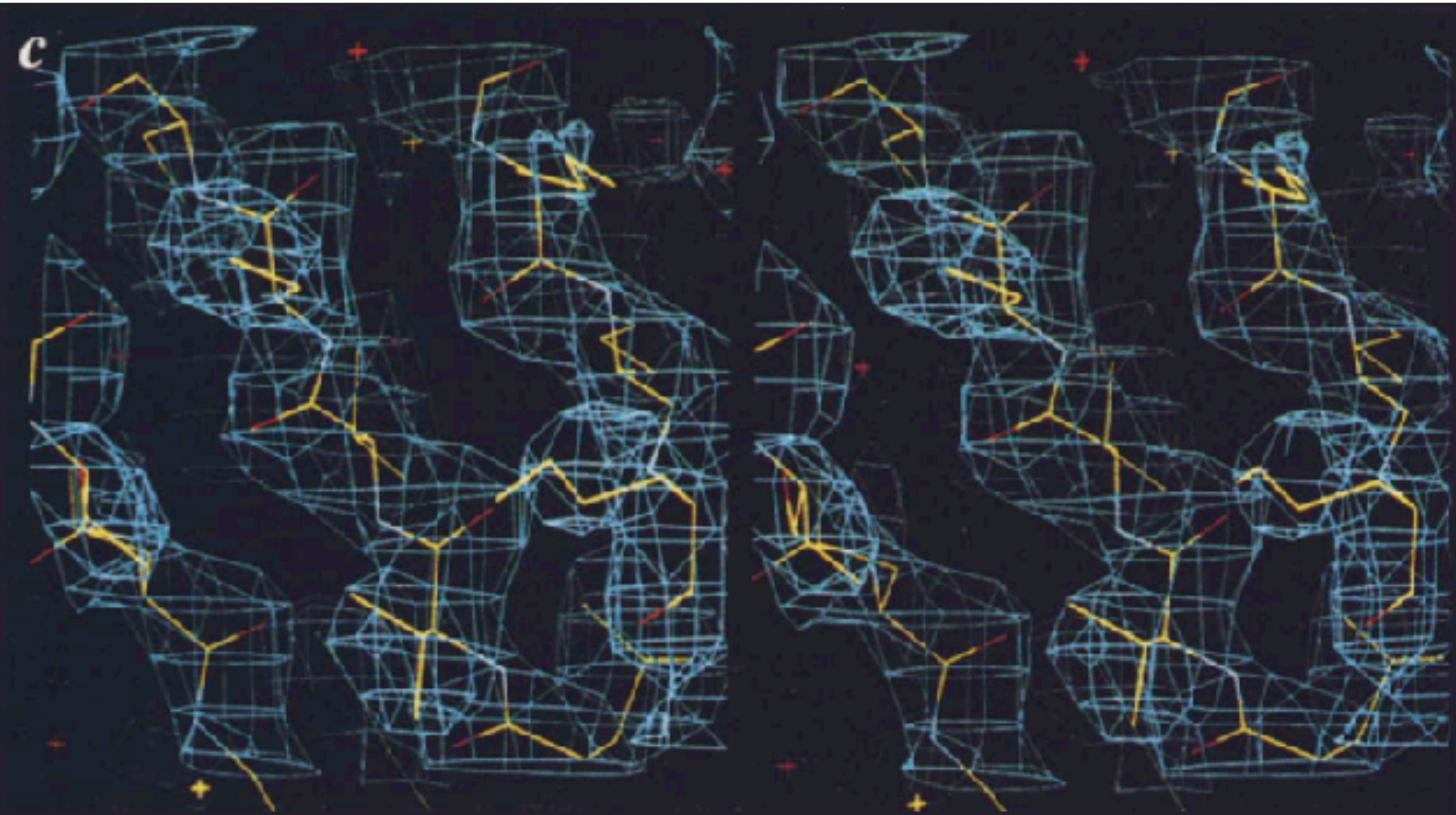
Electron Density Map



Rhodes, 2000

Series truncated at 4.5 A

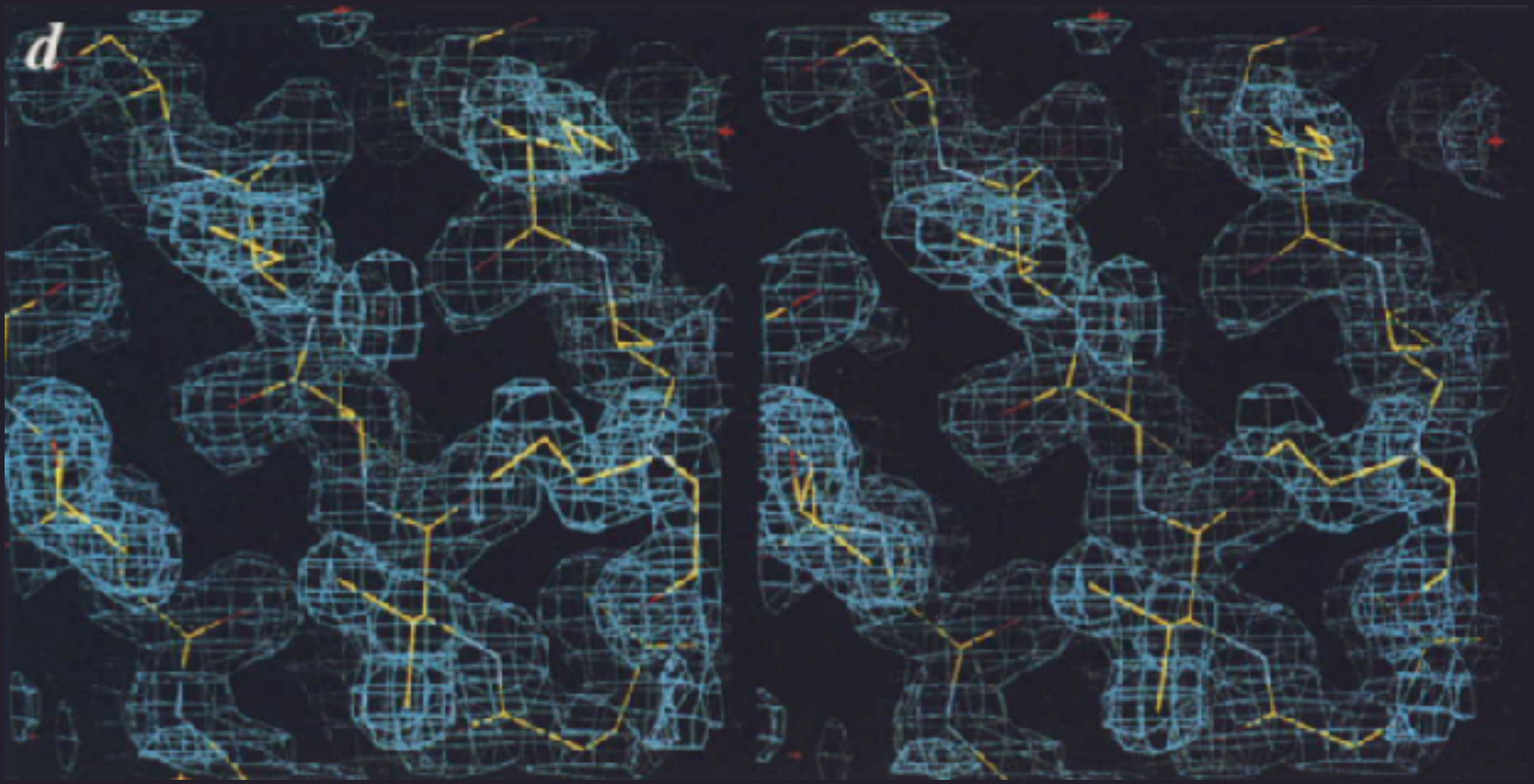
Electron Density Map



Rhodes, 2000

Series truncated at 3.0 A

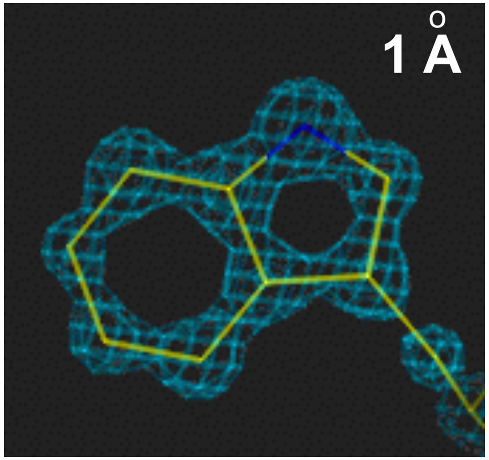
Electron Density Map

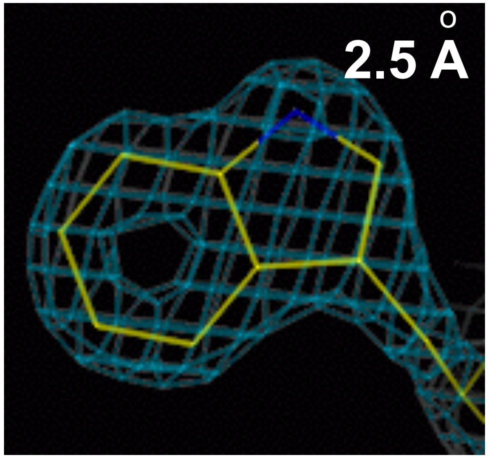


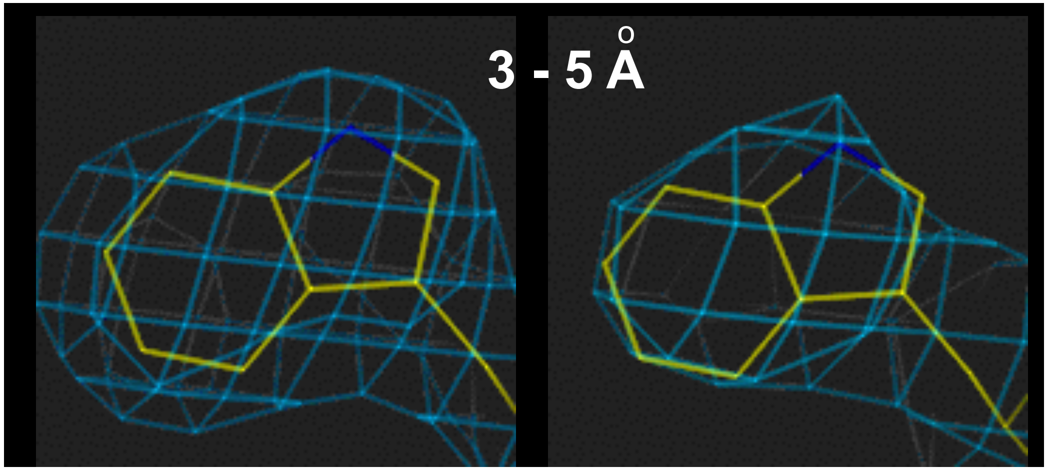
Rhodes, 2000

Series truncated at 1.6 A

Electron Density Map - Tryptophan

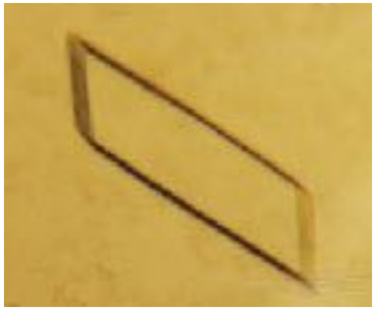


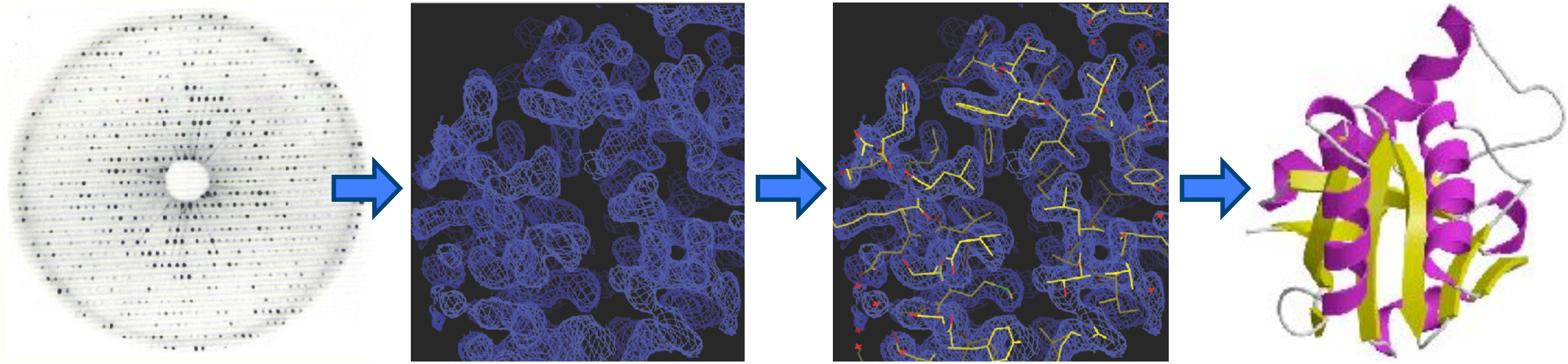




X-Ray Crystallography

1) Overview
2) Diffraction Theory
3) Protein Crystals
4) Collecting Diffraction Data
5) ‘Solving’ Diffraction Data - Phasing
6) Electron Density Map
7) Fitting the Map - Generating the Molecular Structure





Diffraction Data

Elect. Density Map

Fit Elect. Density Map

Structure

Fitting

Phase Extension

Increasing confidence of phases

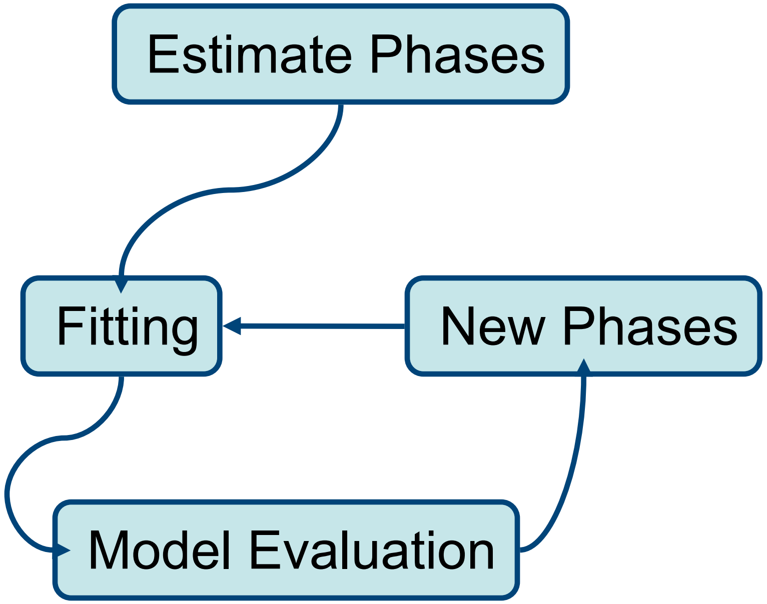
Iterative incorporation of higher resolution terms

Iterative model building and refinement

Use of difference maps ($\mathbf{F}_0$-$\mathbf{F}_c$)

Molecular Replacement Model

Serves as starting point for manual manipulation (changing A -> B)



No Model?

Build from scratch

Fitting / Refinement - Typical

1) Early Fittings Often Done Manually

First trace - disconnected, fragments, low resolution

Ridge lines - through regions of maximum density - backbone?

2) Build Backbone from Trace (find C_{alpha})

3) Align Sequence to the Trace

Find landmarks (ie. characteristic AAs)

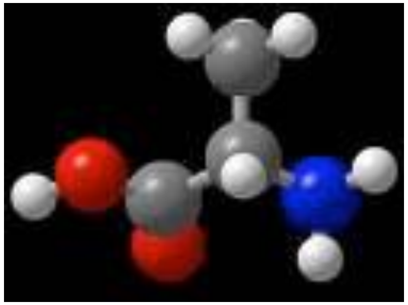
4) Place Side-Chains

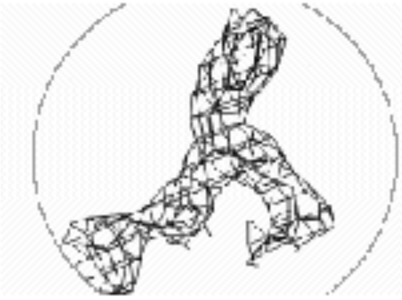
5) Adjust (refine) Structure

atoms ~4A apart, near the center of the main-chain next to bulges representing side-chains

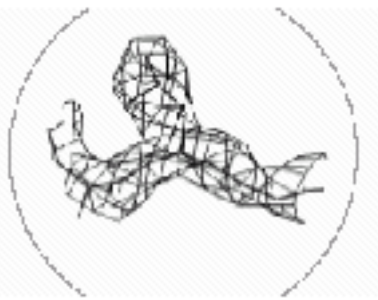
Poly-Alanine

if unknown

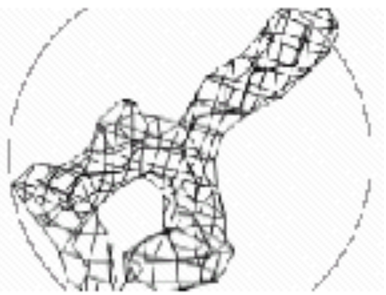




Phe



Leu



Lys

ARP/wARP

Given an Initial Electron Density Map

- Refine phases
- Build a protein model (structure)

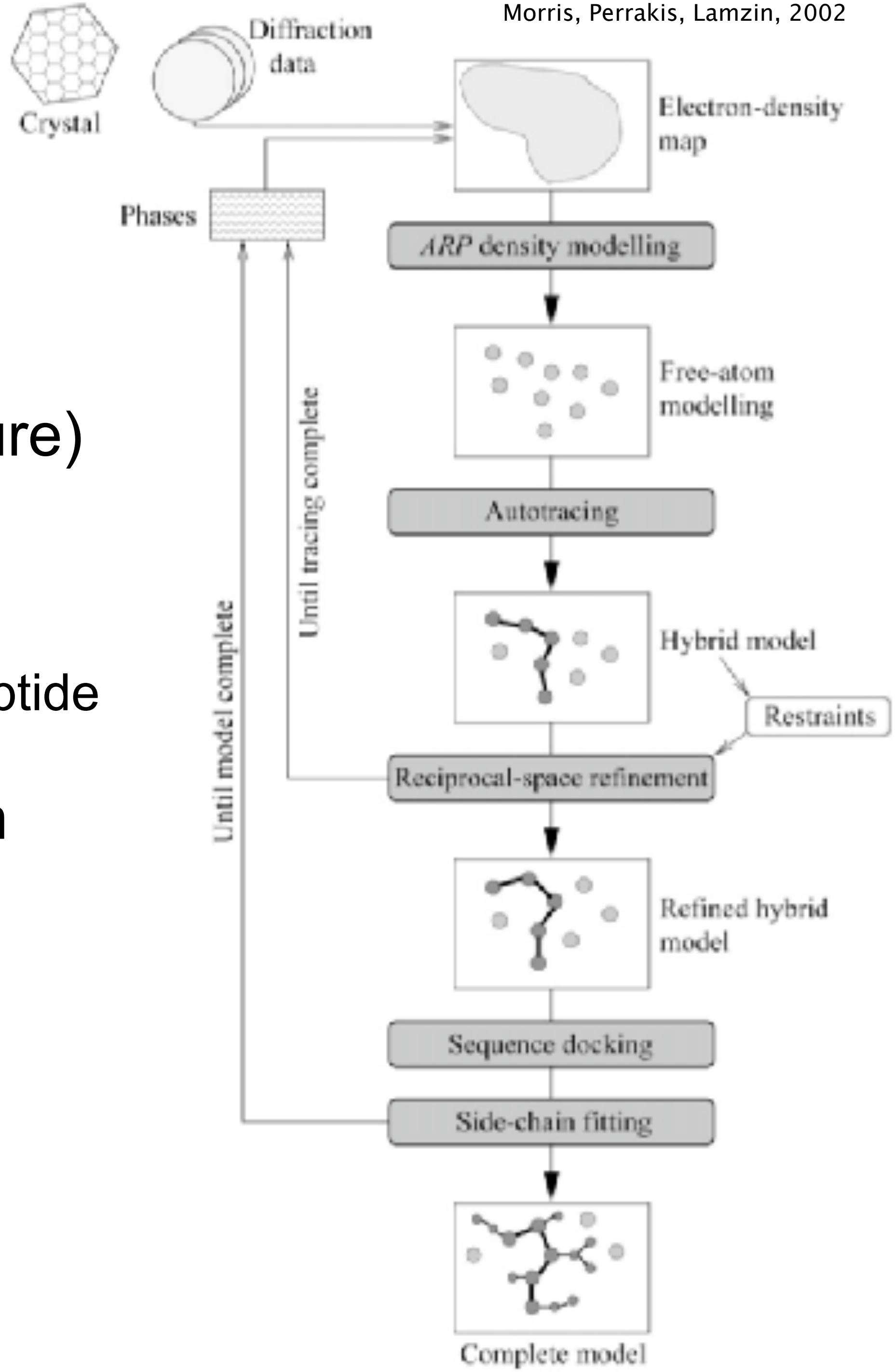
Assumptions:

- Crystal is a protein crystal
  Long single non-branching polypeptide chain
- Accessibility to high-resolution data (2.3A)

General Steps:

- Place Dummy Atoms
- Build Skeleton
- Refine Skeleton
- Add Sidechains

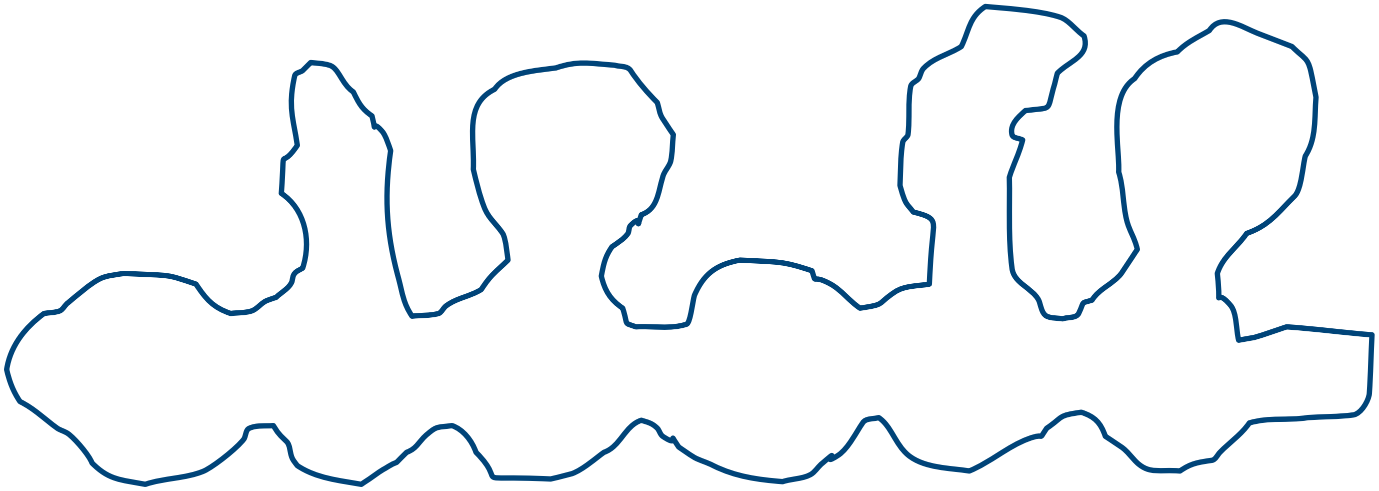
Morris, Perrakis, Lamzin, 2002



ARP/wARP

Flood Electron Density Map with Dummy Atoms

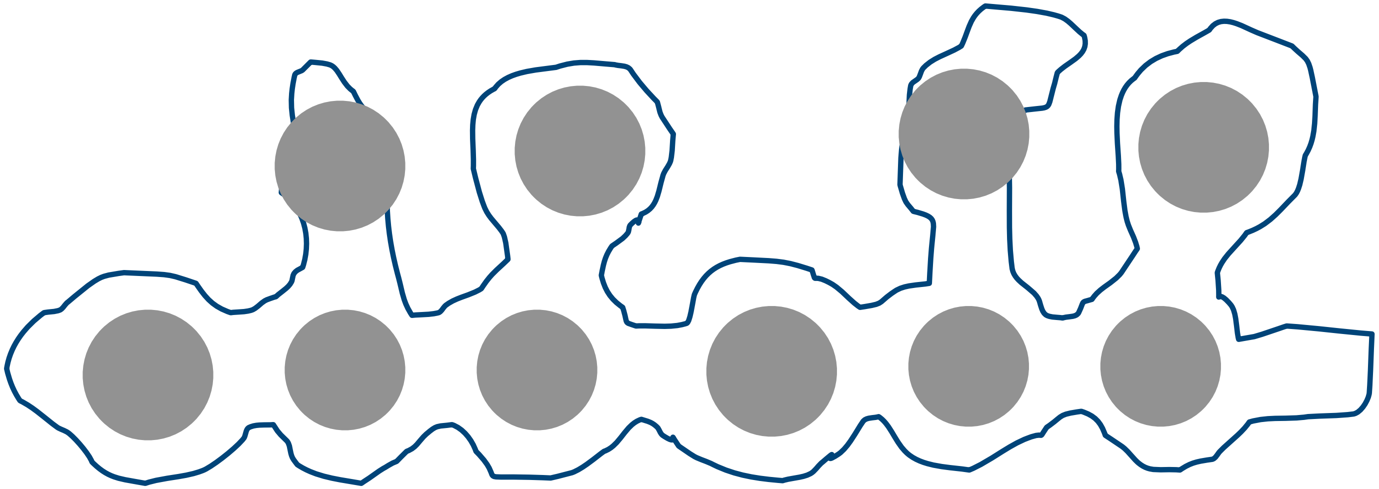
- Atoms placed in regions of high electron density
- Each placed atom is free to move (untethered)
  - Moves: translation, appear, disappear
  - Update phases



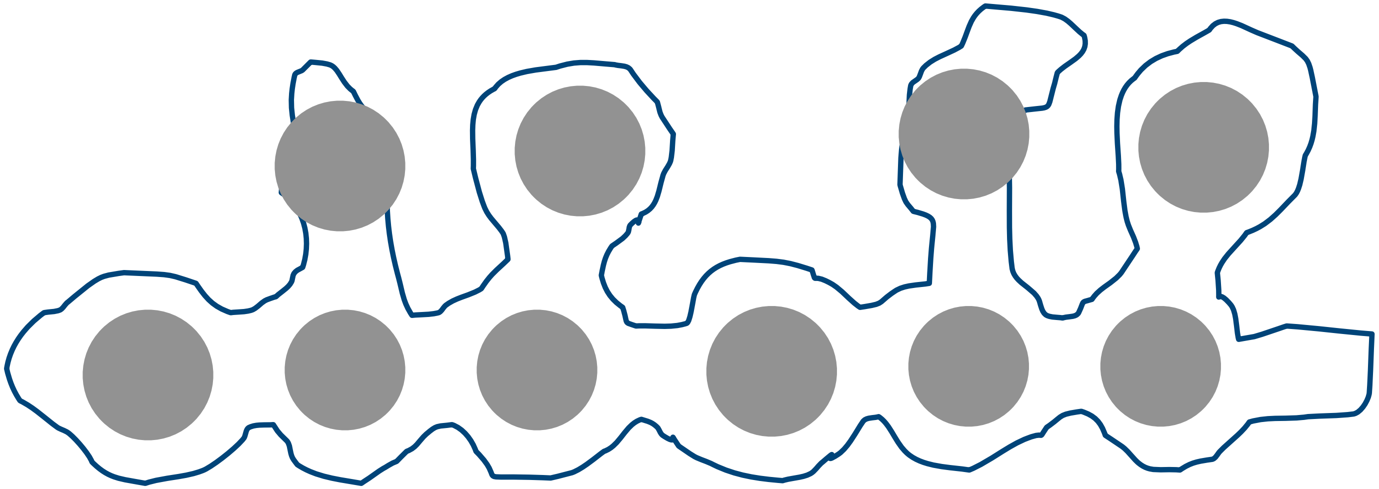
ARP/wARP

Flood Electron Density Map with Dummy Atoms

- Atoms placed in regions of high electron density
- Each placed atom is free to move (untethered)
  - Moves: translation, appear, disappear
  - Update phases



ARP/wARP



Atoms usually within 0.5A of final position

Tasks:

- Identify atom types
- Identify connectivity
- Align to sequence

First: Identify putative C_α positions

ARP/wARP

Each C_α should be connected to at least one other C_α approximately 3.8A away in either:

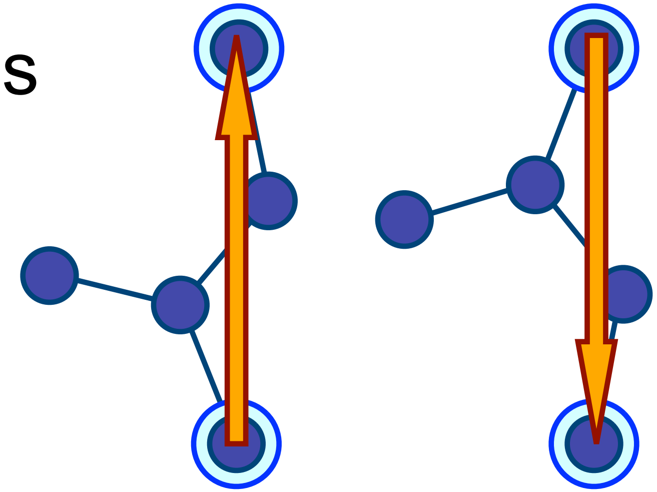
-C(=O)-N-C_a Forward (outgoing)

-N-C(=O)-C_a Backward (incoming)

For all pairs of atoms ~3.8A apart, check intervening electron density

If correlation of electron density is above threshold:

- Make vertex from candidate atoms
- Add edge between atoms



ARP/wARP

Given directed graph (previous slide) of candidate C_a

Generate graph where each vertex represents 4 continuous C_a

- Consider all paths of length 4 in original graph
- Prune 4-mers that are not consistent with protein structure

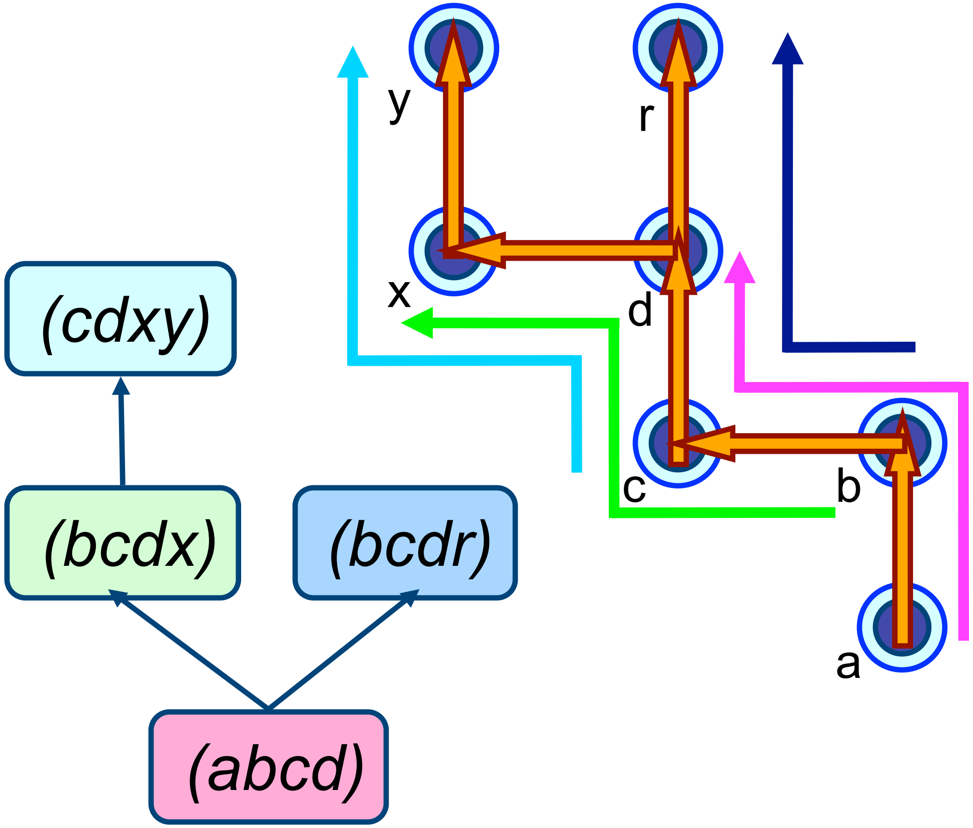
Valence Angle

$$C_\alpha(n) - C_\alpha(n+1) - C_\alpha(n+2)$$

Dihedral Angle

$$C_\alpha(n) - C_\alpha(n+1) \\ - C_\alpha(n+2) - C_\alpha(n+3)$$

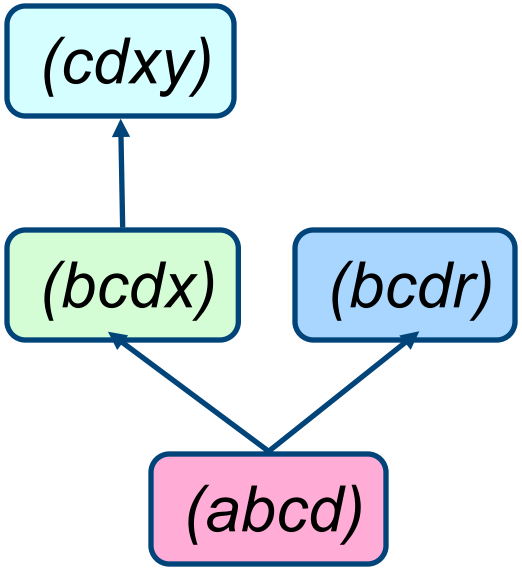
Underlying distribution mined from pdb, represented with Parzen windows of multivariate Gaussians.



ARP/wARP

Optimization problem:

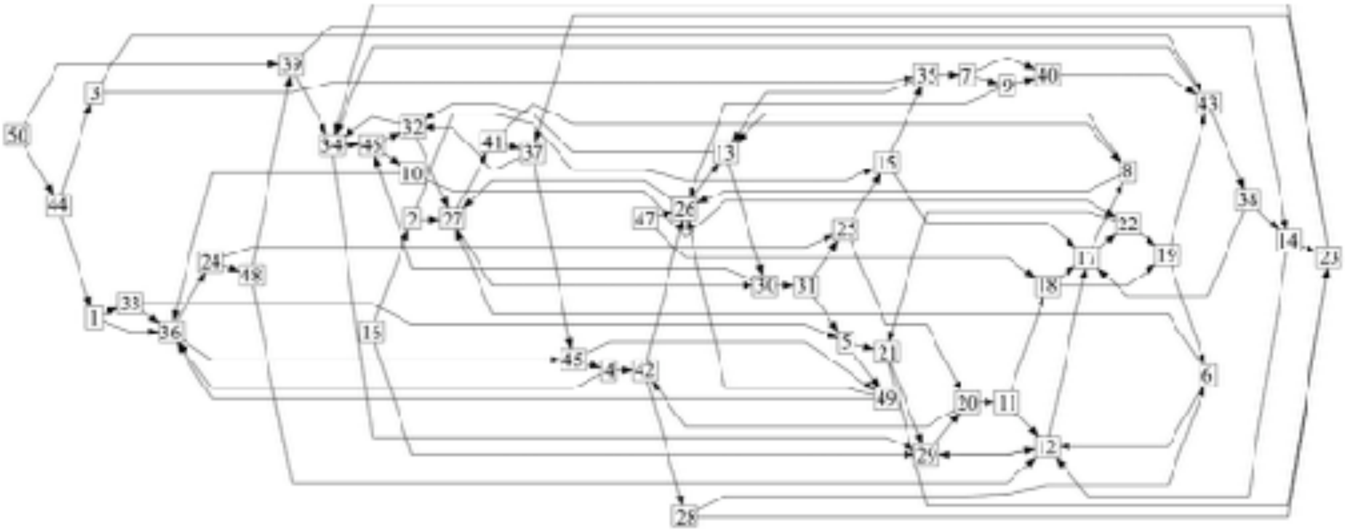
Finding set of chains in a weighted graph with highest score



Vertices - 4 C_a segments

Edges - overlapping fragments

Weights - geometrical scores of fragment and average electron density

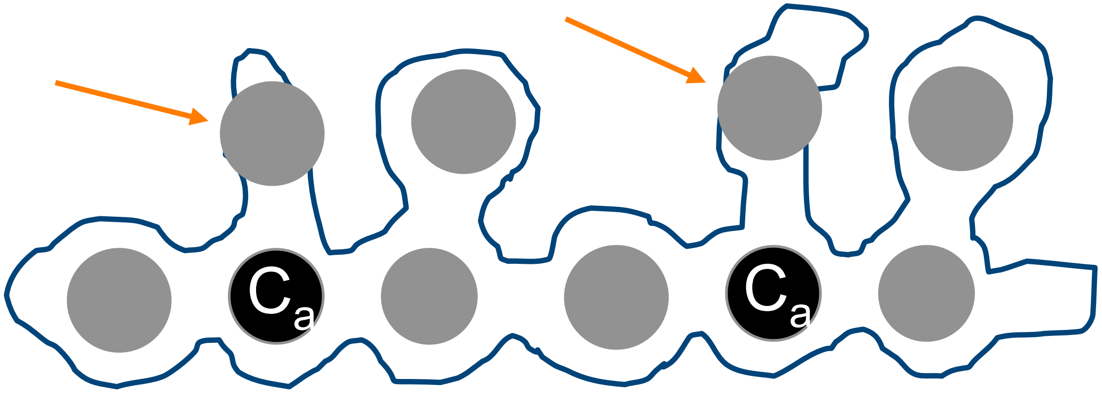


Morris, Perrakis, Lamzin, 2002

- Depth first search from each node to identify ‘best’ scoring chains
- Greedy merging
- Avg branching factor 2-4

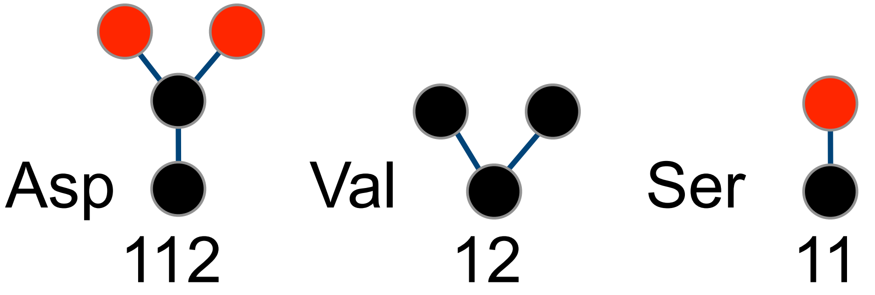
ARP/wARP

Sidechains



Consider atoms neighbouring C_as but not part of the backbone.

Compute a mini-feature vector for each C_a, based on number of atoms hanging off the C_a



$$p(\mathrm{AA}|D_i)$$

Compute probability of each AA type for each C_a density region D_i

Compute score of sliding window over observed densities *D* and known sequence *S*

$$P(D_i, j) = \prod_{k=-m}^{m} p(S_{j+k}|D_{i+k})$$

TEXTAL

Locate putative C_a positions

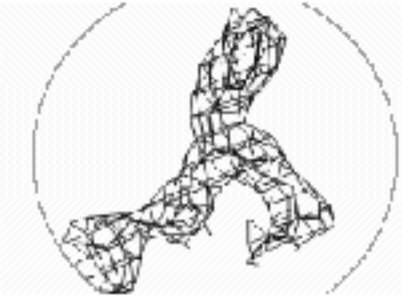
Use of rotation invariant feature vectors

- Average Density / Distance to center of mass
- Moment of Inertia Based, Skewness (magnitudes and ratios)
- Tubes (C_a should have 3 regions of density extending out)

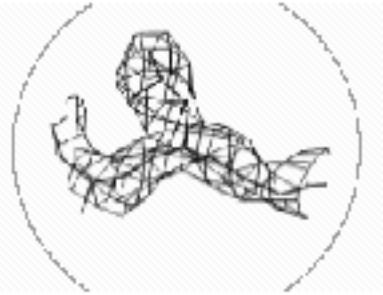
19 Features per Radius (4 radii used)

Compare feature vectors to classify each C_a into
Structure and AA type

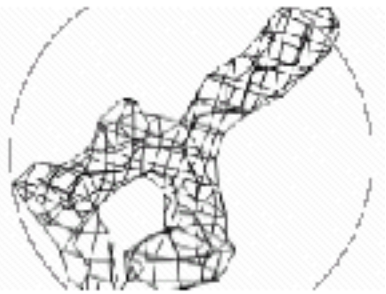
Match against fragments from the PDB database



Phe

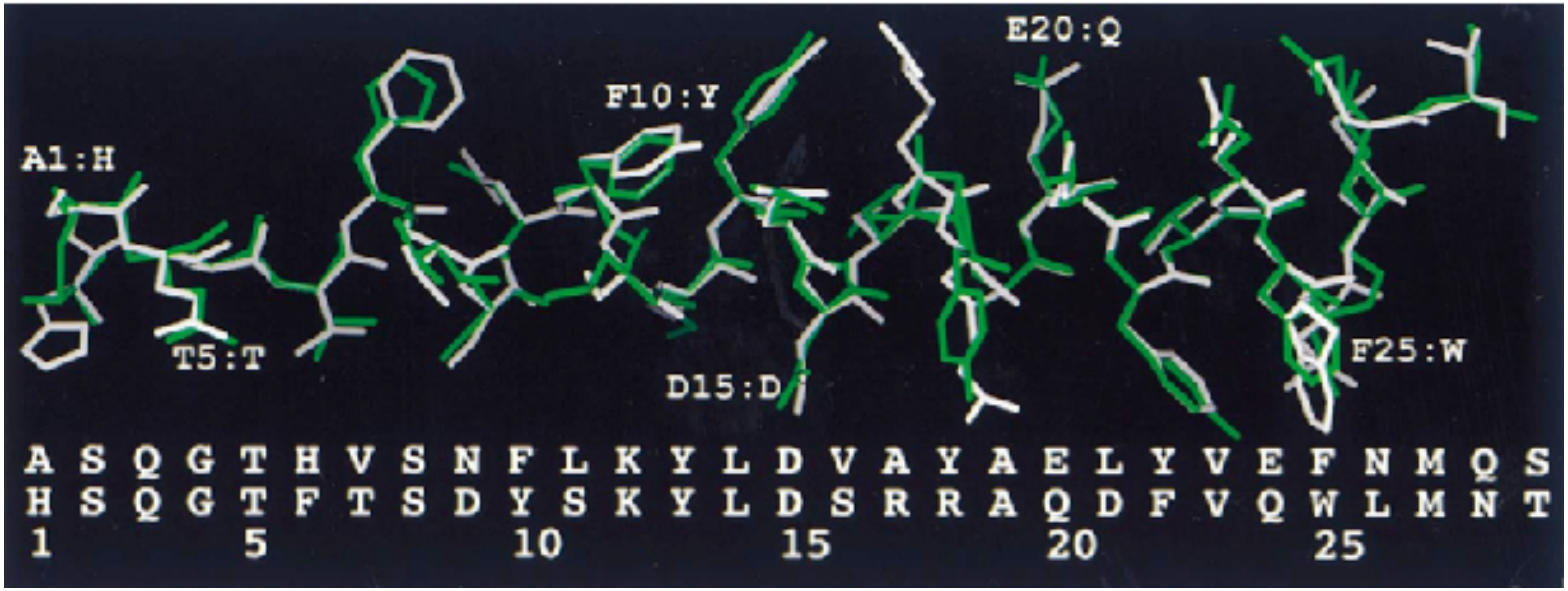


Leu



Lys

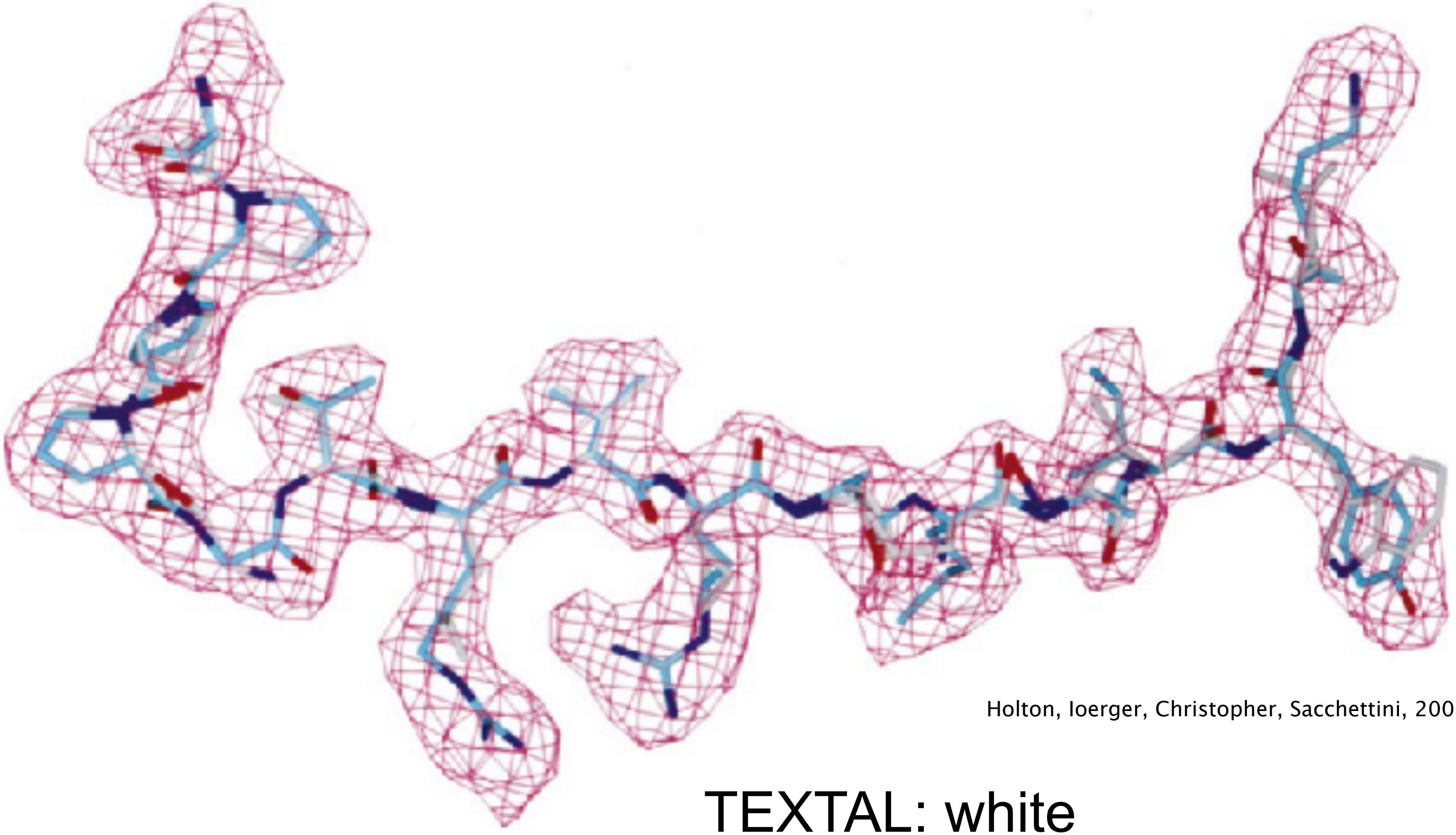
TEXTAL


A1:H
T5:T
F10:Y
D15:D
E20:Q
F25:W
A S Q G T H V S N F L K Y L D V A Y A E L Y V E F N M Q S
H S Q G T F T S D Y S K Y L D S R R A Q D F V Q W L M N T
1 5 10 15 20 25

Holton, Ioerger, Christopher, Sacchettini, 2000

TEXTAL: green structure, top sequence
Correct / Refined: white structure, bottom sequence

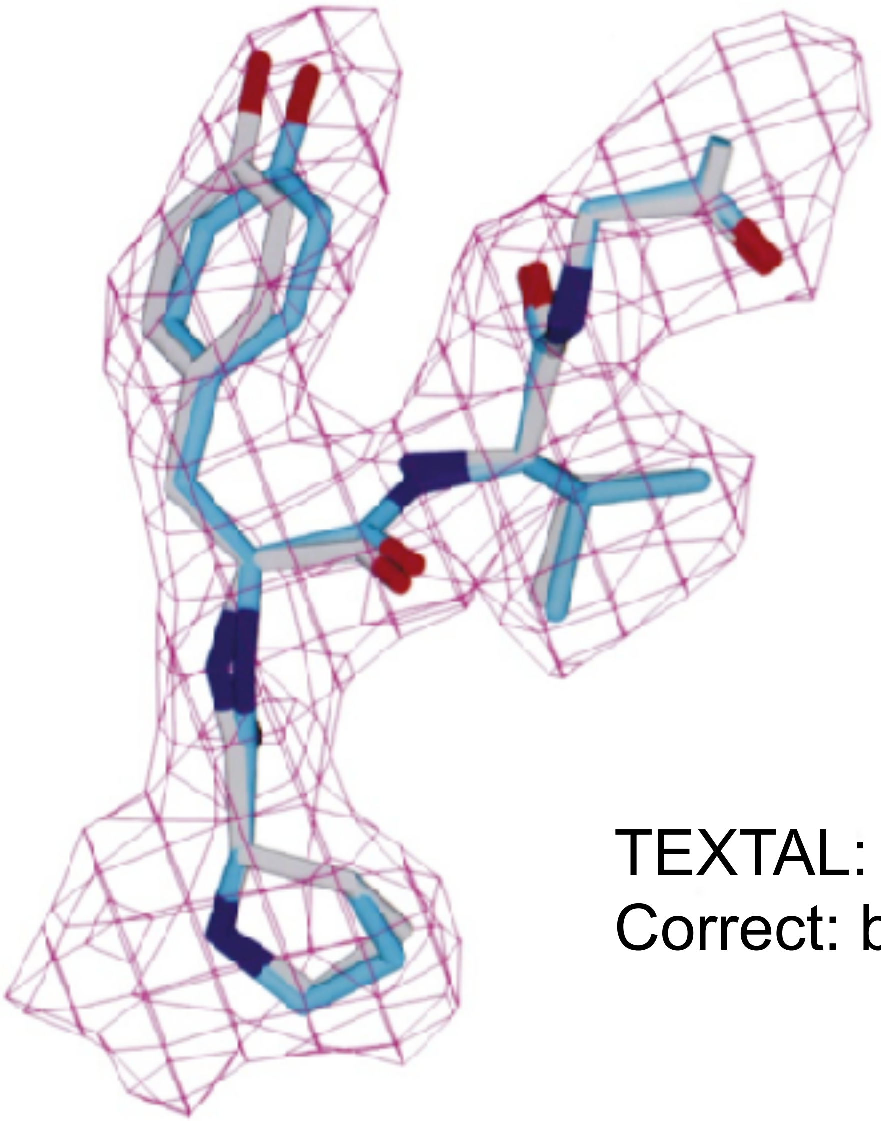
TEXTAL



Holton, Ioerger, Christopher, Sacchettini, 2000

TEXTAL: white
Correct: blue

TEXTAL



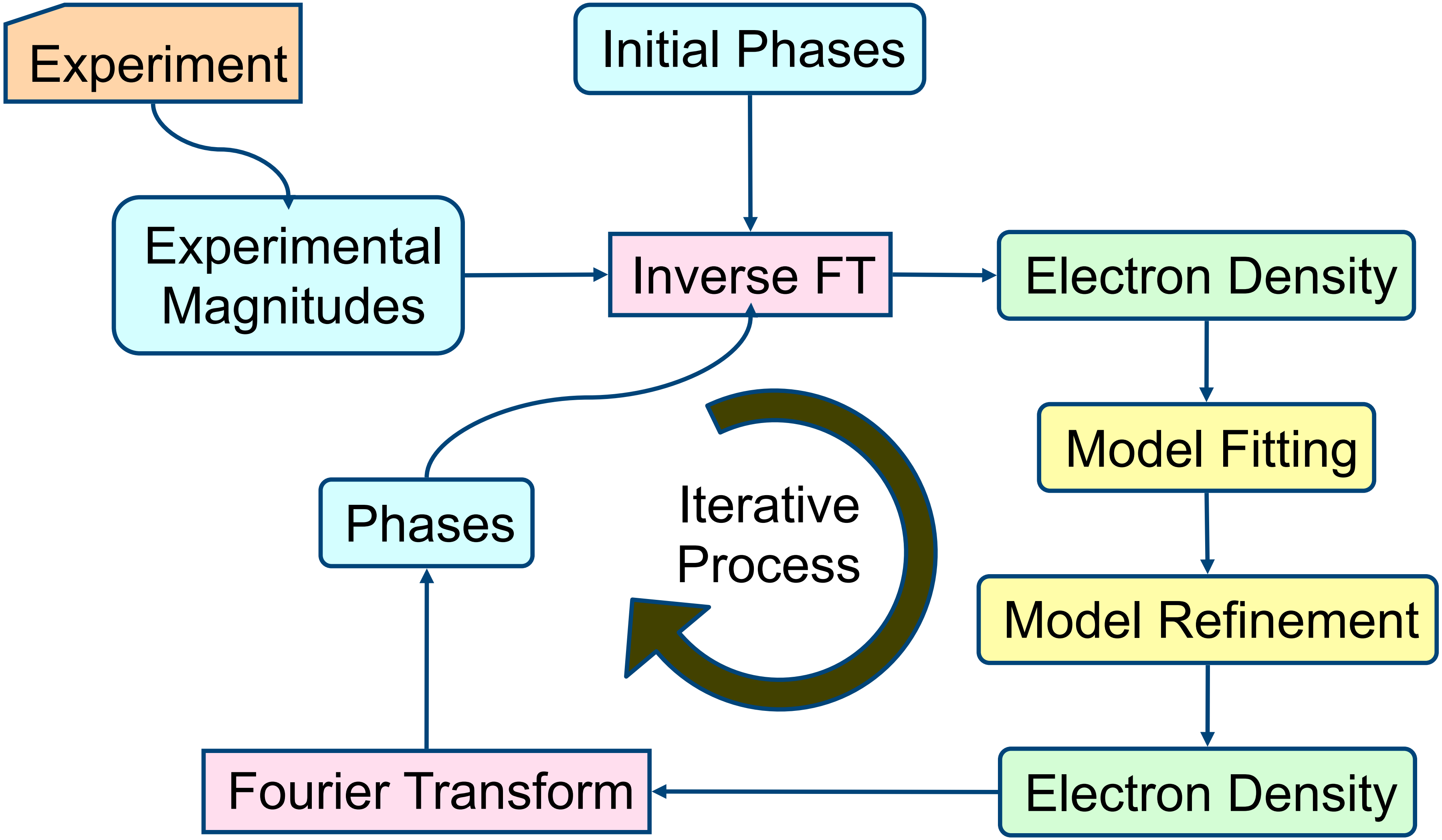
Results Building 12 Proteins

Mean Ca RMSD = 0.96A
All atom RMSD = 1.04A

TEXTAL: white
Correct: blue

Holton, Ioerger, Christopher, Sacchettini, 2000

Iterative Structure Solution - XRC



Experiment

Initial Phases

Experimental Magnitudes

Inverse FT

Electron Density

Model Fitting

Phases

Iterative Process

Model Refinement

Fourier Transform

Electron Density

Nuclear Magnetic Resonance Spect.

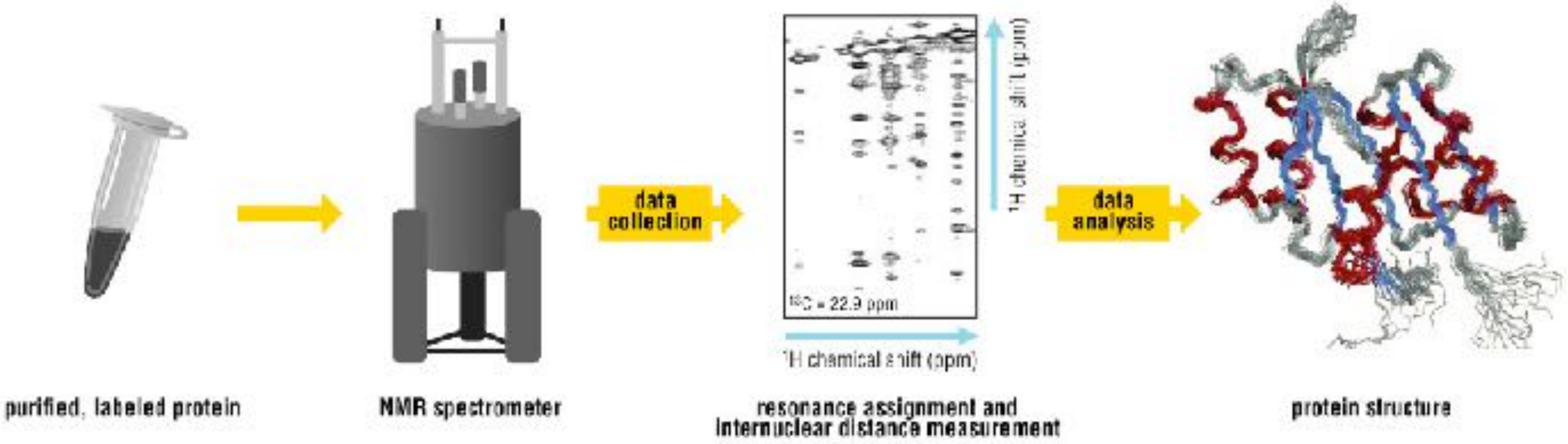
Proteins in Solution - high concentration, but don’t want crystal

Two broad classes of experiments:

- Get dictionary of resonances
- Measure geometric constraints (bond, angle, space)

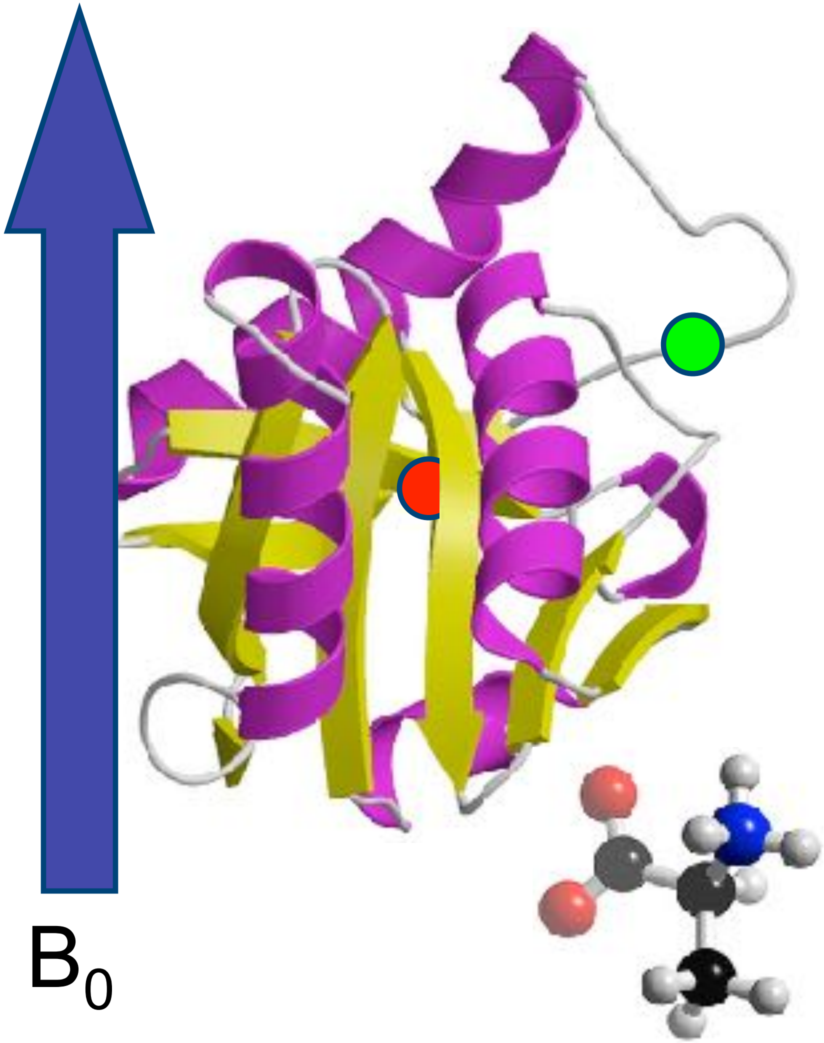
Generate ensemble of conformations consistent with constraints

Can measure protein dynamics



Effect of Local Environment

IMPORTANT



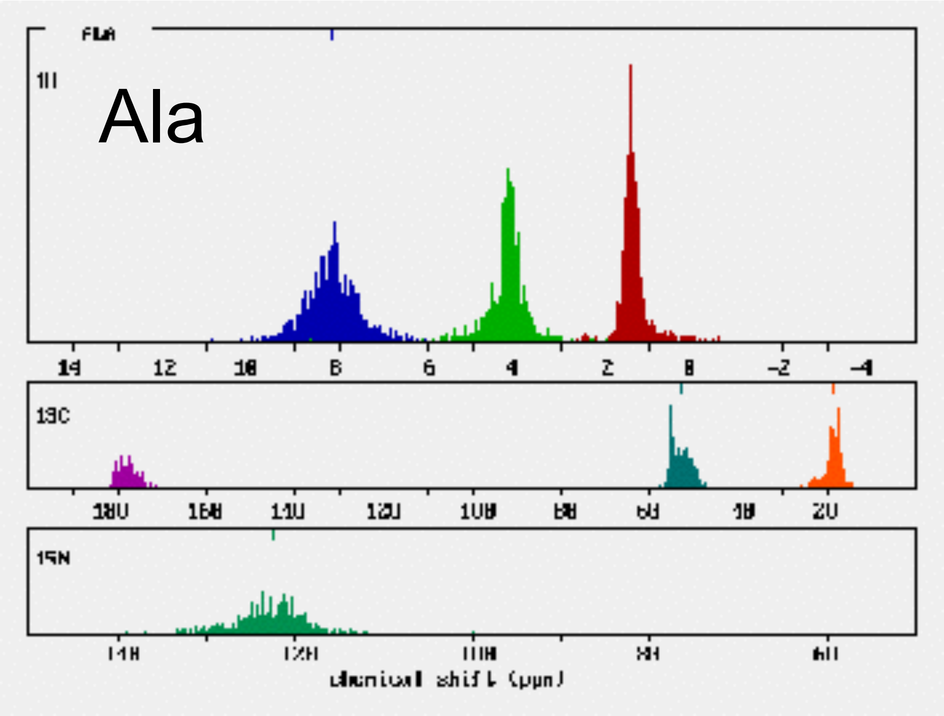
Different Atoms

Different Electronic Environments

Atoms experience B_0 differently

Resonate at different frequencies

slightly different frequencies

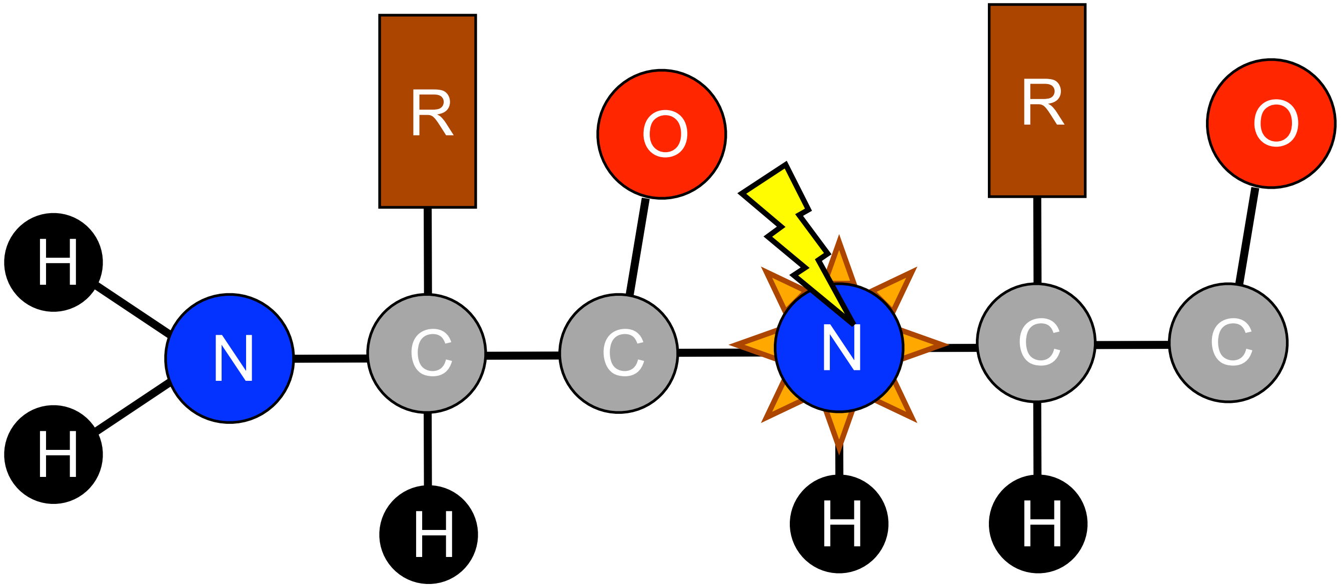


NMR

Resonance Transfer

Provides Information on:

Connectivity, Torsion Angles, Proximity

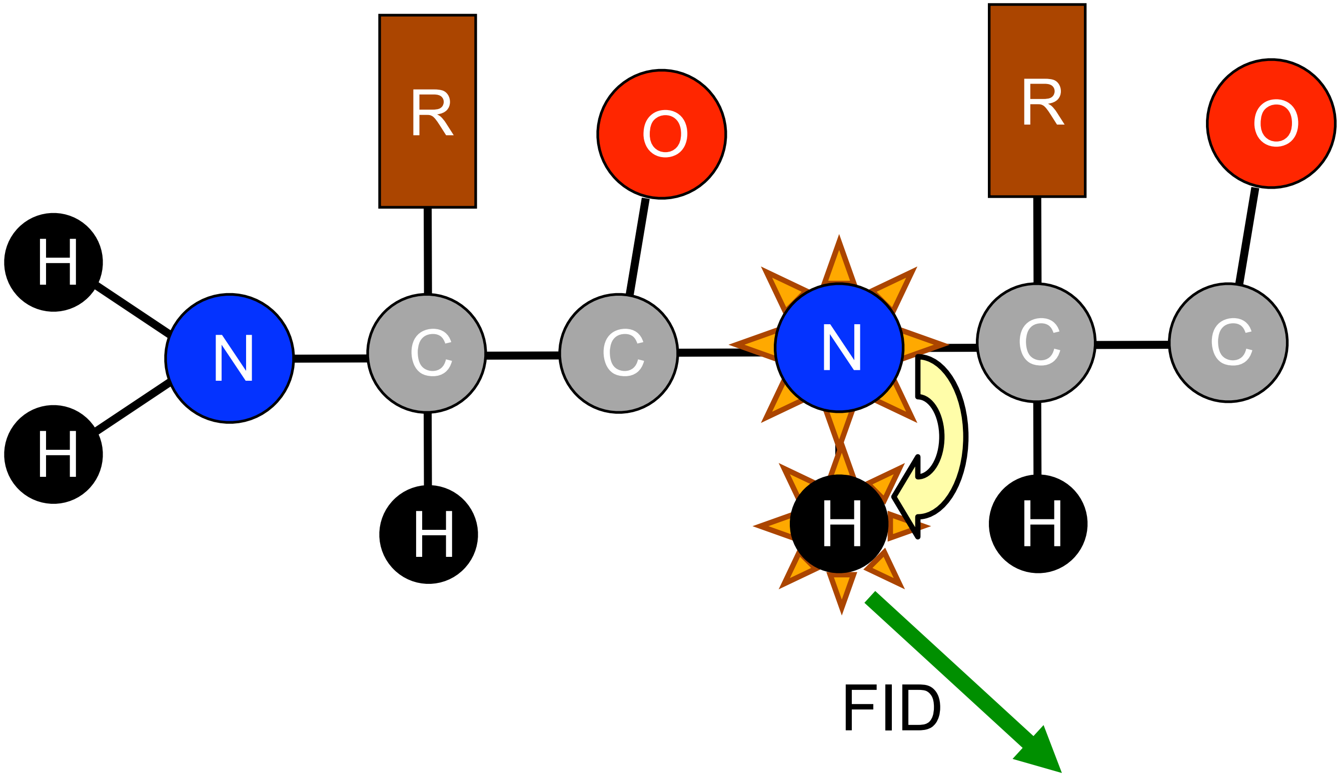


NMR

Resonance Transfer

Provides Information on:

Connectivity, Torsion Angles, Proximity

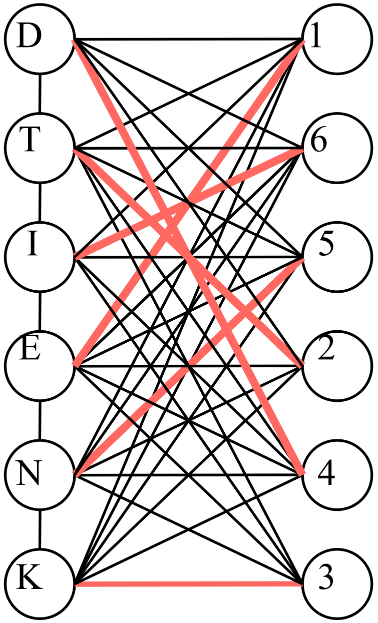


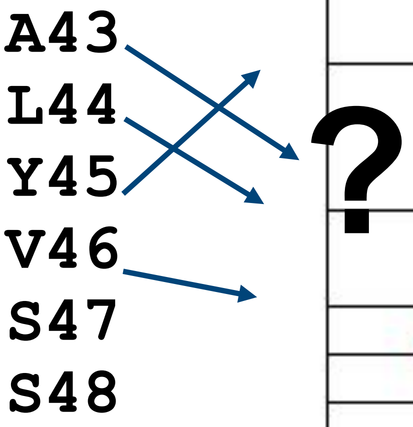
NMR

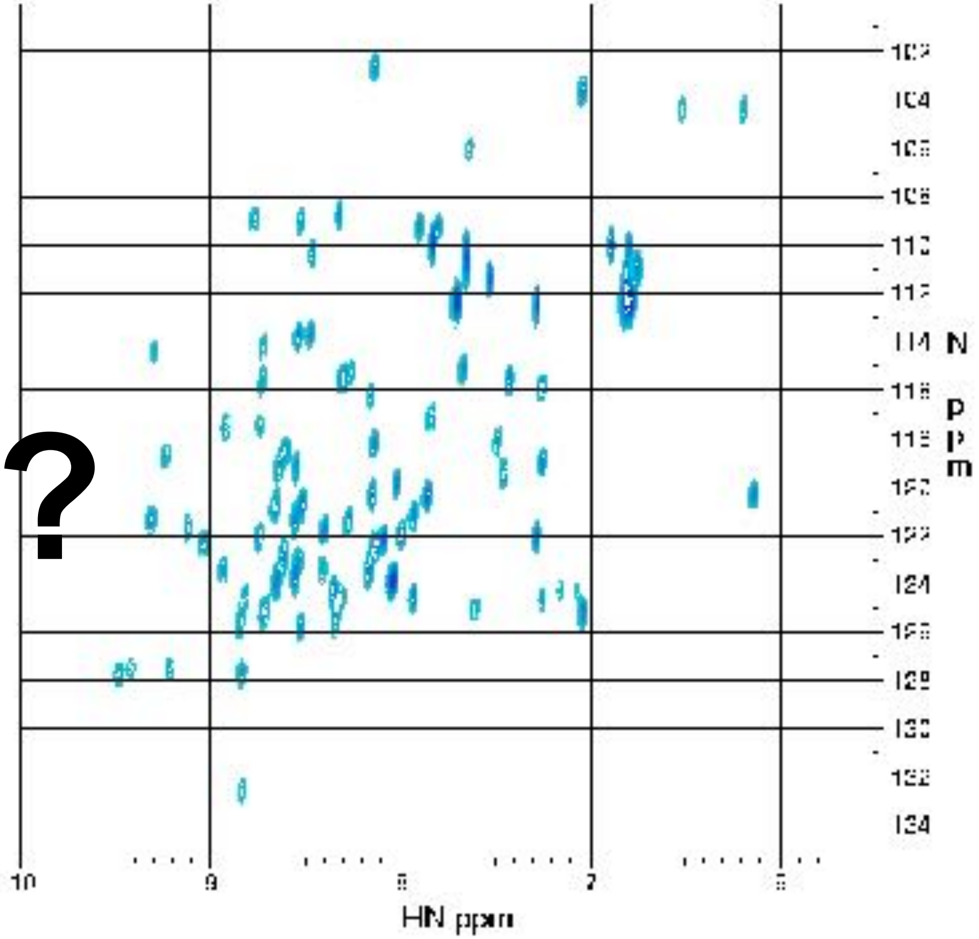
Assignment Problem!

Spectra are Unassigned!

Unknown correspondence between spectral peak and residue

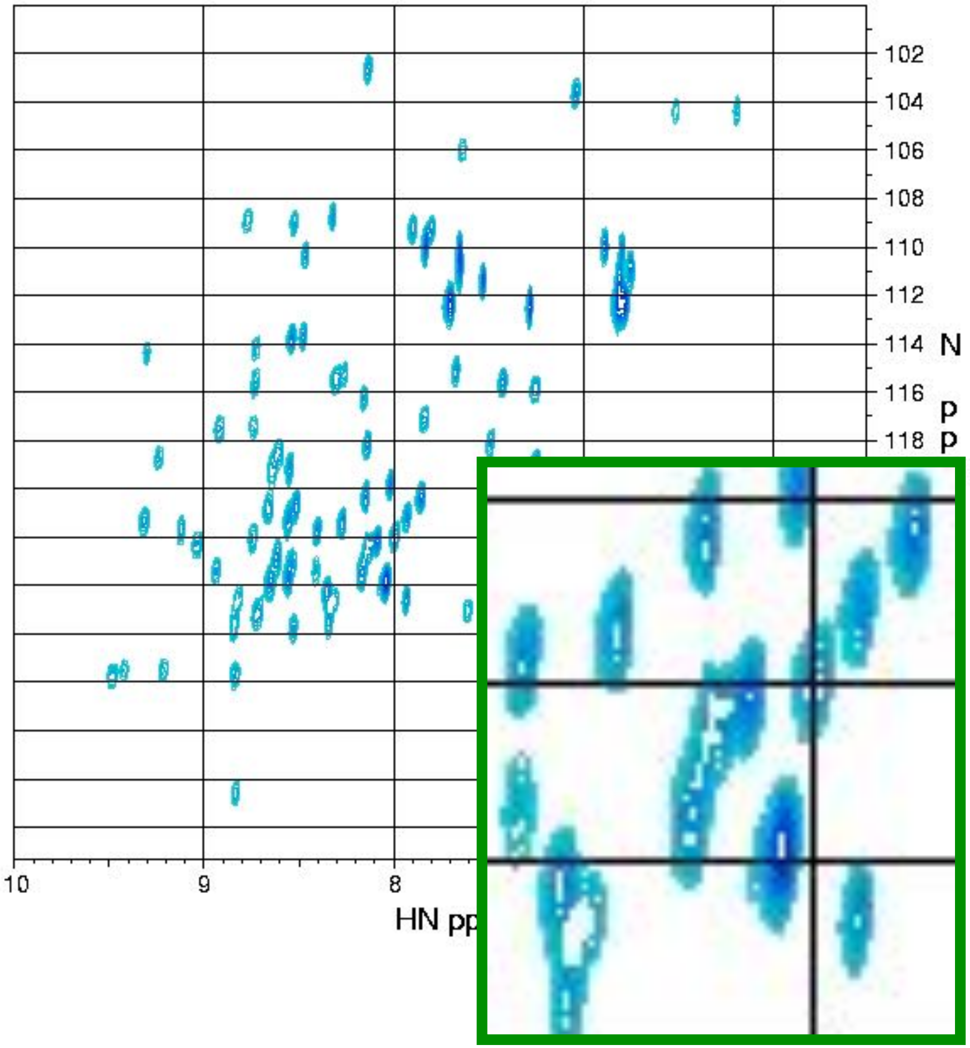


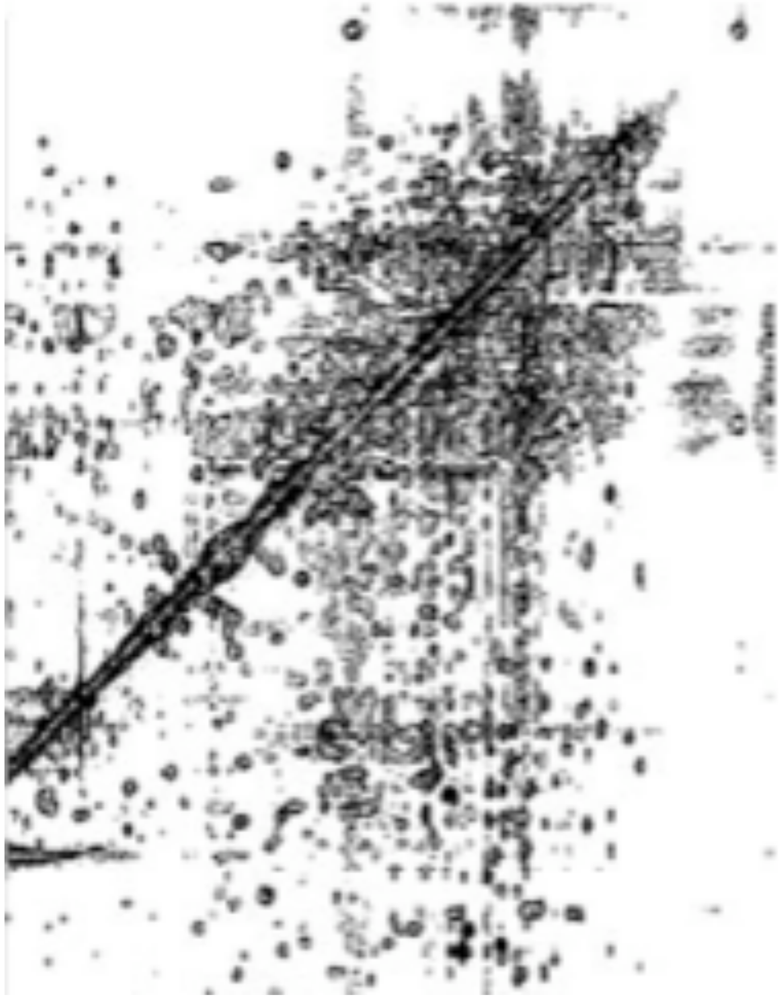




NMR

Peak Picking Problem!



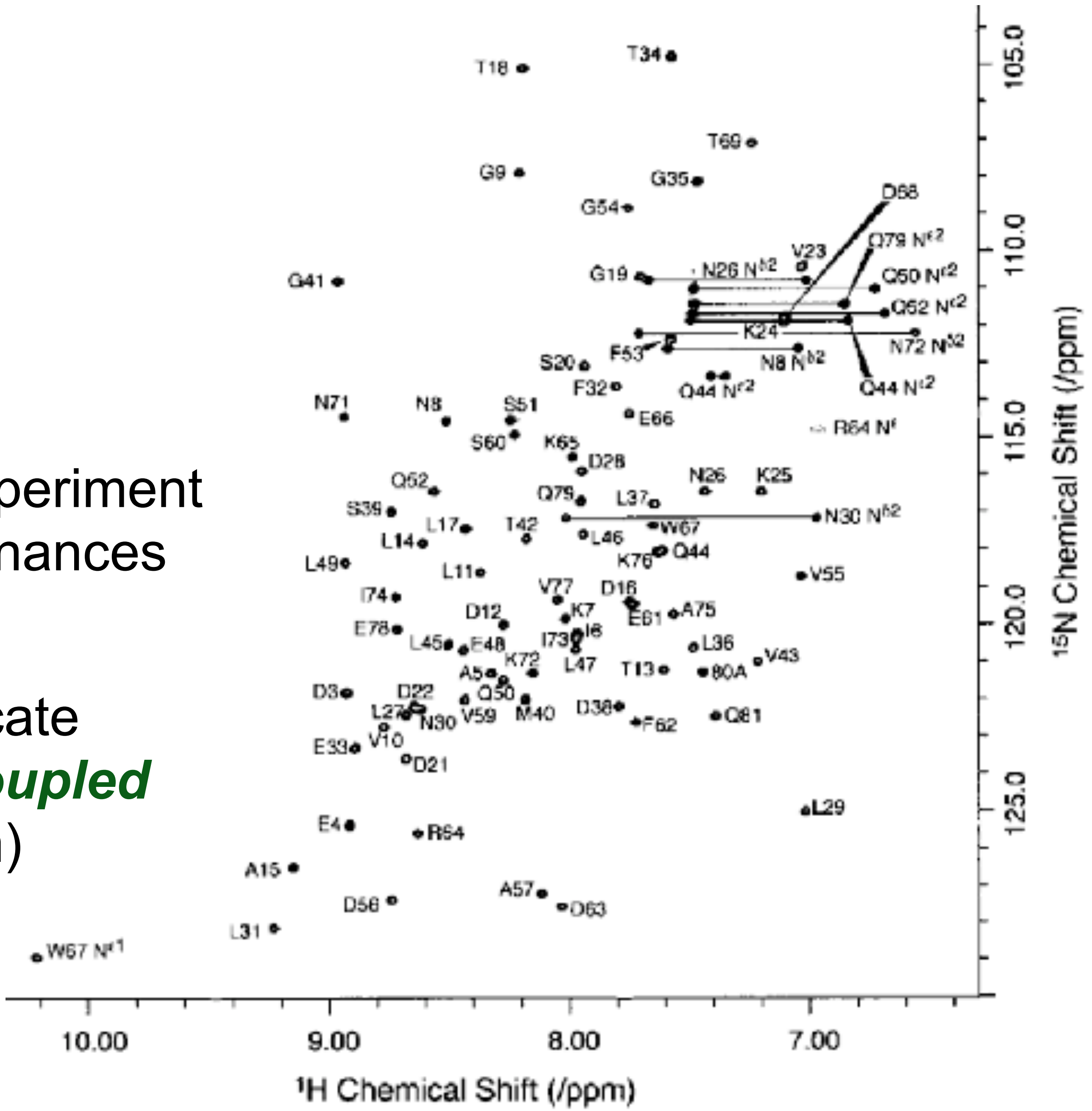


HSQC

Heteronuclear Single-Quantum Correlation

Through Bond Experiment
Identifies NH Resonances

Cross-Peaks indicate that atoms are ***coupled*** (aka Spin System)



NMR

Three Main Stages

Resonance Assignments (assume peaks picked)

Geometric Constraints

- **Dihedrals**: J-couplings - interaction of dipoles
- **Interatomic Distances**: NOEs
- **Relative Bond Vector Orientation**: RDCs

Structure Generation

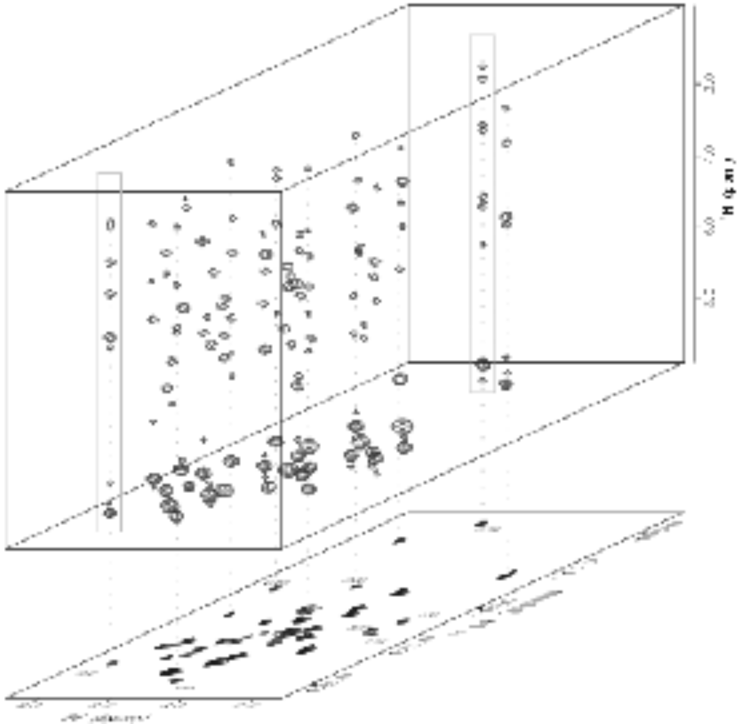
2D vs 3D vs ...

Multidimensional NMR

- Vary transfer times

Spreads peaks out

- Allows better peak picking



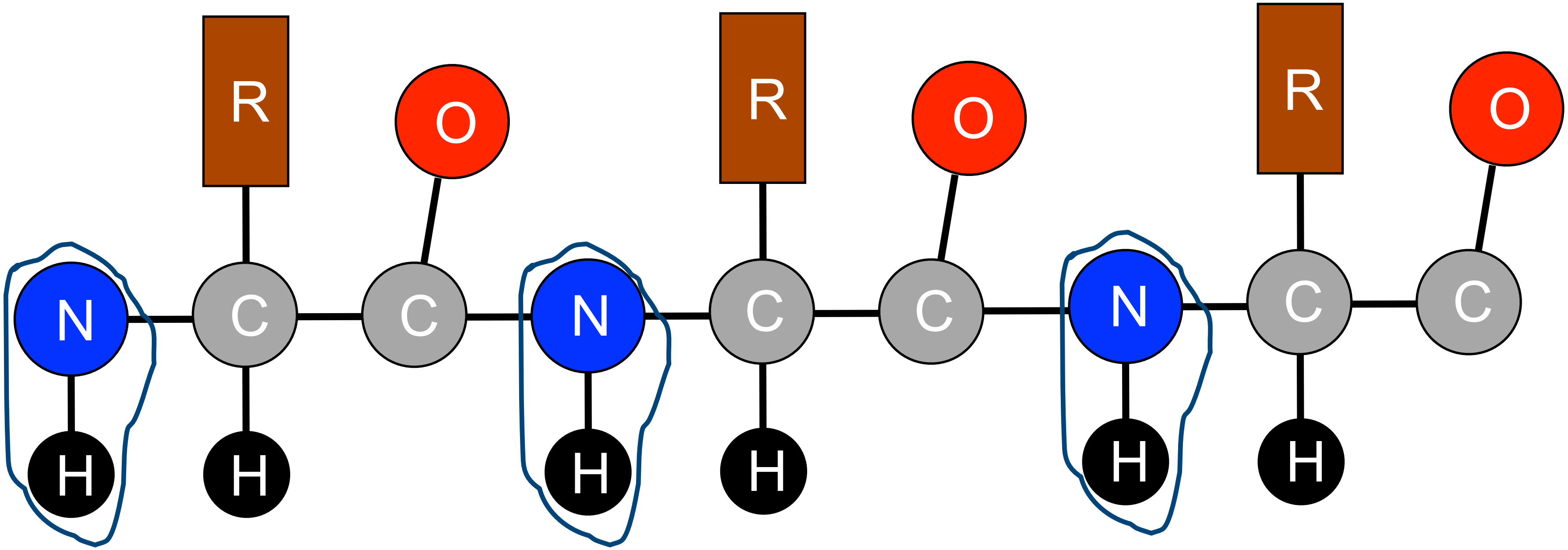
NMR - Experiment Types

HSQC - (HN(i), N(i))

HNCA - (HN(i), N(i), C_a(i)) & (HN(i), N(i), C_a(i-1))

HNCOCA - (HN(i), N(i), C_a(i-1))

HNCO, HNCACO, CCONH, CBCACONH, HNCACB



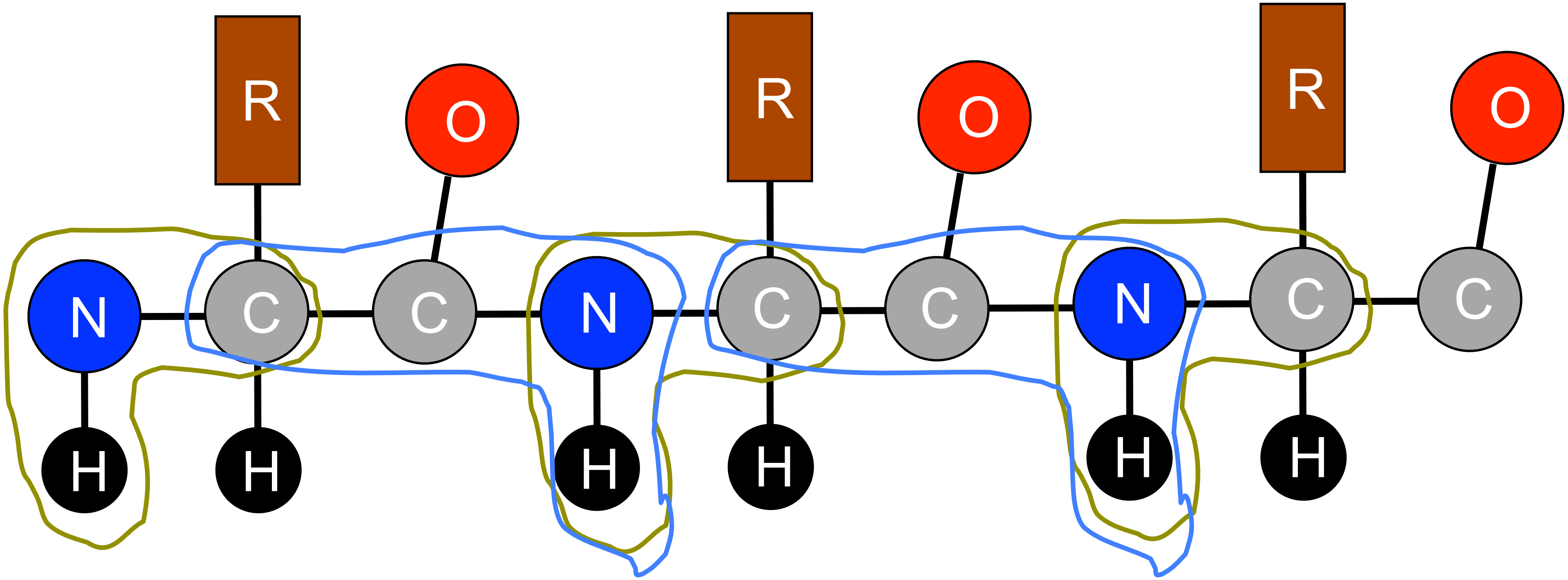
NMR - Experiment Types

HSQC - (HN(i), N(i))

HNCA - (HN(i), N(i), C_a(i)) & (HN(i), N(i), C_a(i-1))

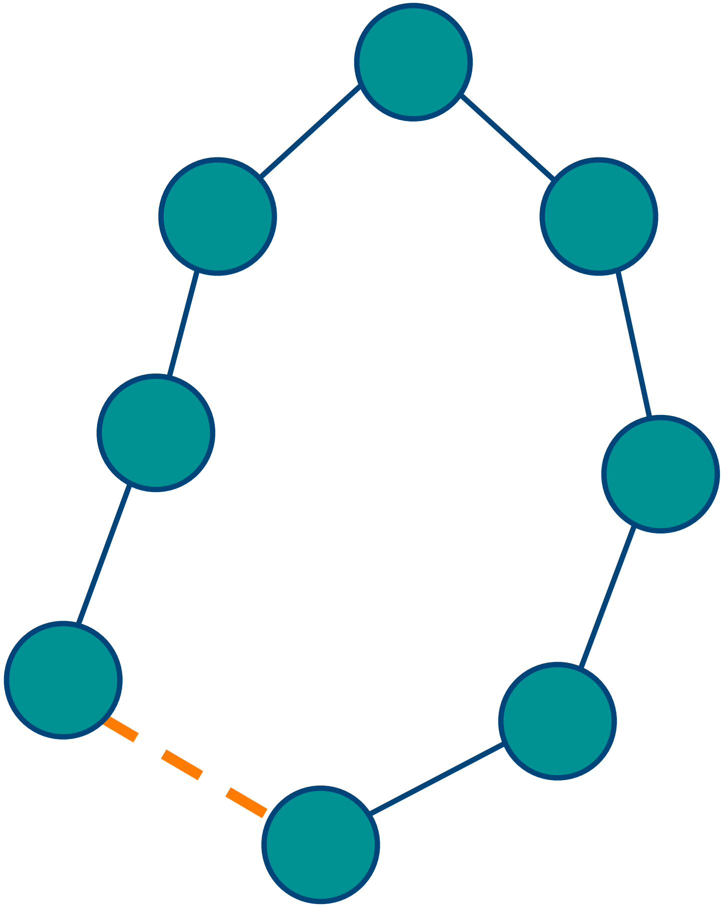
HNCOCA - (HN(i), N(i), C_a(i-1))

HNCO, HNCACO, CCONH, CBCACONH, HNCACB



NMR

Through Space Resonance Transfer

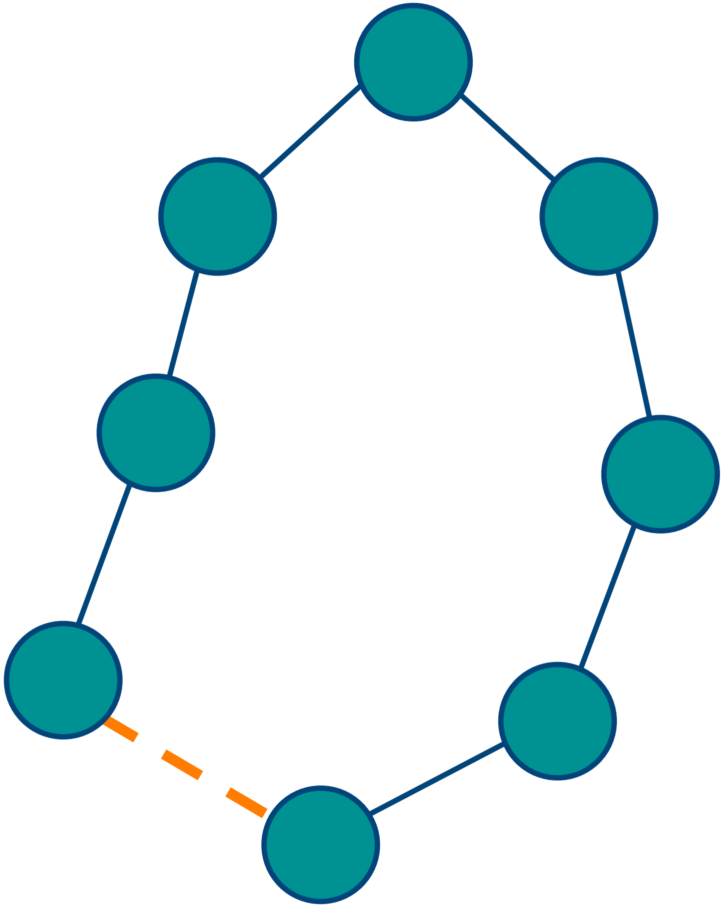


NOESY

Nuclear Overhauser Effect (NOE)
Through Space
Resonance transferred between
two non-bonded hydrogens.
Strength falls off as r^6
Atoms must be <6A apart

NMR

NOESY



Nuclear Overhauser Effect (NOE)
Through Space
Resonance transferred between two non-bonded hydrogens.
Strength falls off as r^6
Atoms must be <6A apart

Crude Distance Measurements

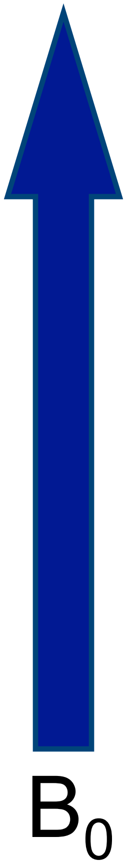
Large Peaks	0 - 2.5 A
Medium Peaks	0 - 3.5 A
Smaller Peaks	0 - 5.0 A

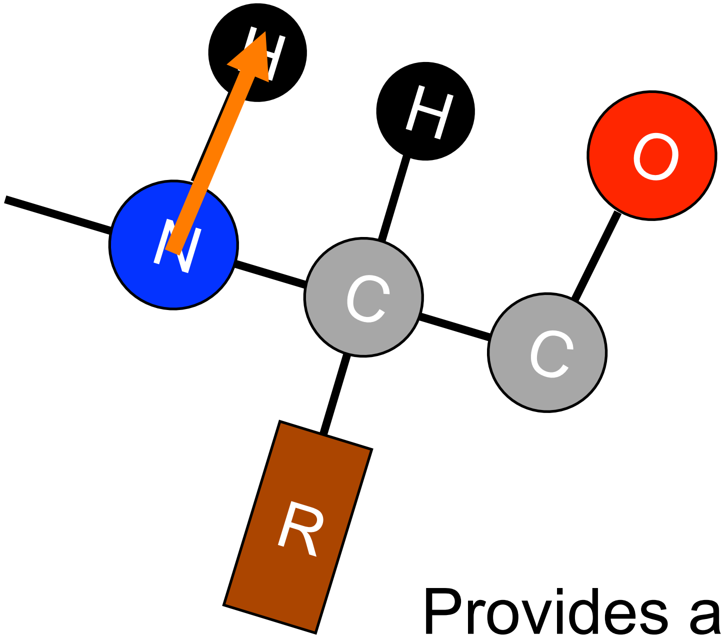
NMR

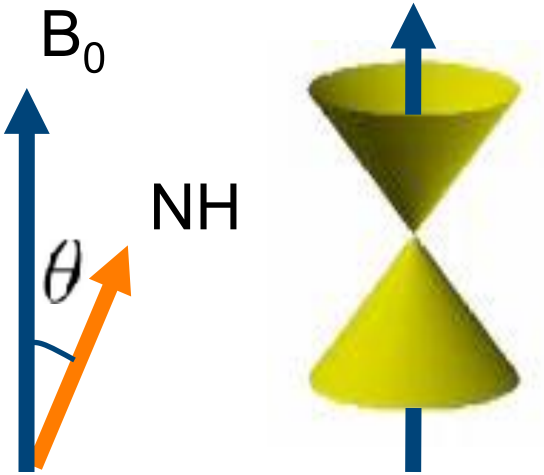
Residual Dipolar Couplings

Measures angle of bond vector wrt B_0

Use of partially aligning media







Provides additional geometric constraint

NMR

Available Information

Sequential Connectivity

HSQC, HNCA

Residue Type 'Assignment'

TOCSY

Through Space Distance Constraints

NOEs

Bond Vector Orientations

RDCs

Geometric Constraints

Dihedrals: J-couplings - interaction of dipoles

Interatomic Distances: NOEs

Relative Bond Vector Orientation: RDCs

NMR - Structure Generation

Challenges:

Missing information
False information

Typical Approach: MC or Molec. Dynamics

$$V_{\text{total}} = V_{\text{bonded}} + V_{\text{nonbonded}} + V_{\text{NMR}}$$

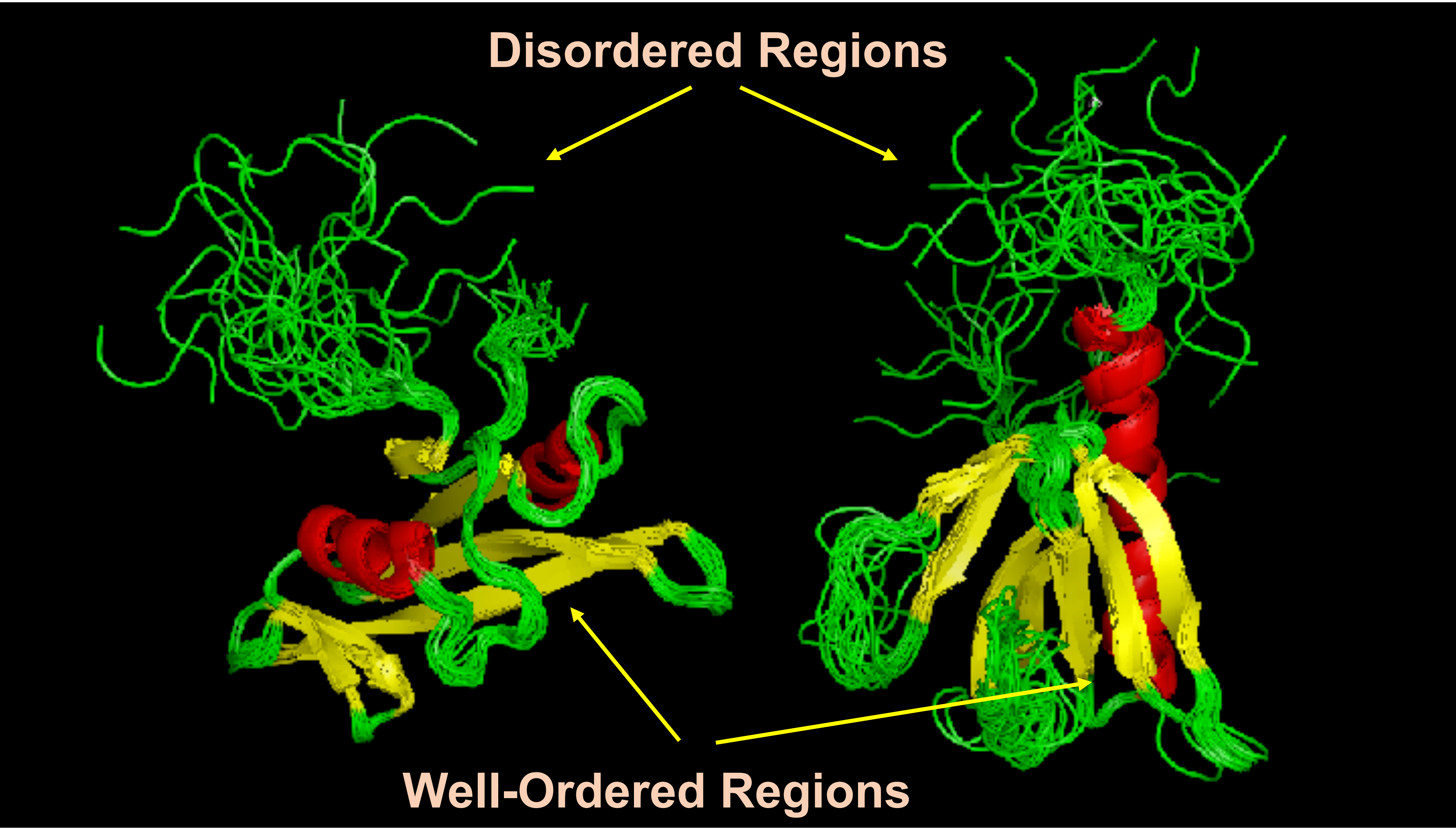
DYANA

Start with ‘random’ conf.
Energy function of PE, KE
Torsion Angle Optimization
MD with Simulated Annealing

$$V = \overbrace{\sum_{u,l,v} \sum_{(\alpha,\beta)\in I_c} f_c(d_{\alpha\beta}, b_{\alpha\beta})}^{\text{distance constraints}} + \underbrace{w_d \sum_{k\in I_d} \left(1 - \frac{1}{2}\left(\frac{\Delta_k}{\Gamma_k}\right)^2\right) \Delta_k^2}_{\text{torsion angle constraints}}$$

NOE

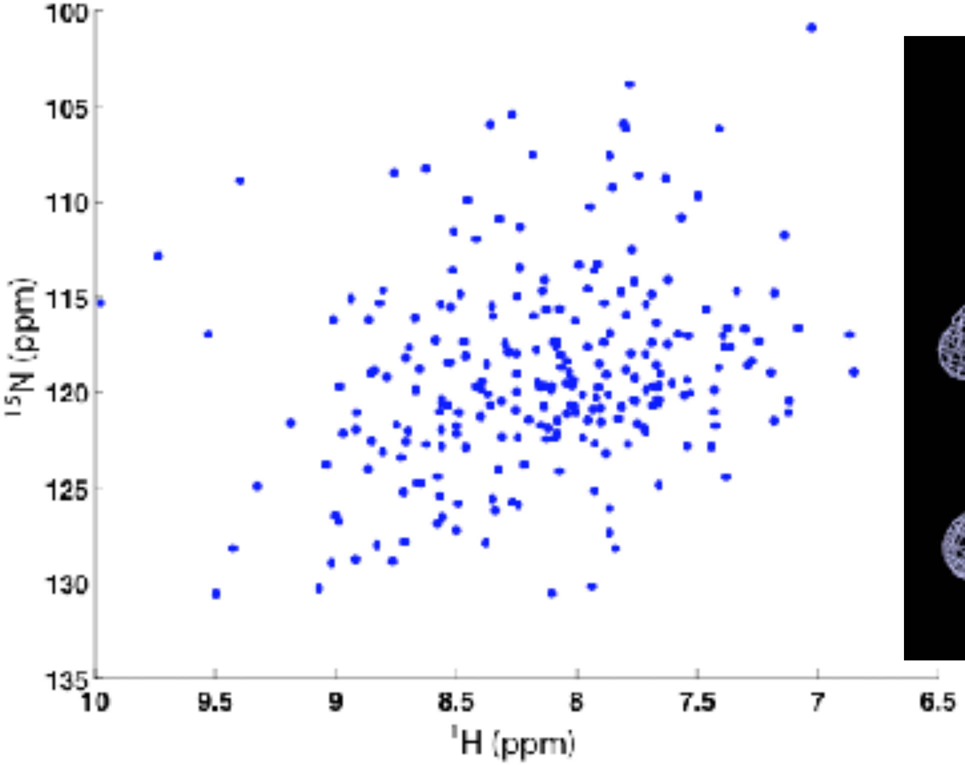
NMR

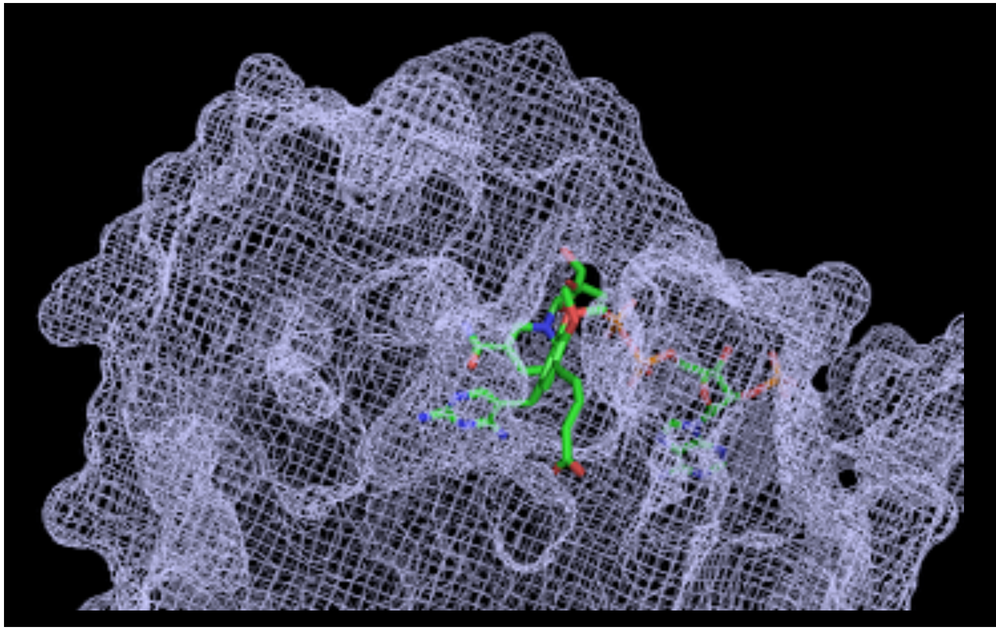


SAR-by-NMR

Structure-Activity-Relationship or Chemical Shift Perturbation

Assists in Ligand Binding and Protein-Protein Interactions

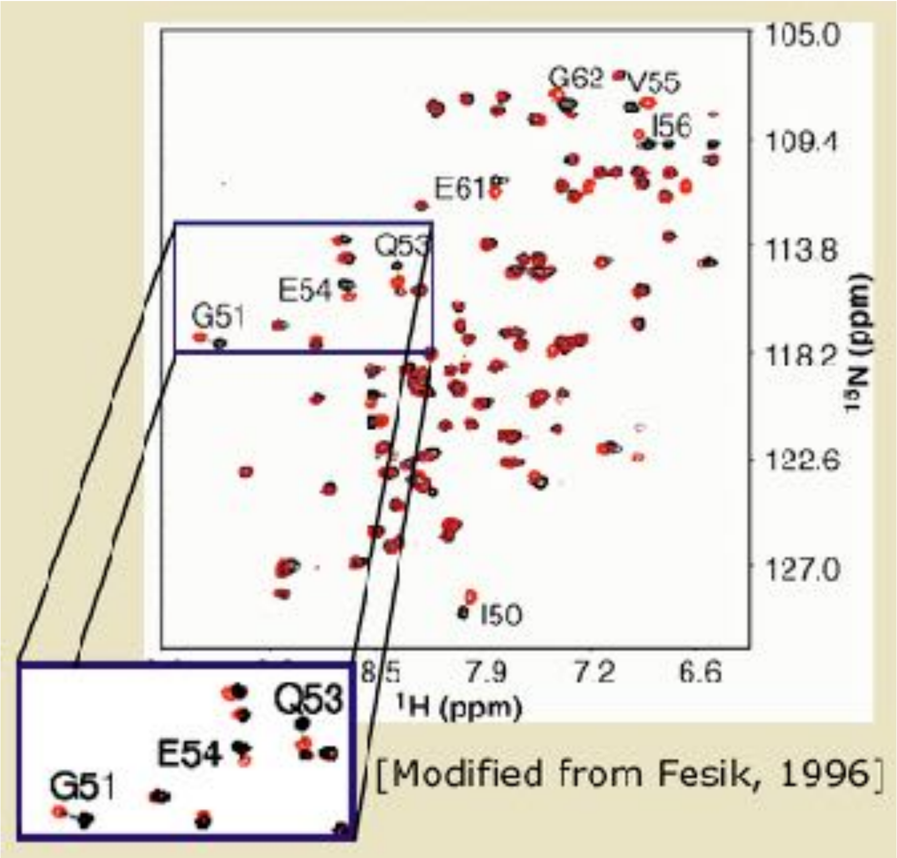




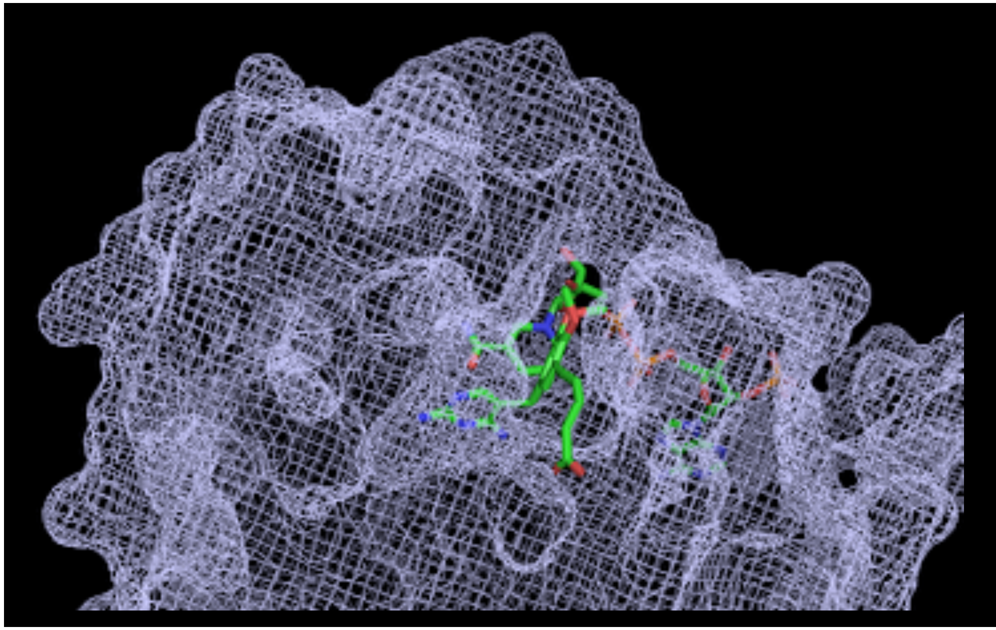
SAR-by-NMR

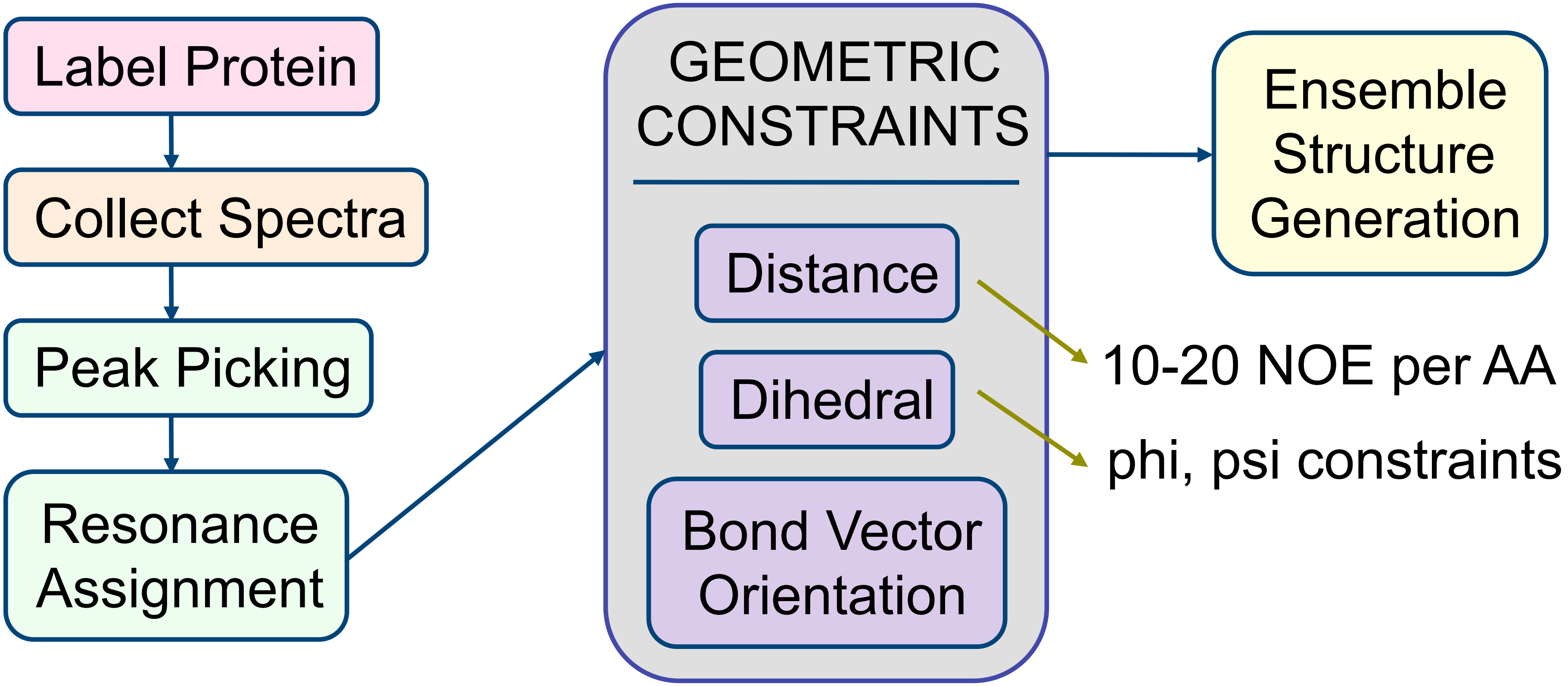
Structure-Activity-Relationship or
Chemical Shift Perturbation

Assists in Ligand Binding and Protein-Protein Interactions



[Modified from Fesik, 1996]





Label Protein

Collect Spectra

Peak Picking

Resonance Assignment

GEOMETRIC CONSTRAINTS

Distance

Dihedral

Bond Vector Orientation

10-20 NOE per AA

phi, psi constraints

Ensemble Structure Generation

Assignment Problem

- Noise, Degeneracy
- Often cast as graph algorithm
- Locate Mutually Consistent Information

Structure Generation

- Identify structures consistent with most geometric constraints
- Must ignore some constraints
- Utilize prior knowledge

Interpreting Dynamics Information

- Model time evolution of spin-systems

Experimental Struct. Determination

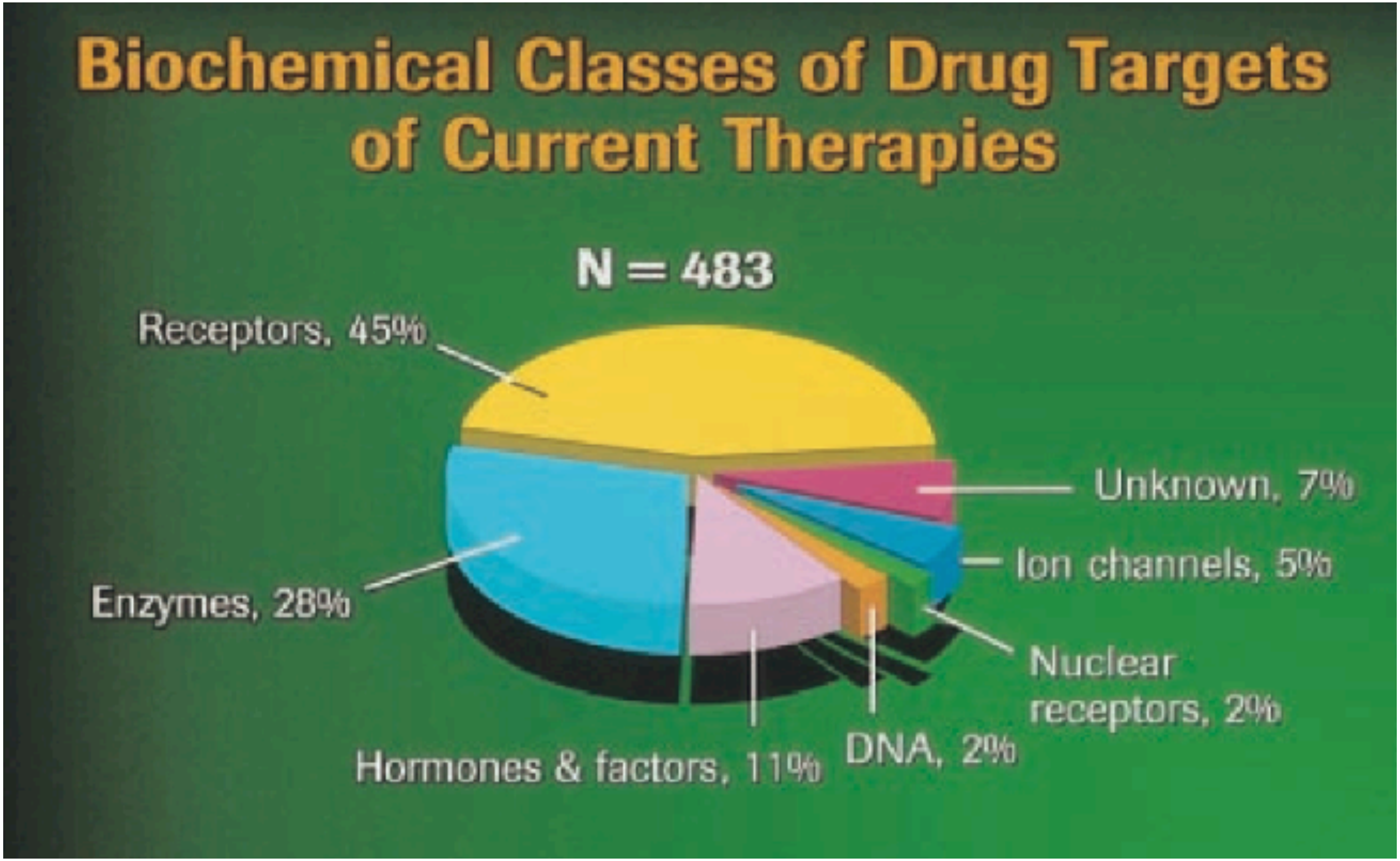
	Advantages	**Limitations**
XRC	Protein size, Accuracy	Must grow crystals, Limited dynamics information, Rare to see hydrogens, Potentially non-physiologic folds, Phase problem, **Cost, Time**
NMR	Solution (no crystals), Some dynamics information, Some sparse-data applications (ie. folding), More physiologic conditions	Size limits, Isotopic labeling required, Assignment problem, **Cost, Time**

Open Computational Problems:

XRC: cryst condition prediction, phasing, model building and refinement

NMR: pulse sequences, assignment, utilizing novel geometric information (ie. RDCs), model building and refinement

Drug Targets

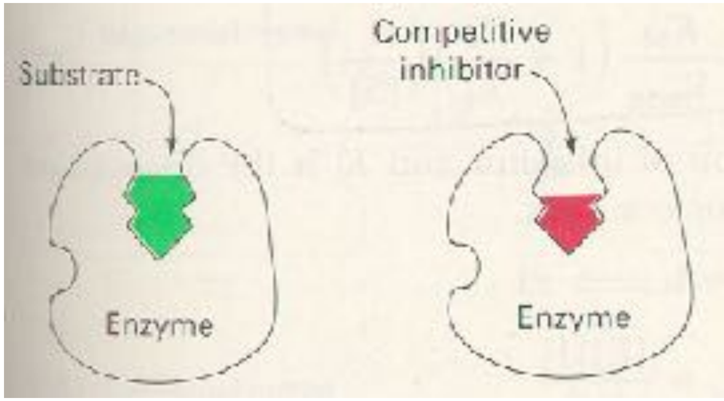


Drews, Science, March 17 2000

Drug Design

Traditional Drug Design

Identify small molecule capable of binding protein active site and inhibiting protein function



Active Site:

- Small compared with rest of protein
- Three dimensional crevice
- Binding specificity based on functional groups of active site residues (obvious)

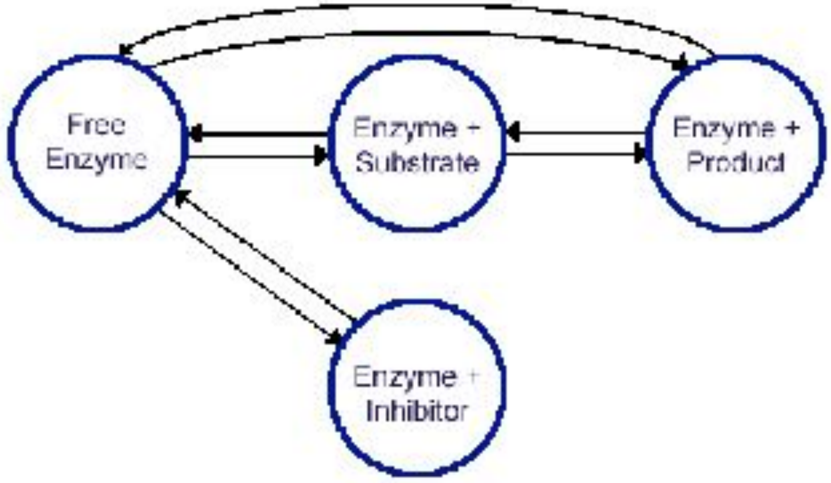
Ligand:

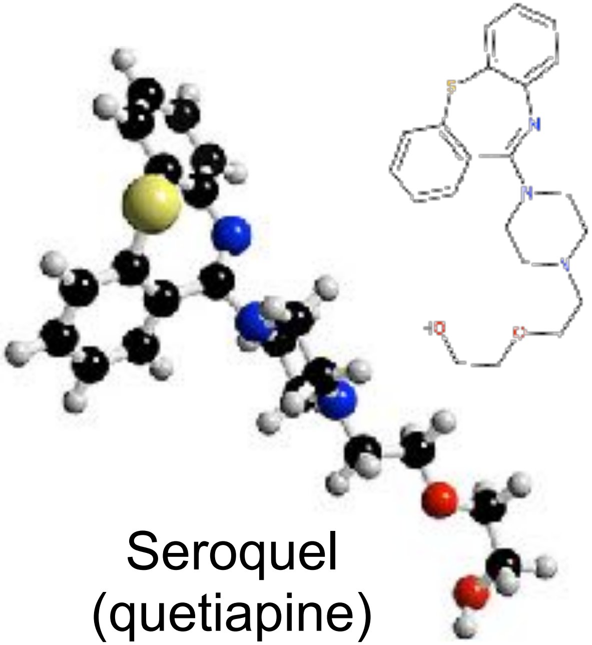
Any small, non-protein molecule capable of binding something

Typically <50 atoms

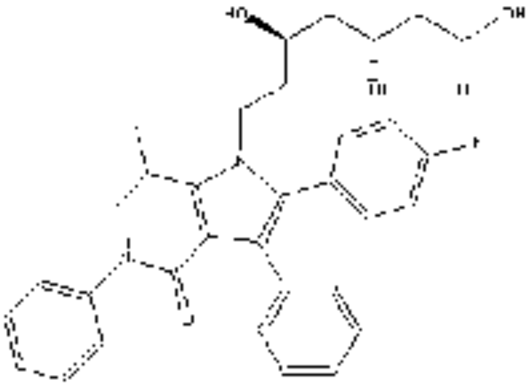
Inhibitors are usually analogs of natural substrate

Reversible Competitive Inhibitor States



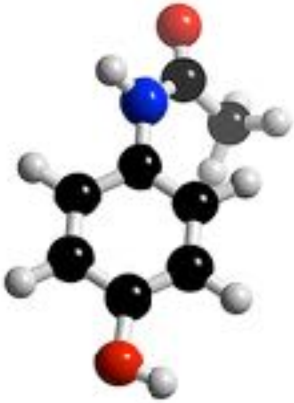


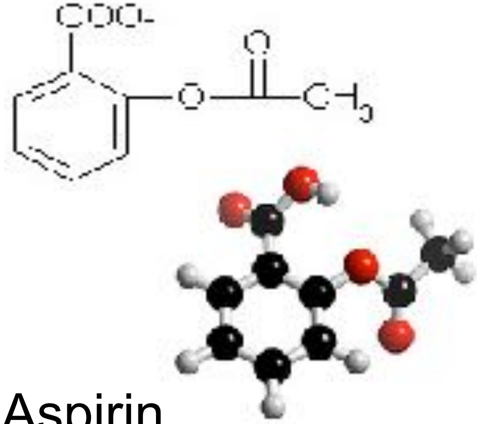
Seroquel
(quetiapine)



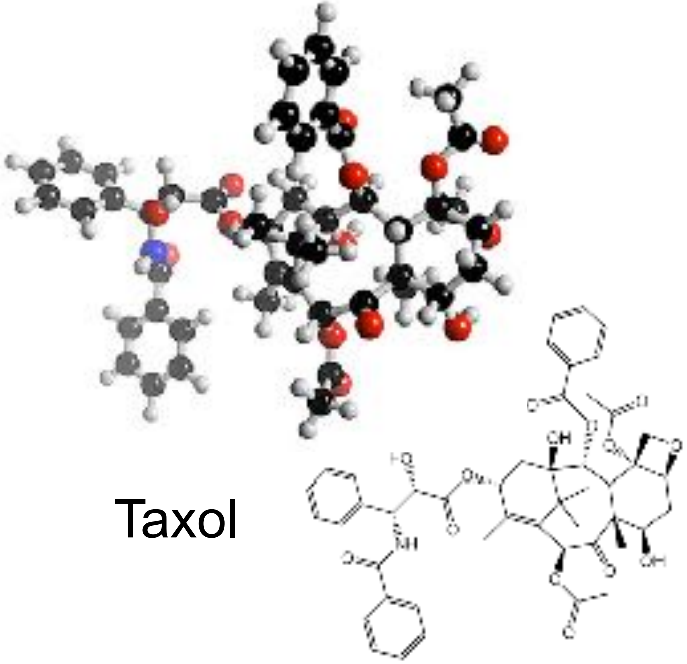
Lipitor
(atorvastatin)

Tylenol
(acetaminophen)

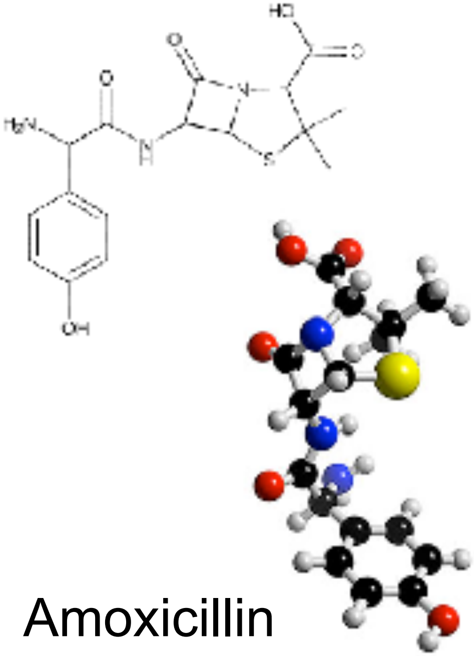




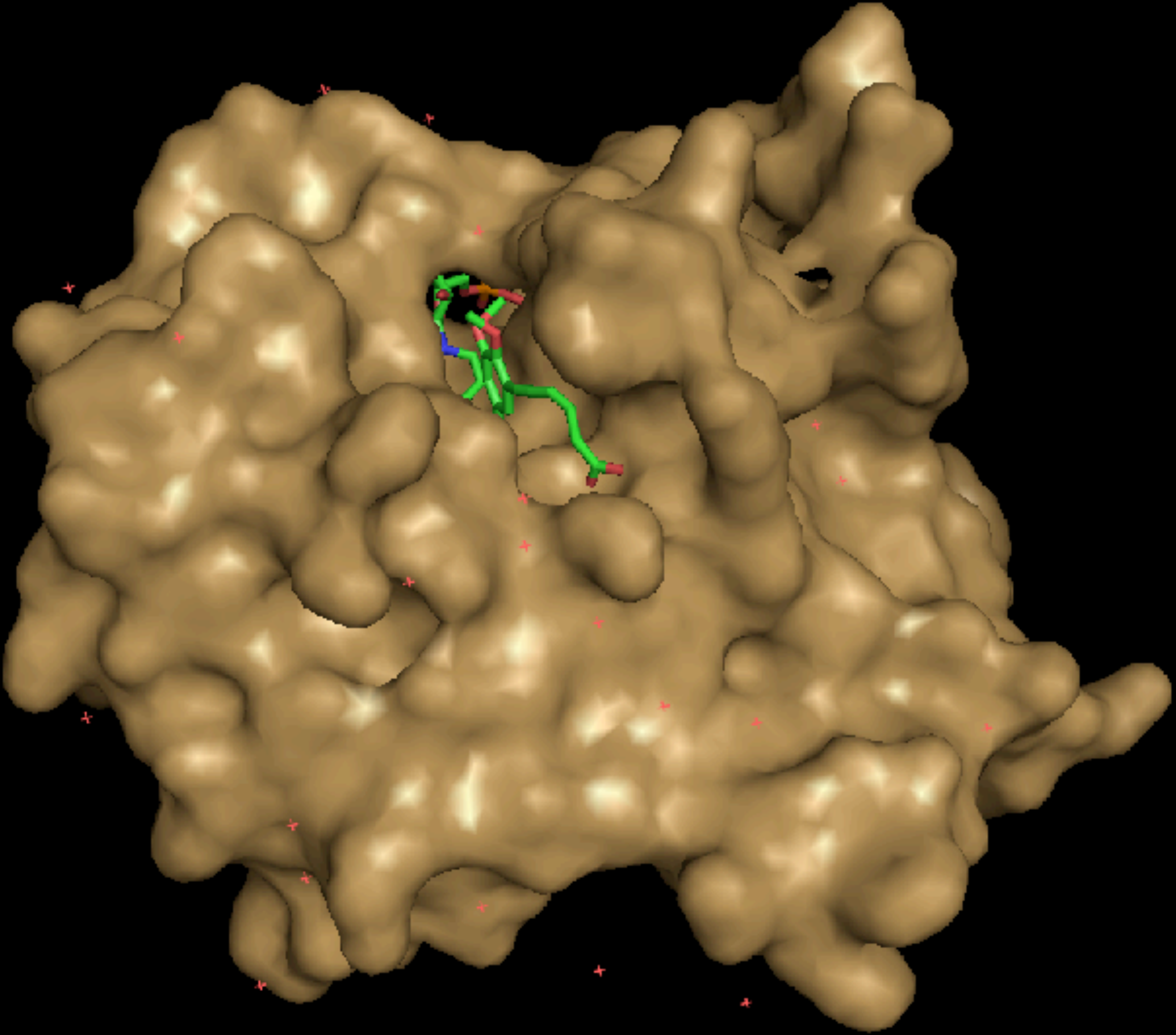
Aspirin

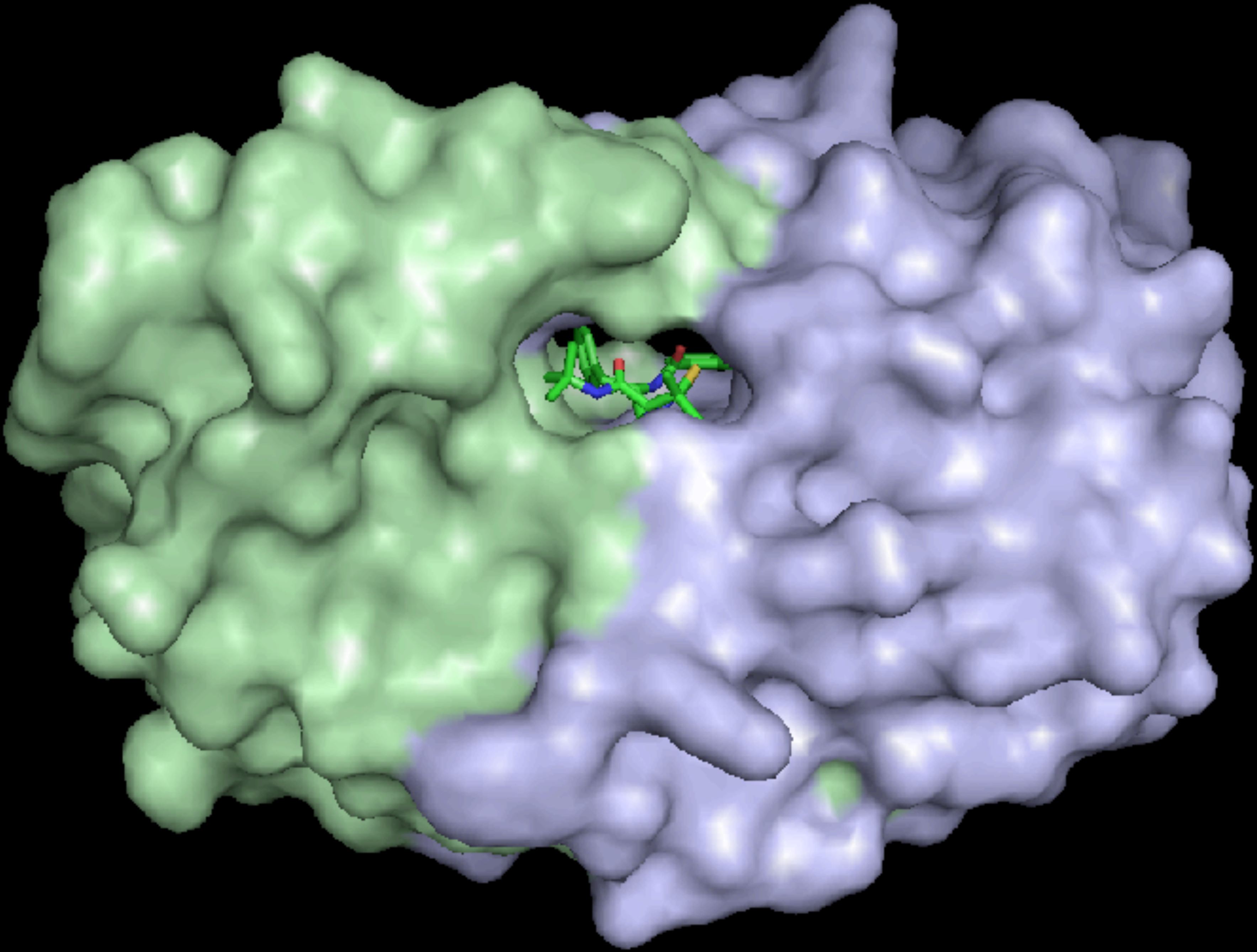


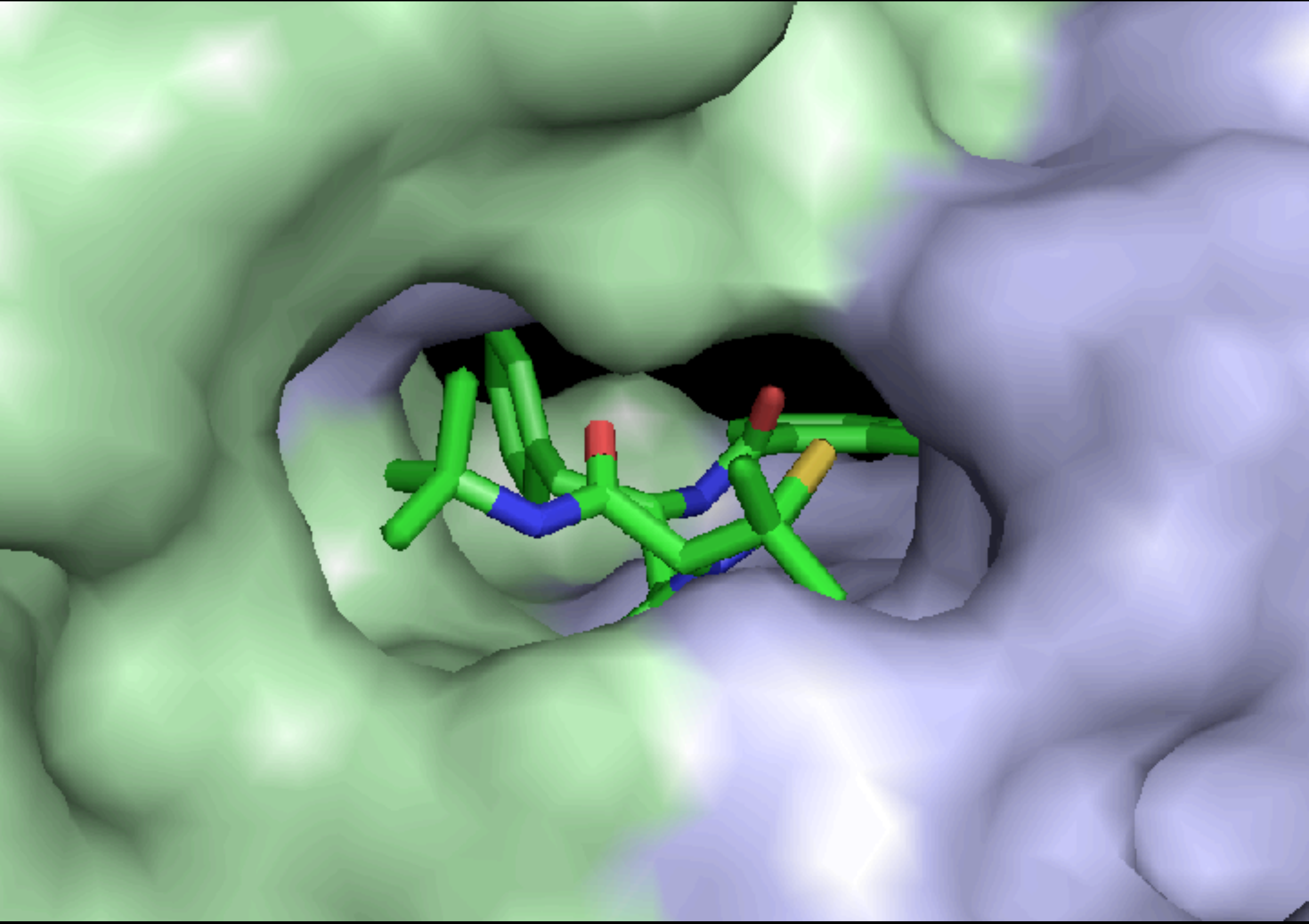
Taxol



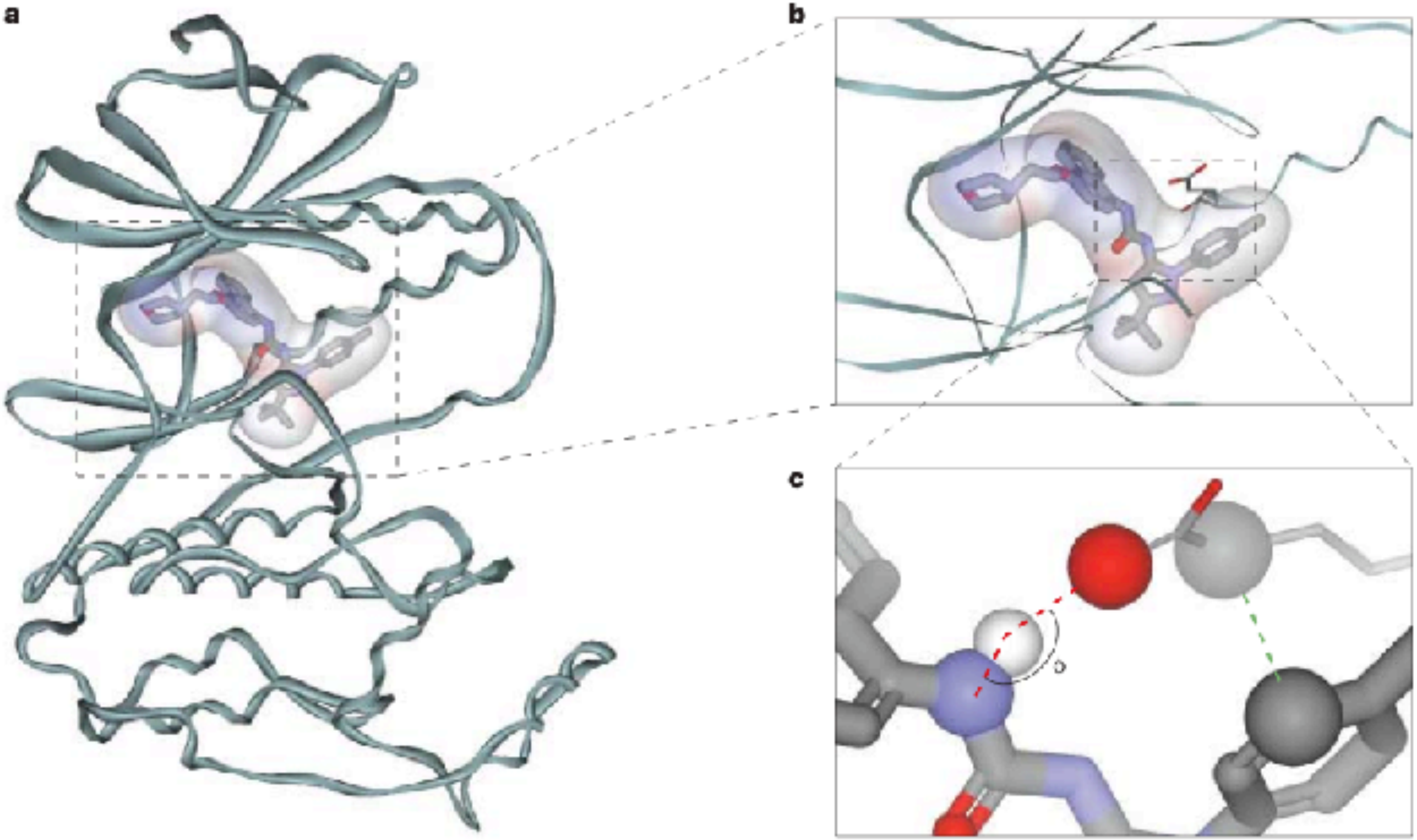
Amoxicillin







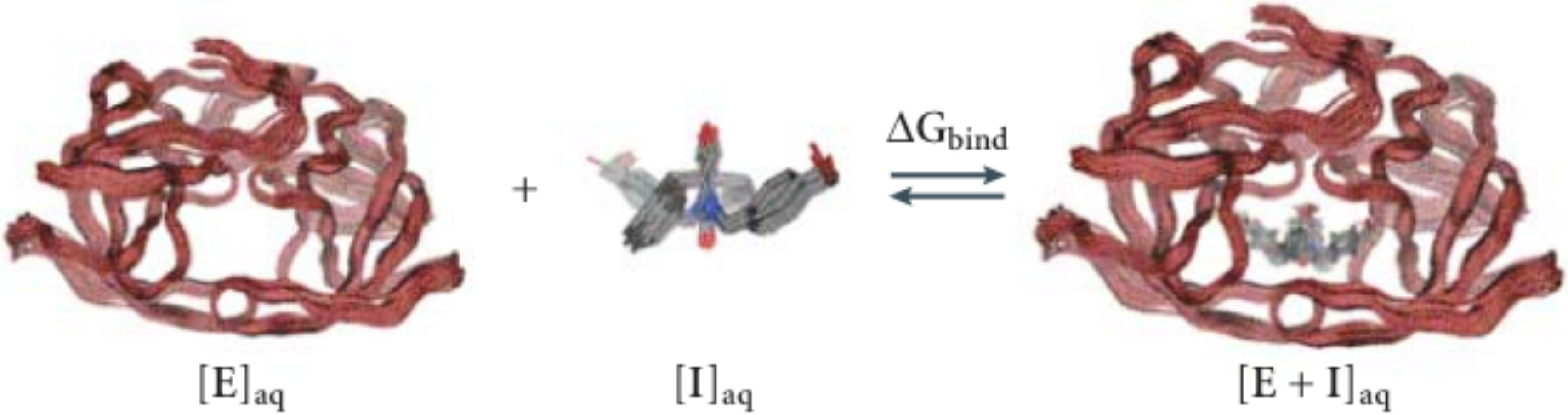
Protein-Ligand Interactions



Kitchen, Decornez, Furr, Bajorath, Nature Reviews Drug Disc, 2004

Protein Ligand Binding

$$[\mathrm{E}]_{\mathrm{aq}} + [\mathrm{I}]_{\mathrm{aq}} \rightleftharpoons [\mathrm{EI}]_{\mathrm{aq}}$$



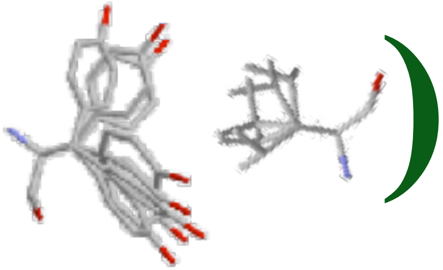
Kitchen, Decornez, Furr, Bajorath, Nature Reviews Drug Disc, 2004

$$\Delta G = -RT \ln K_A \qquad K_A = \frac{1}{K_D} = \frac{[\mathrm{EI}]}{[\mathrm{E}][\mathrm{I}]}$$

Protein Ligand Binding

Maximum Likelihood

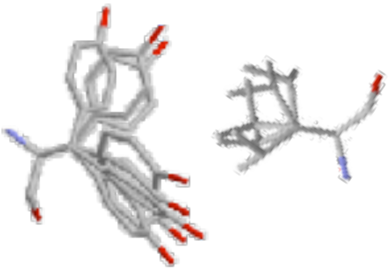
(pick most probable)

$\mathit{min}($  $)$

Global Minimum Energy Conformation

Bayesian

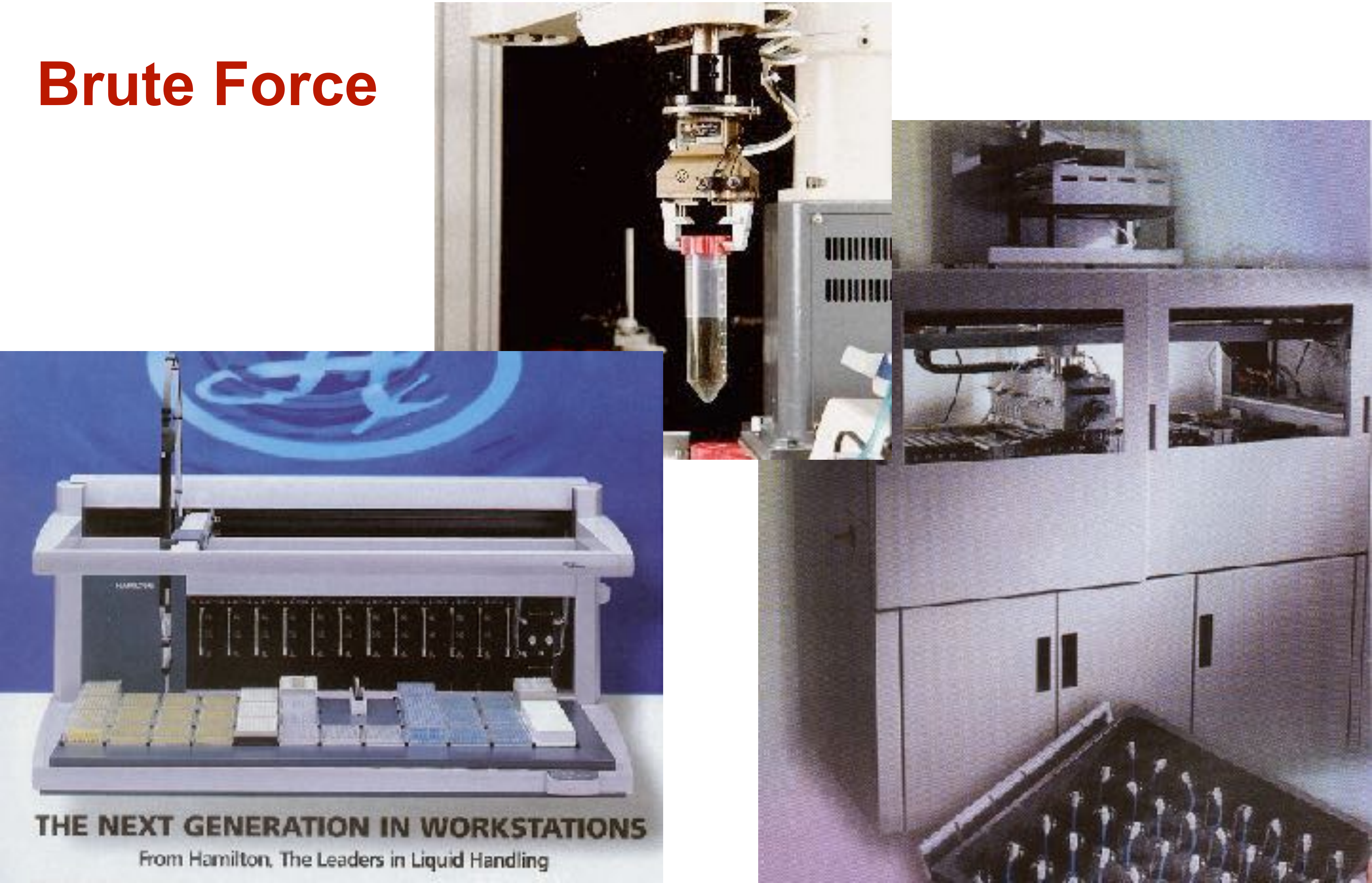
(average over all conformations)

$\frac{1}{Z}\int$ 

Probability ↔ *Energy* using Boltzmann distribution

High Throughput Screening (HTS)

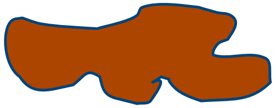
Brute Force


THE NEXT GENERATION IN WORKSTATIONS
From Hamilton, The Leaders in Liquid Handling

SBDD Process

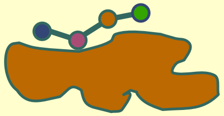
3D Ligand Database

Choice of Protein Target



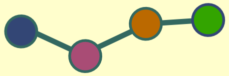
Protein Structure: XRC, NMR, Modeling

Docking



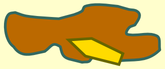
Linking ⟷ Building

Synthesis



Screening

Lead Compound



Struct of Prot-Lig Complex

Redesign, improve affinity, specificity, pharmacodynamics, toxicity

If promising

Pre-Clinical Trials

Clinical Trials

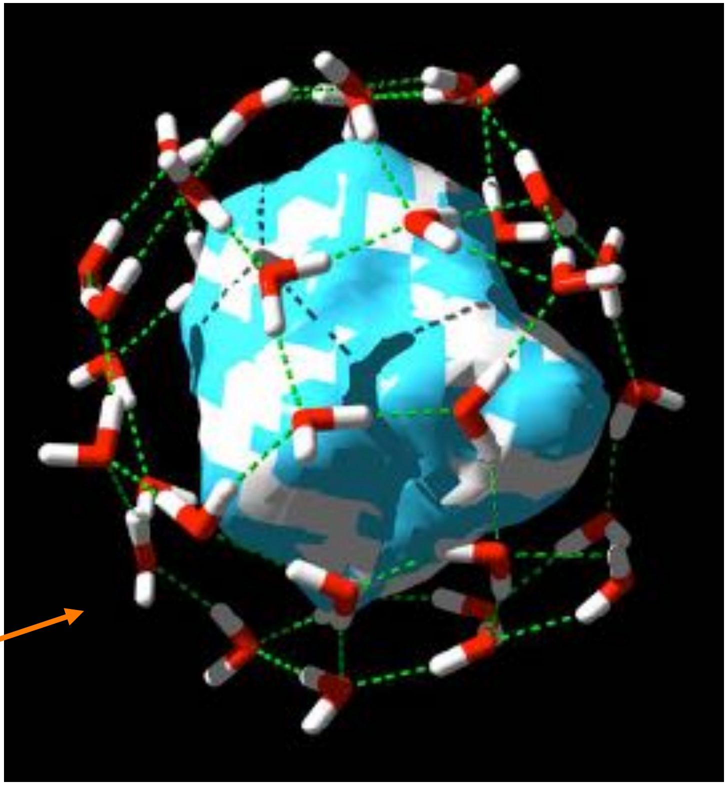
Drug

SBDD Approaches

Structure Based Drug Design

Find (or design) a ***ligand which will tightly bind*** the active site and ***determine where the ligand binds***

Input: Model of AS, set of candidate ligands or fragments, energy function
Output: Set of binding ligands with their bound conformations



Issues

Scoring Function

Flexibility (Backbone/Sidechain)

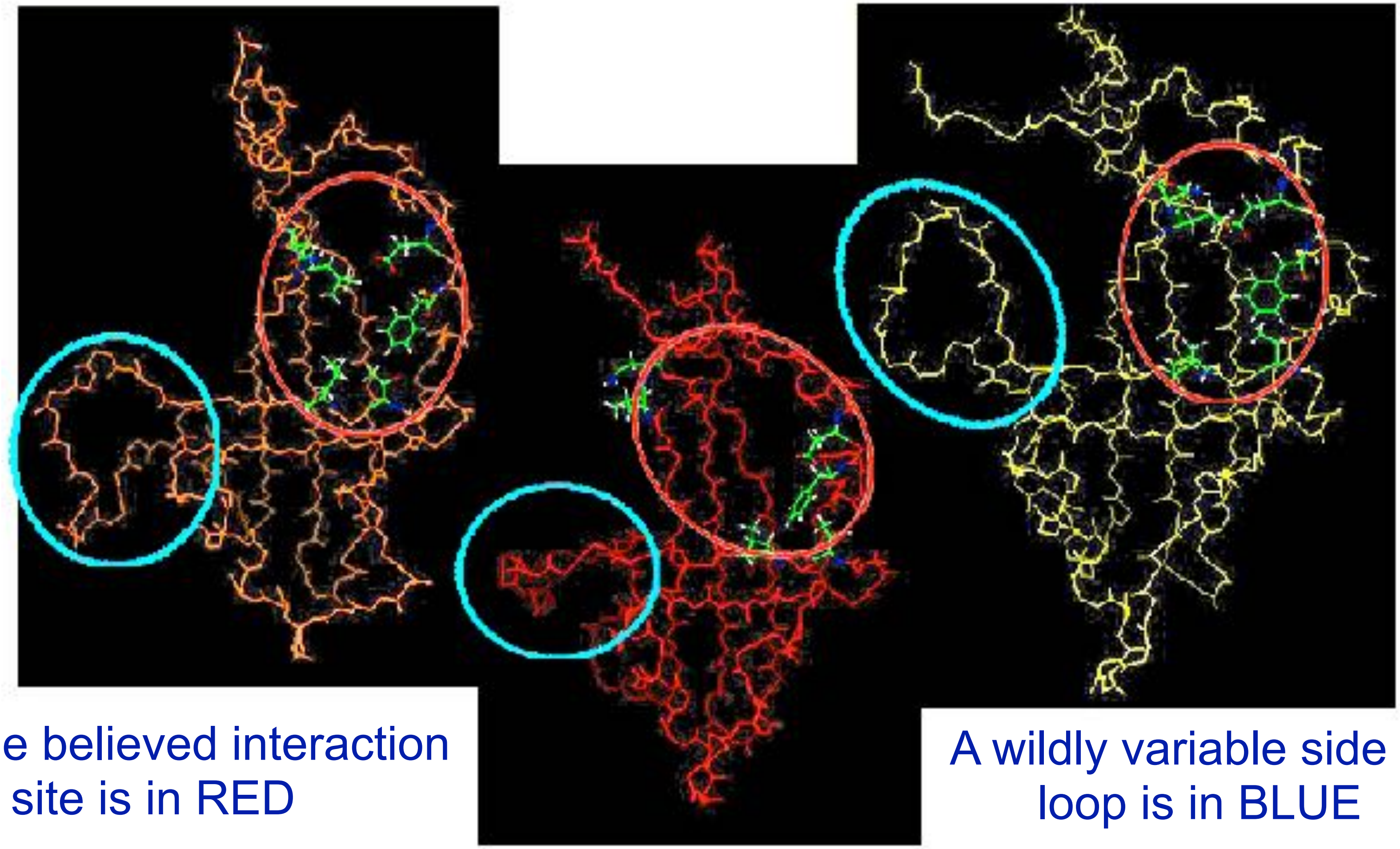
- ligand (rigid / flexible)
- receptor (rigid / flexible)

Solvent Modeling (explicit/implicit)

usually ignored, why?

Molecular Flexibility

3 ‘Snapshots’ of CBFb



The believed interaction site is in RED

A wildly variable side loop is in BLUE

SBDD Approaches

Structure Based Drug Design

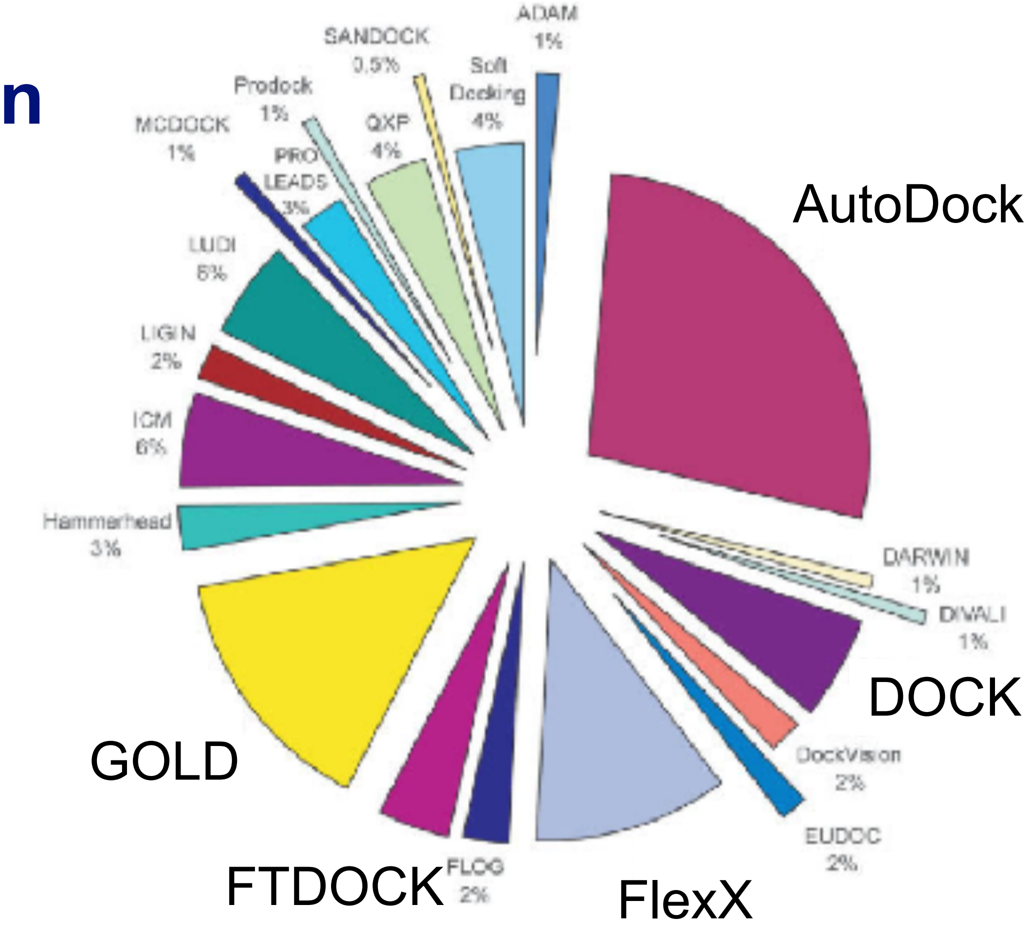
Database Search

Docking - Virtual Screening

De Novo Ligand Design

Building vs. Bridging

Sousa, Fernandes, Ramos, PROTEINS, 2006



Database Search

Screen DB of 100,000 molecules - ***Dock*** ligand into active site

Energy function to evaluate goodness of fit

Ligand score represented by:

Minimum energy over all conformations -
the Global Minimum Energy Conformation (GMEC)

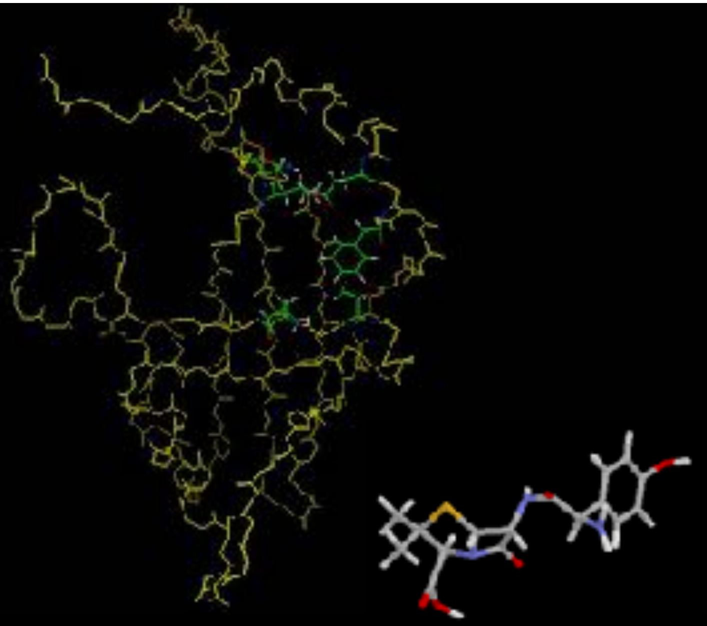
$$\Delta G_{\text{bind}} = \Delta G_{P} + \Delta G_{L} + \Delta G_{PL} + \Delta G_{\text{solvent}} + \Delta G_{\text{entropy}}$$

Direct handle to binding strength

Brute Force

6-DOF Search (no internal DOF)
20x20x20A grid (0.5A spacing)
100-sample points per rotation axis
100^3x40^3 = 6.4x10^{10} conformations

This is one molecule without protein or ligand flexibility



Database Search

Docking Search Methods

Random Methods

Monte Carlo / Simulated Annealing
Genetic Algorithms (state variables ‘genes’)
Tabu Search (avoid previously seen solutions)

Simulation Methods

Molecular Dynamics

Minimization Methods

Energy Minimization (rarely used alone)

Docking Scoring

Empirical Energy function (varying types)
Some with explicit hydrogen-bond terms

Database Search

Ligand Flexibility

Ensemble-Based

Generate multiple conformations of each ligand

Dock each conformation

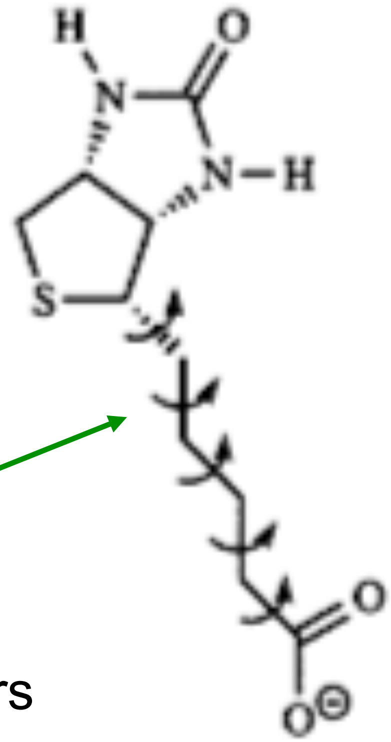
Compute some consensus score (weighted average)

Explicitly Modeled with Hinges

Maintain information on rotatable dihedrals

Allow them to move during docking

May need to utilize ‘rotamers’ to get over energy barriers



Protein (Receptor) Flexibility

Systematic modeling not feasible

Some approaches

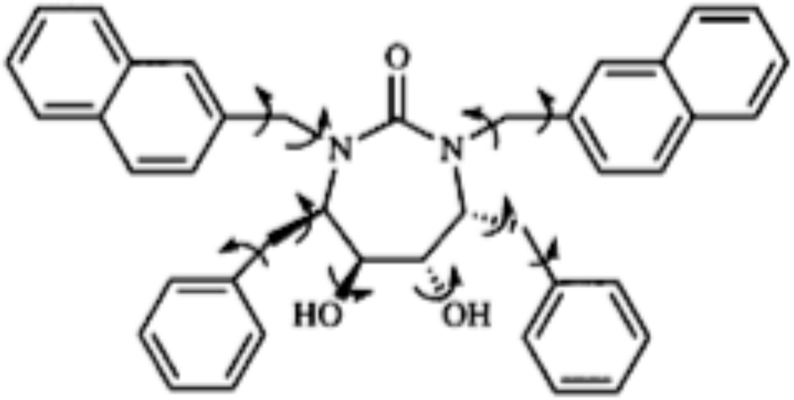
Explicit Backbone vs Sidechain Flexibility

Dock against **Ensemble** (FlexX, FlexE)

Multiple ‘static’ conformations

Harmonic (Normal) Mode Analysis

Soft-Receptors (dampen vdW term)



Docking

AutoDock

Search: Lamarckian Genetic Algorithm
Scoring: 5-term Energy Function (with explicit h-bond term)
Ligand Flex: Random search, MC/SA
Receptor Flex: Sidechain Flexibility
Notes: Freely available to academic community

DOCK

Search: GA, First fragment placed via sterics, grow
Scoring: 3 scoring functions (none with explicit h-bond term)
Ligand Flex: Systematic, Fragment-based flexibility (incremental)
Receptor Flex: Limited, Can now dock to ensembles
Notes: Very fast, but limited accuracy, Free to academics

GOLD

Search: Genetic Algorithm
Scoring: Empirical Energy Function (with explicit h-bond term)
Ligand Flex: Random search, GA
Receptor Flex: Limited

Docking

Performance

Decent at ***enrichment***

Not so good at absolute binding strength

Most able to predict known protein-ligand poses with 1.5-2A RMSD 70-80% of the time

Performance drops dramatically with >7 rotatable bonds

Only 20-30% within 1.5-2A

No major methodology change over past 10 years

Challenges

Scoring function

Solvent modeling

Deterministic search (better branch-bound algorithms)

Micro-Flexibility (Multi-resolution rotamers?)

Macro-Flexibility (NMR?, Harmonic Mode Analysis?)

de novo

General Scheme

- Based on identification and satisfaction of ***interaction sites***
- **Select interaction sites**
- **Satisfy interaction sites** with functional groups
- **Join functional groups** (Bridging technique)
- **Refine structure**

h-bond donors
h-bond acceptors
electrostatic
hydrophobic

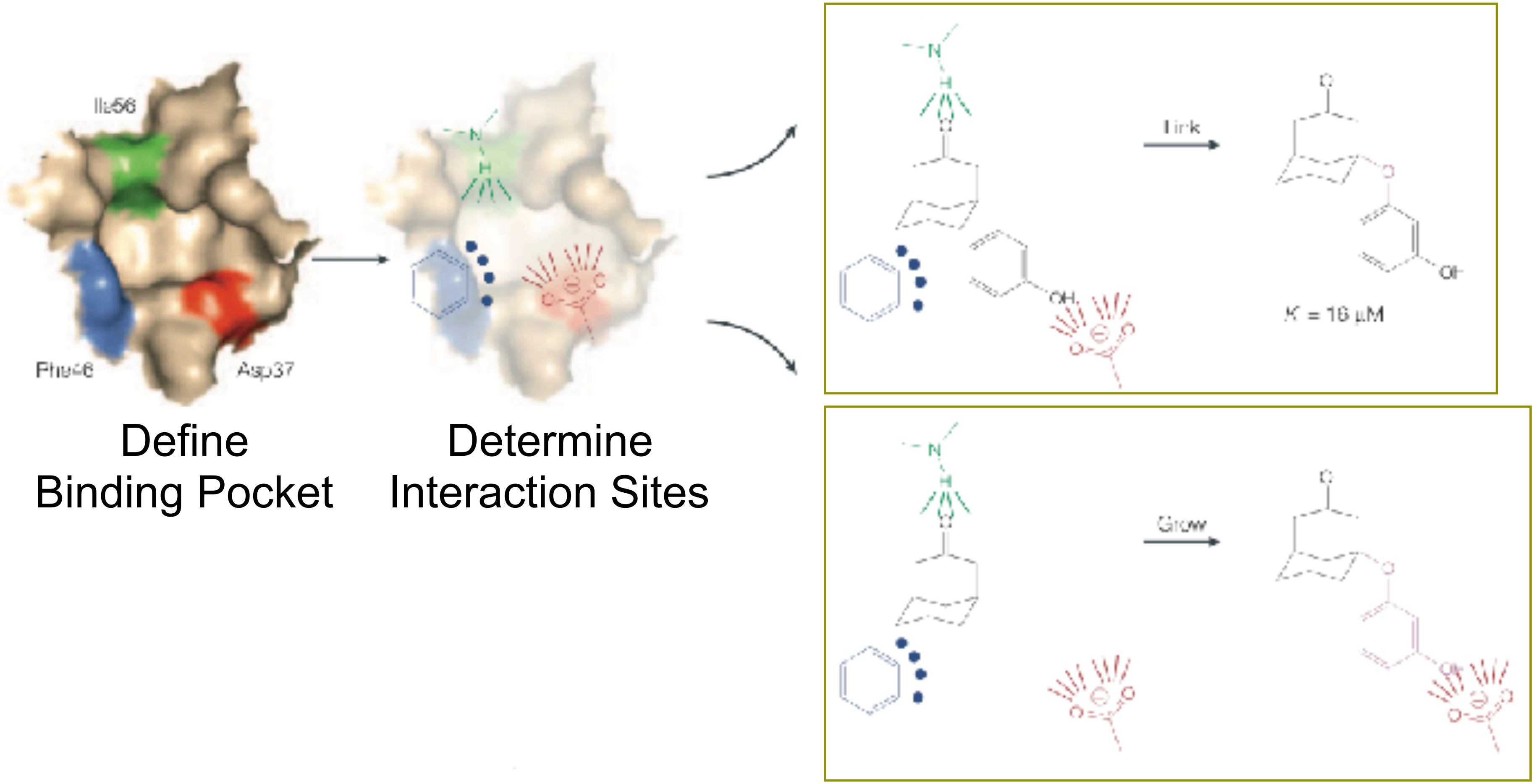
Building Methods (Grow methods)

Start with seed fragment
Selectively add atoms (fragments)

Bridging (Linking) Methods

Dock multiple fragments
Connect by bridging

de novo



Define Binding Pocket

Determine Interaction Sites

Bridging (Linking) Methods

Building (Grow) Methods

Schneider, Fechner, Nature Rev Drug Disc, 2005

de novo

Major Challenges

- Problems when interaction sites are far away
- Very difficult to model receptor flexibility
- Synthetic accessibility
- Suggested molecules may not be chemically stable
- Pharmacodynamic / Pharmacokinetic properties of ligands

Components / Parameters

Building Blocks: Atoms vs. Fragments
Search Strategy: Deterministic (DFS, BFS), Random (MC, GA)
Construction: Bridging vs. Building
Scoring Function: Empirical Energy Force Field

de novo - Buildup

Monte Carlo de Novo Ligand Generator (MCDNLG)

Building Blocks: Atom

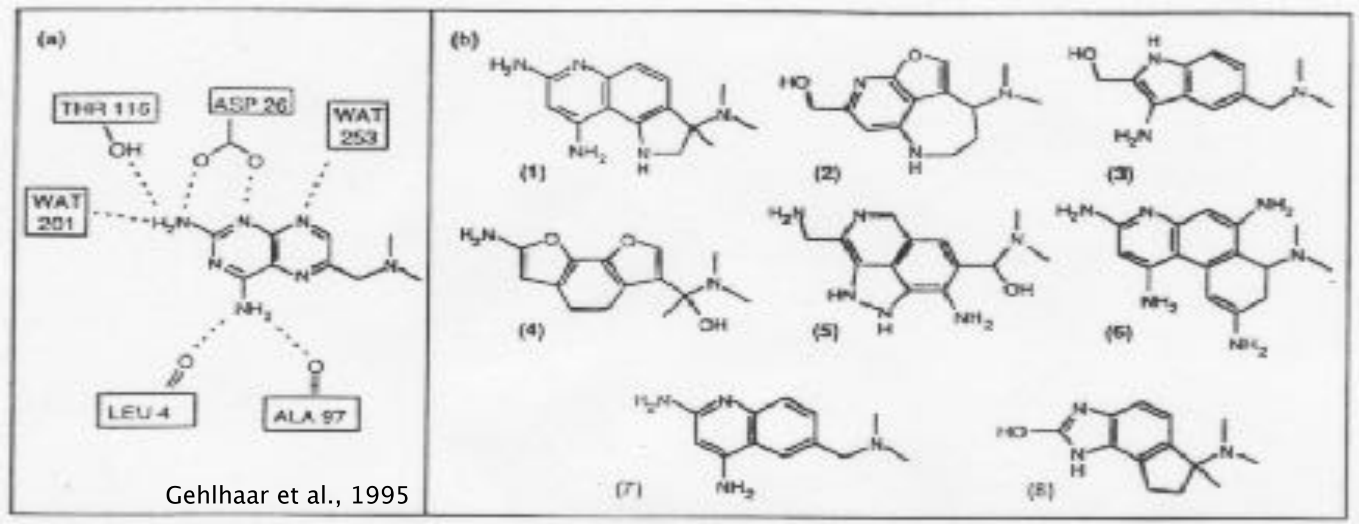
Search Strategy: Random MC

Active Site starts filled with Carbons

Monte Carlo Steps

- Change atom occupancy (on/off)
- Change atom position
- Change bond type (off/single/double)
- Change atom type (C,N,O)
- Rotate/Translate a fragment

Heuristic Penalties and Rewards
300,000 steps in typical run



Gehlhaar et al., 1995

de novo - Buildup

GROW

Building Blocks: Fragments

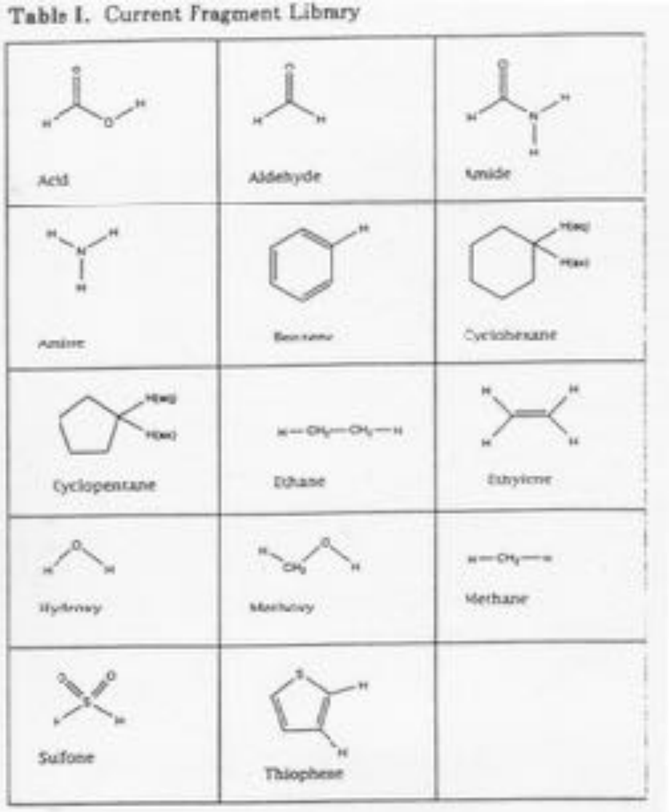
Search Strategy: Beam Search

Attach new fragment

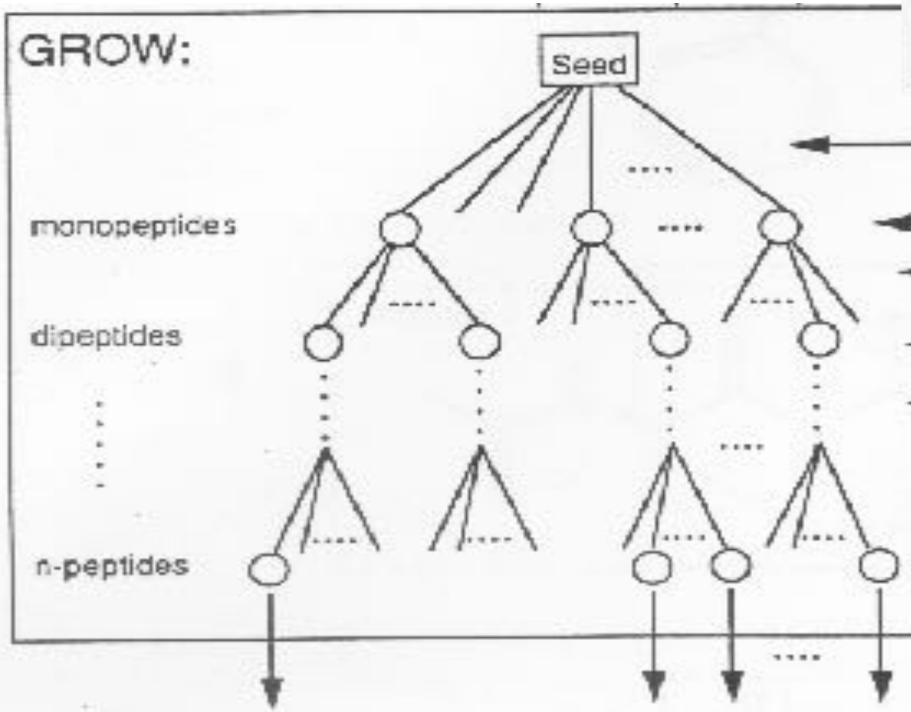
Rotate around new bond

Energy minimize

Table I. Current Fragment Library



Acid	Aldehyde	Amide
Amine	Benzene	Cyclohexane
Cyclopentane	Ethane	Ethylene
Hydroxy	Methoxy	Methane
Sulfone	Thiophene	



A: attach each template to seed: score

B: keep 10 best constructs

C: attach each template to each construct kept: score

D: keep 10 best

E: iterate over C and D

F: stop at requested peptide length, keep 10 best

Moon and Howe, 1991

de novo - Bridging

SPROUT

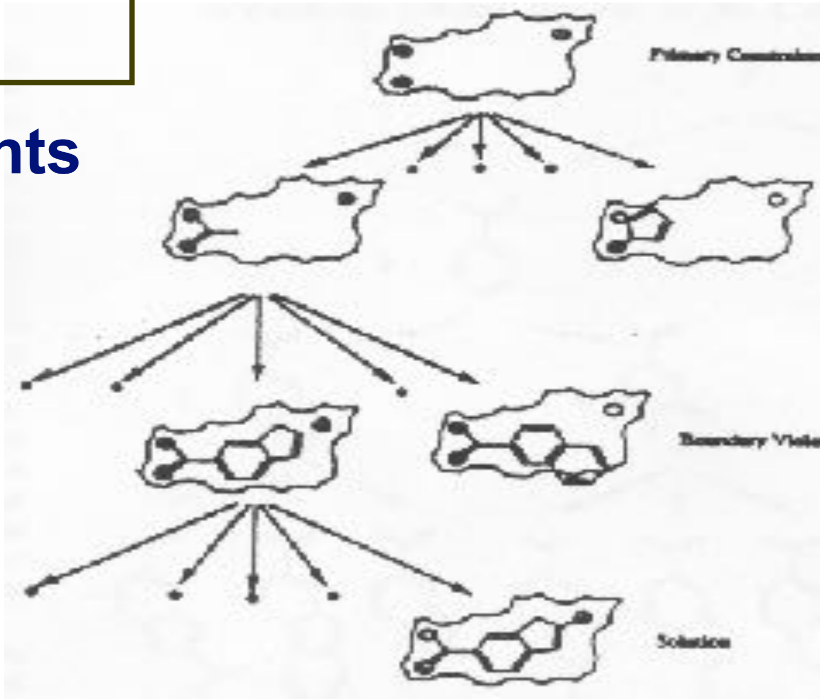
Building Blocks: Fragments
Search Strategy: DFS/BFS, A* Search

Find ‘target sites’

Known ligand binding site
Manual ligand docking
Multiple Copy Simultaneous Search (MCSS)
Pharmacophore

Generate Skeletons of 3D Fragments

- No notion of element type
- Anchor one vertex of template, rotate (15^o) increments
- Continue to add fragments until some fraction of sites linked
- All templates added in all ways
- A* search (branch-and-bound)

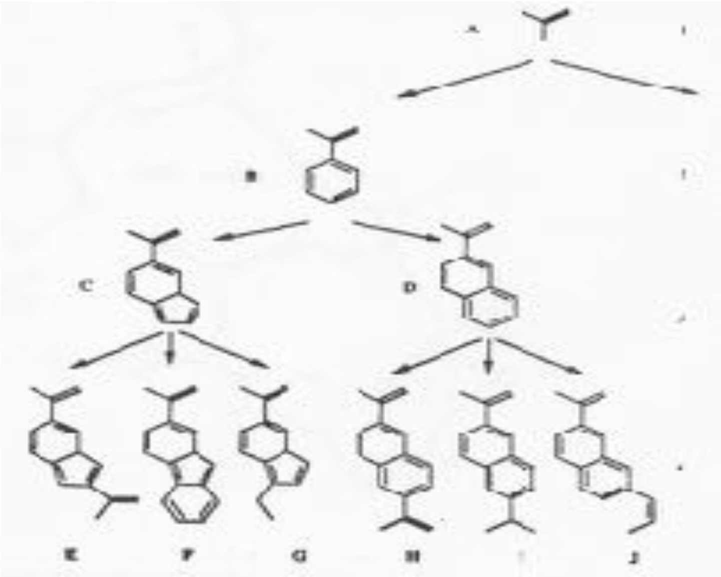


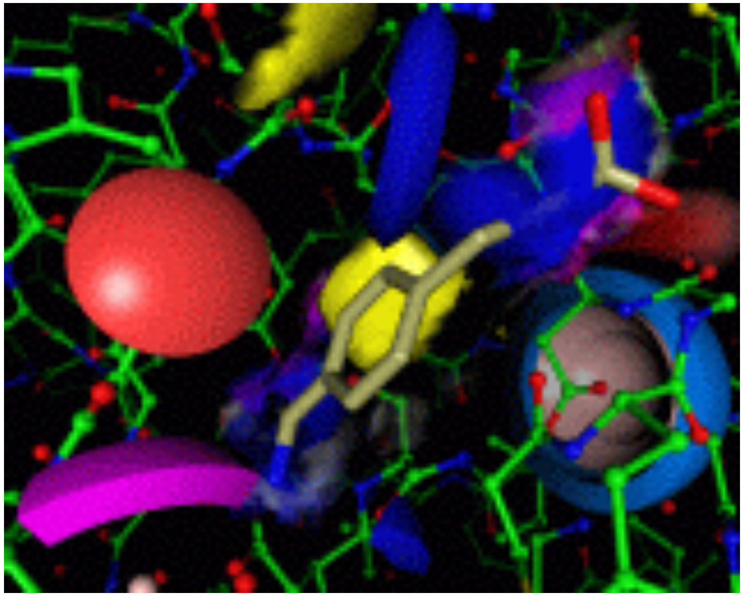
de novo

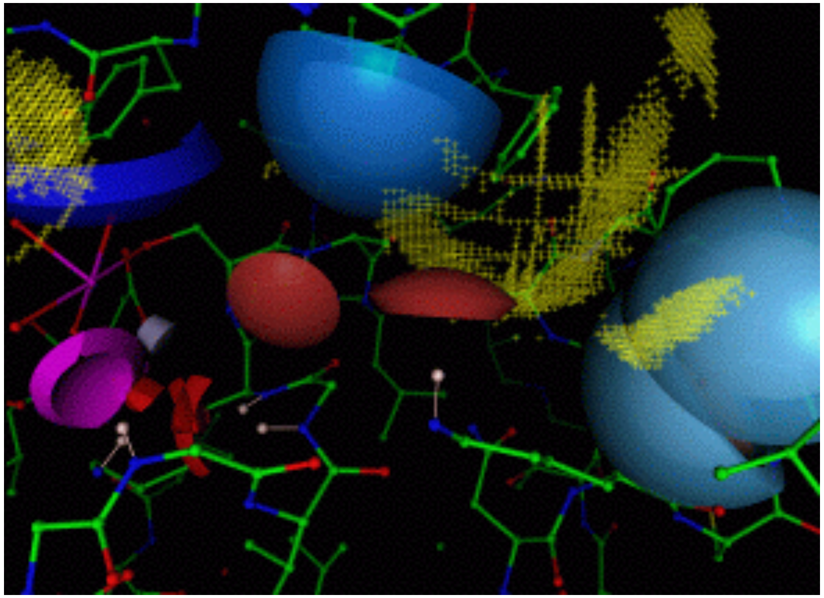
SPROUT

Substitute Real Atoms into Skeleton

- Based on binding character (H-Donor/Acceptor)
- Conformations grouped by common ancestors





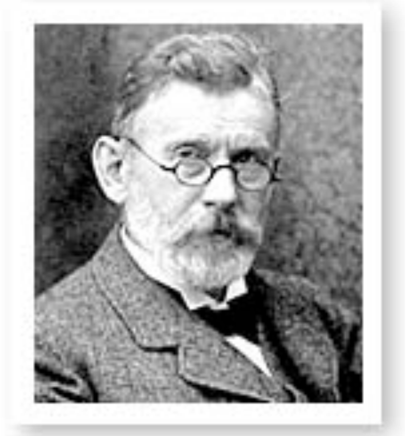


http://chem.leeds.ac.uk/ICAMS/SPROUT/zsolt/sprout_galery.html

Pharmacophores

Pharmacophore:

A molecular framework that carries (phoros) the essential features responsible for a drug’s (=pharmacon's) biological activity -Paul Ehrlich



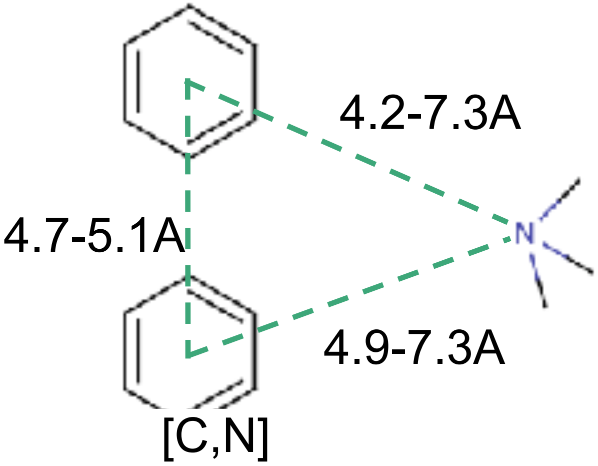
Paul Ehrlich (1854–1915)

Useful when

Active Site structure unknown

Have Positive and Negative Ligand Examples

3D Pharmacophore

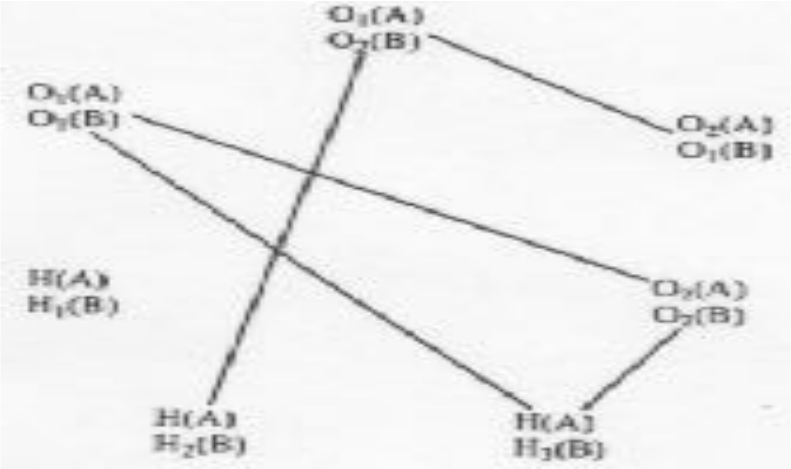


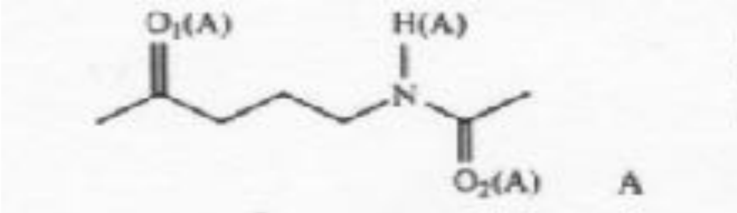
Can Reduce Pharmacophore Matching Problem to Clique

Pharmacophore as Clique

Start with set of Active Molecules

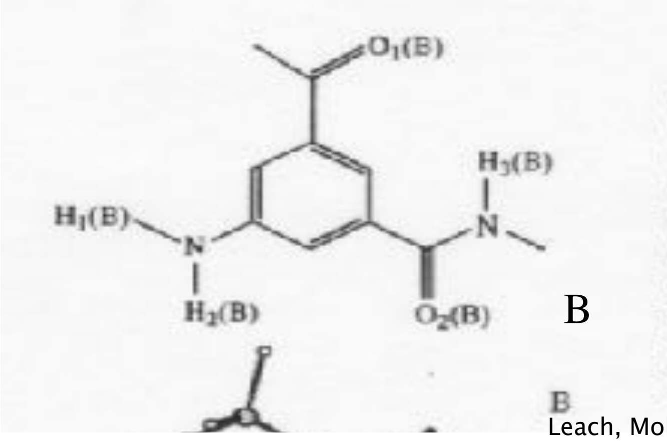
- We don’t know which functional groups actually bind nor which distances are favored
- Nodes are equivalent functional groups
- Edges are between distance consistent functional groups





O_1–H = 4.73 Å
O_1–O_2 = 6.84 Å
O_2–H = 3.14 Å

A



O_1–O_2 = 6.37 Å
O_1–H_1 = 6.22 Å
O_1–H_2 = 6.88 Å
O_1–H_3 = 4.19 Å
O_2–H_1 = 6.06 Å
O_2–H_2 = 4.93 Å
O_2–H_3 = 3.14 Å
H_1–H_2 = 1.73 Å
H_1–H_2 = 7.06 Å
H_2–H_3 = 6.64 Å

B

Leach, Molecular Modeling, 2003

Cliques represent sets of common (mutually consistent) features

Pharmacophore

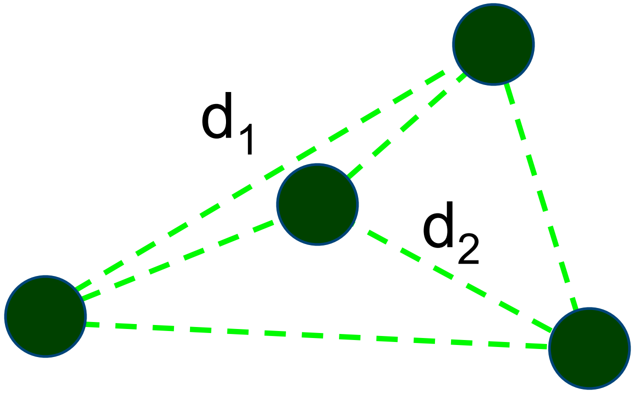
Constrained Systematic Search

Goal: Identify arrangements of functional groups accessible to all positive binding examples

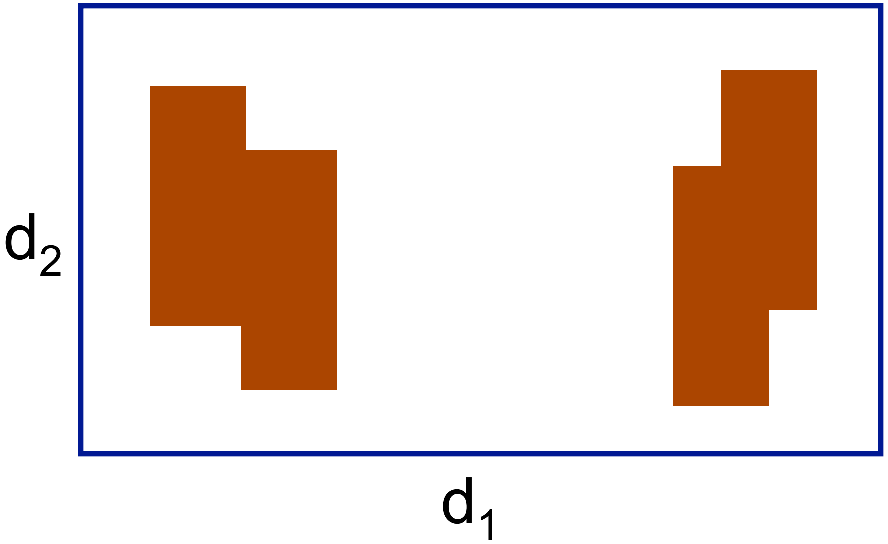
Determine regions of *k* dimensional hyperspace accessible for first molecule

For n^{th} molecule, determine torsion angles that place functional groups in allowed regions

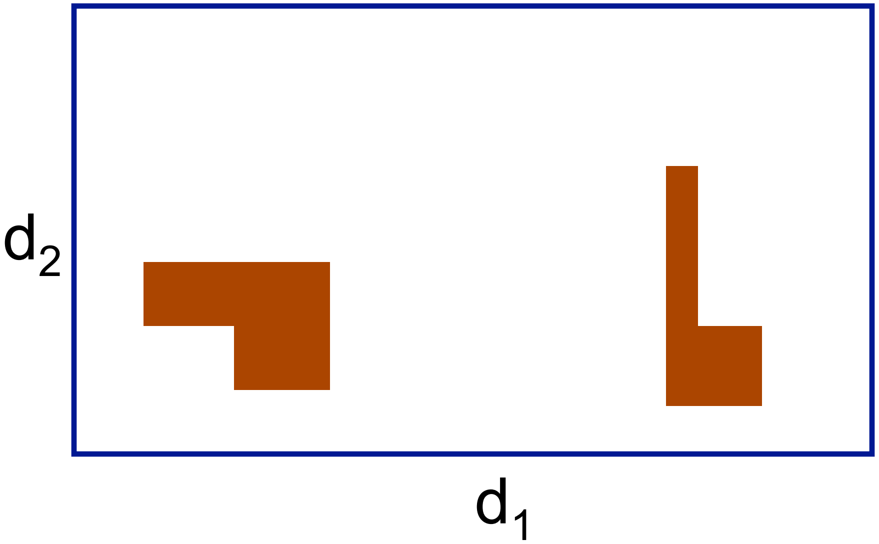
Intersect, Maintain common regions



Molecule 1



Molecule 1 and Molecule 2



Extensions

Pharmacokinetics / Pharmacodynamics

ADMET

Absorption
Distribution
Metabolism
Excretion
Toxicity

ADMET problems kill most drugs

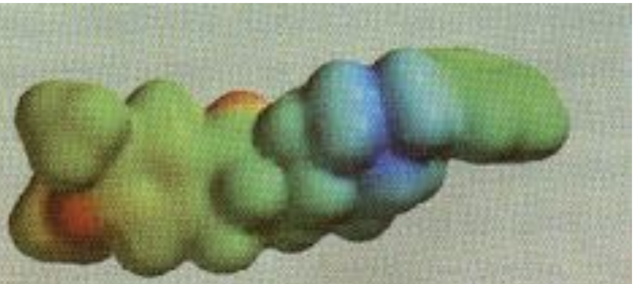
WARNING
JUMPING INTO TOXIC WASTE DOES NOT GIVE YOU SUPERPOWERS
MAY CAUSE ADVERSE HEALTH EFFECTS - SEE FIG 1A
FIG 1A

www.netlash.com

Lead Optimization

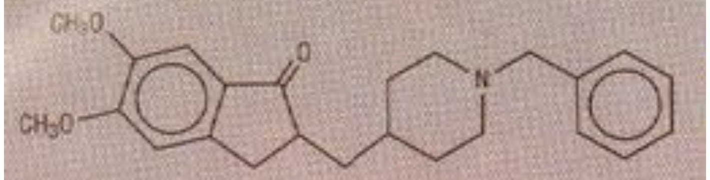
Given lead compound (virtual screening, HTS)
Suggest changes to improve binding
May or may not have structure of lead bound active site

Some Successes...

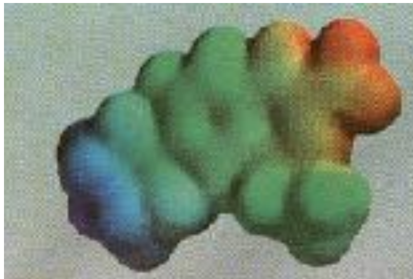


Alzheimer's disease treatment

Molecular modeling, QSAR, molecular shape analysis, and docking played a role in the discovery of donepezil hydrochloride, an acetylcholinesterase inhibitor (18). Eisai markets this compound as Aricept for patients with Alzheimer's disease.

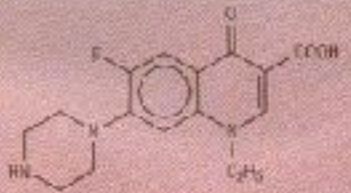


Donepezil

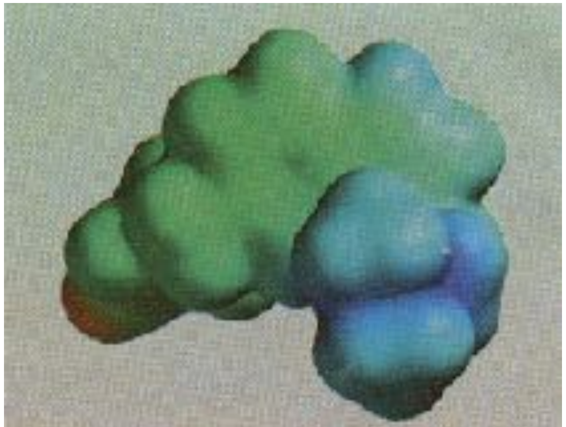


Antibacterial agent

The earliest example of a compound designed using rational techniques, to my knowledge, is norfloxacin. Structural modifications that led the chemists at Kyorin Pharmaceutical Co. to this compound were made with the assistance of QSAR (16). The compound has been on the market since 1983 under various brand names including Noroxin. Spurred by this advance, the 6-fluoroquinolones became a major class of antibacterial agents.

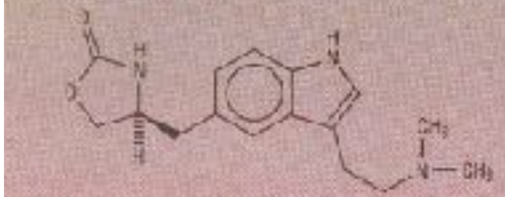


Norfloxacin



Migraine treatment

Zolmitriptan, a drug for migraine, is a 5 HT_{1D} agonist: it was discovered at Wellcome and is marketed by Zeneca under the brand name Zomig. Molecular modeling and the active analogue approach helped define the pharmacophore (21).



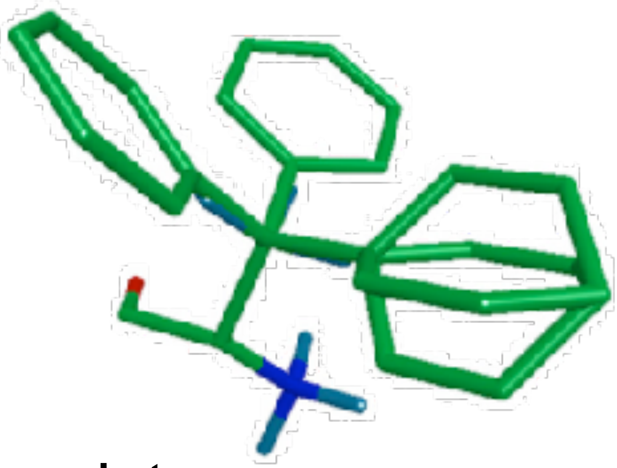
Zolmitriptan

Modern Drug Discovery, Nov/Dec 1998

Protein Design

Suggest a sequence of amino acids capable of folding into a desired conformation or possessing a desired function

Inverse protein folding problem



Two Problems

De Novo Design

Very Difficult

ReDesign

Use of existing protein (backbone) template

Improve (thermal) stability

Change substrate

Protein design with the use of rotamers and a pairwise energy function is NP-Hard

Typically Maximum Likelihood:

For each mutation sequence look for the Global Minimum Energy Conformation (GMEC)

Dead End Elimination

One of the only deterministic, non-trivial, and effective combinatorial optimization algorithms in Computational Structural Biology

Prunes rotamers that are provably NOT part of the GMEC

Used For

Side-Chain Placement (tertiary structure prediction)
Protein Design

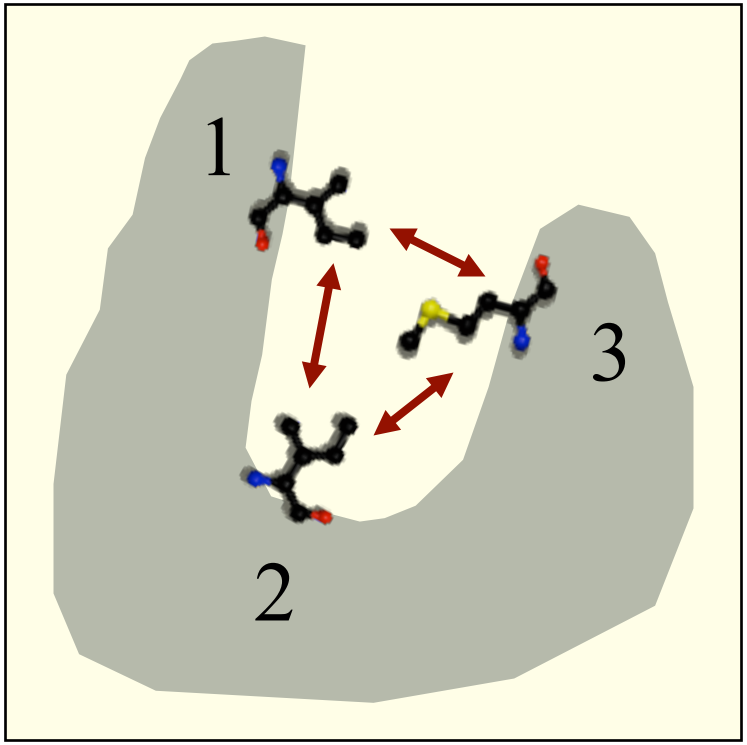
Original DEE

$$E(i_r) + \sum_{j \neq i}^{N} \min_{s} E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max_{s} E(i_t, j_s)$$

Dead End Elimination

Total Energy

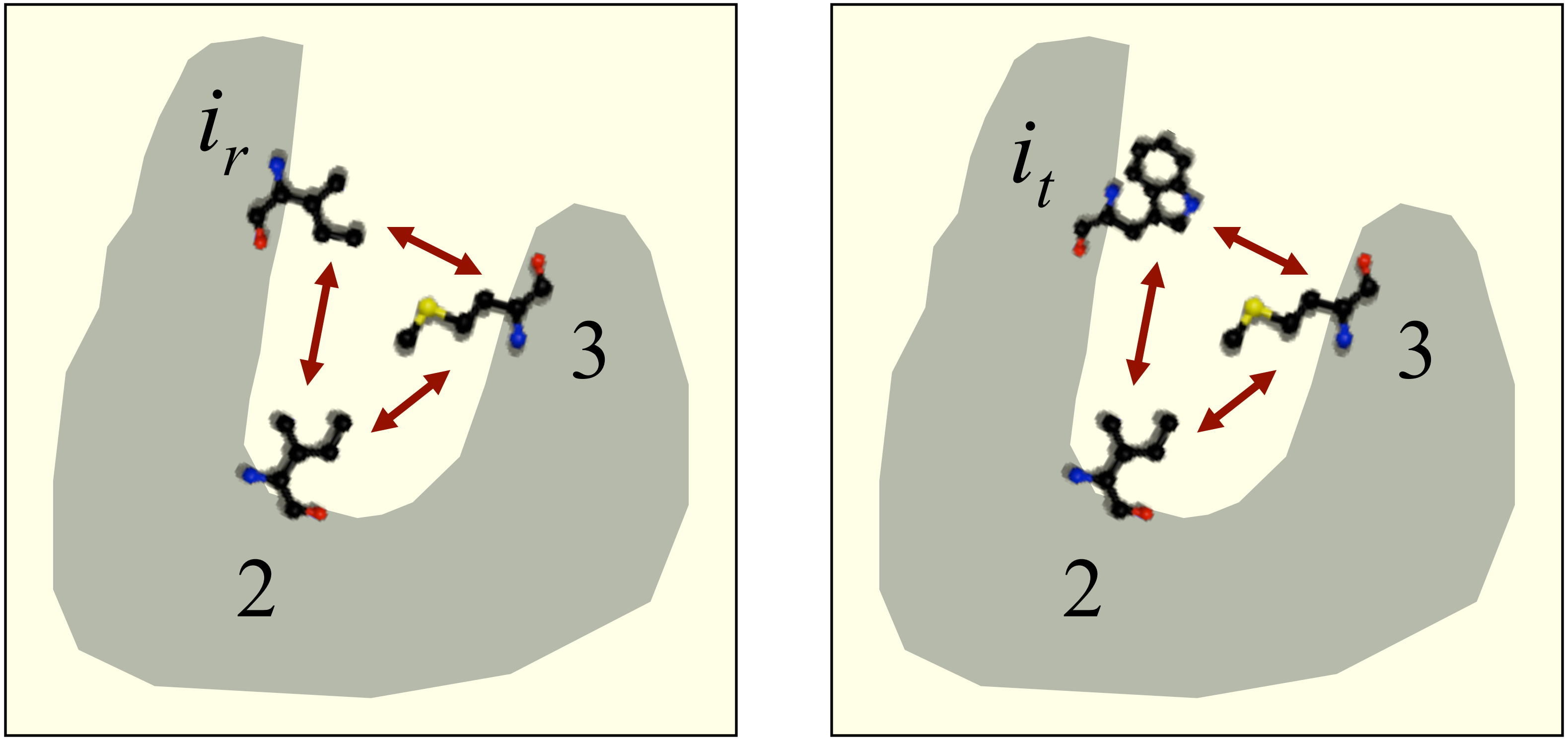
$$E_T = \sum_i \sum_j E(i_r, j_s); \qquad i < j$$



Dead End Elimination

Total Energy

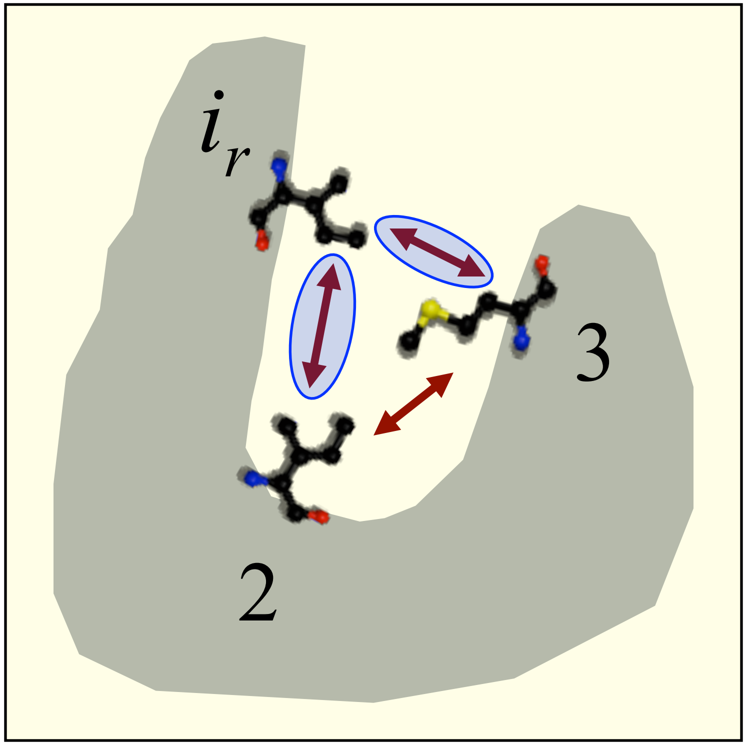
$$E_T = \sum_i \sum_j E(i_r, j_s); \qquad i < j$$

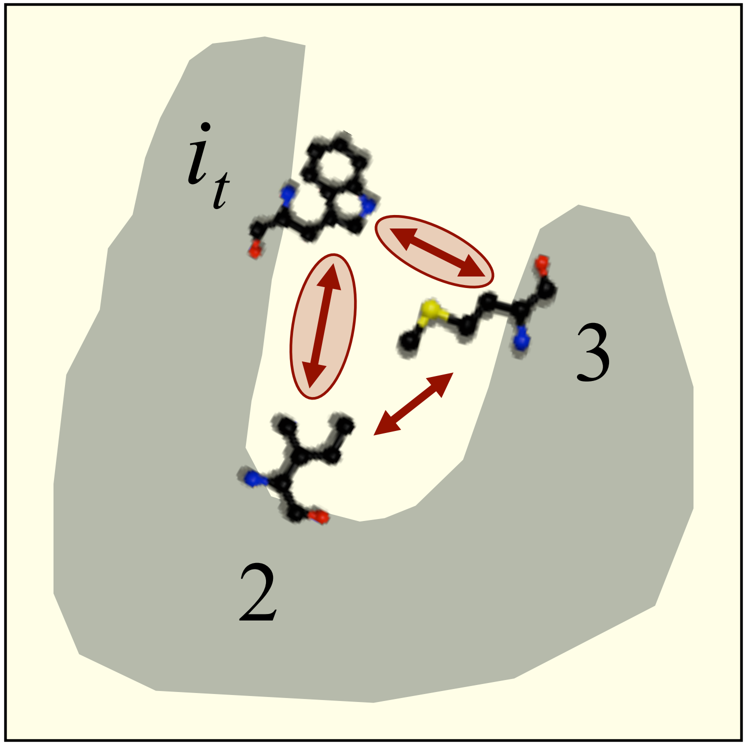


Dead End Elimination

Total Energy

$$E_T = \sum_i \sum_j E(i_r, j_s); \qquad i < j$$

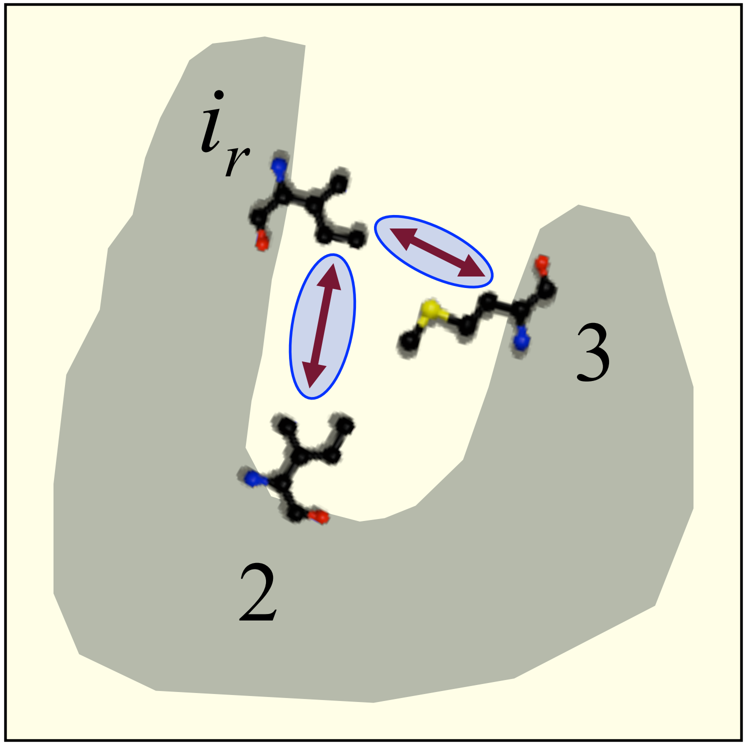


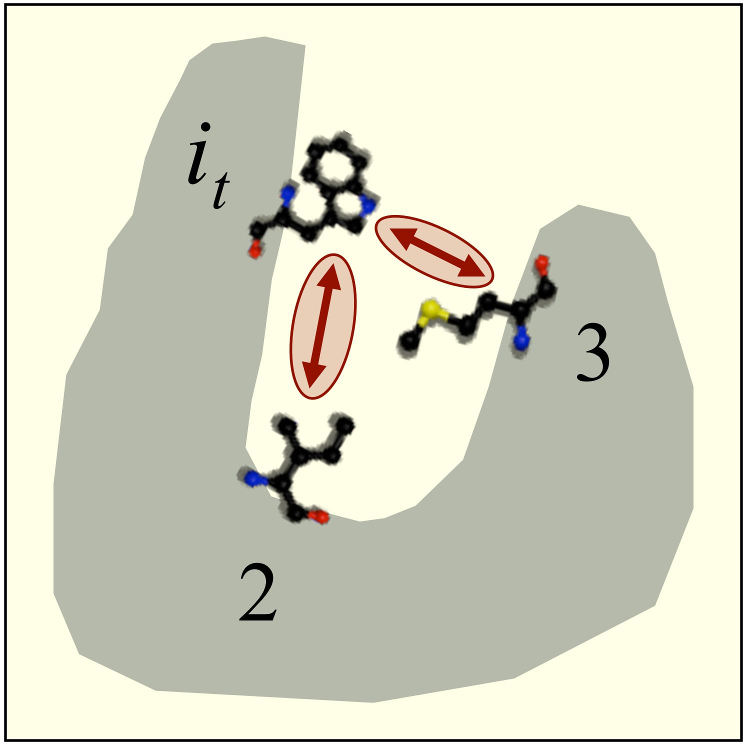


Dead End Elimination

Total Energy

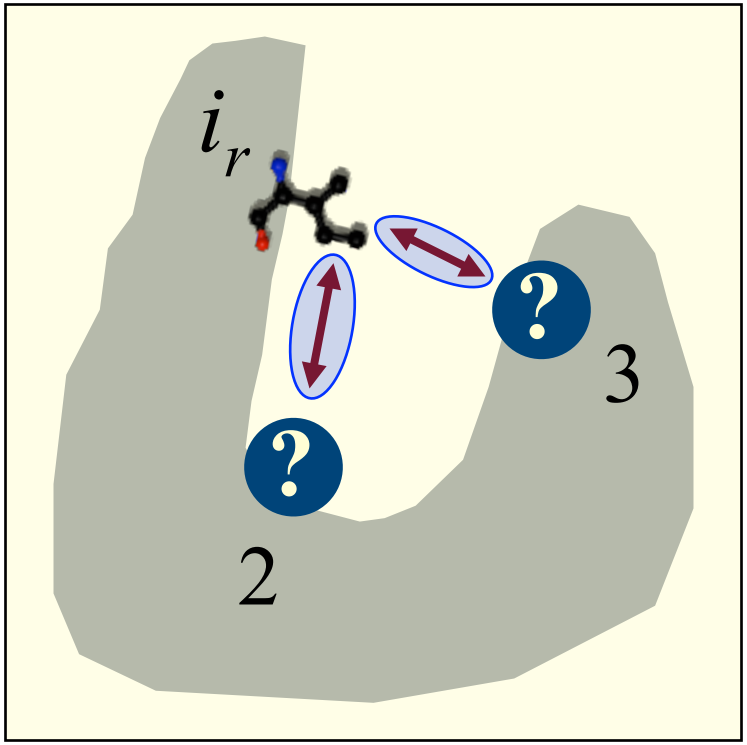
$$E_T = \sum_i \sum_j E(i_r, j_s); \qquad i < j$$

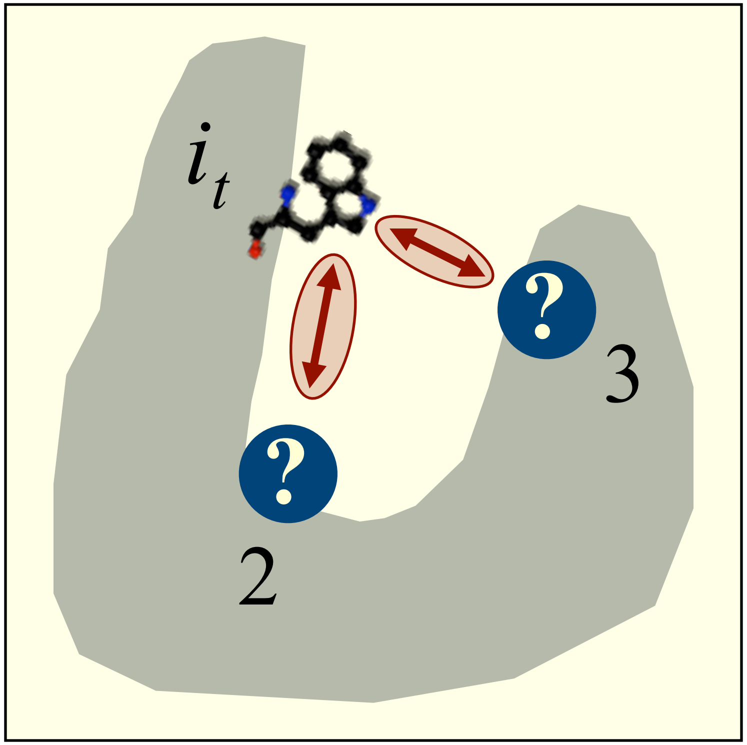




Dead End Elimination

Original DEE (Simplified)

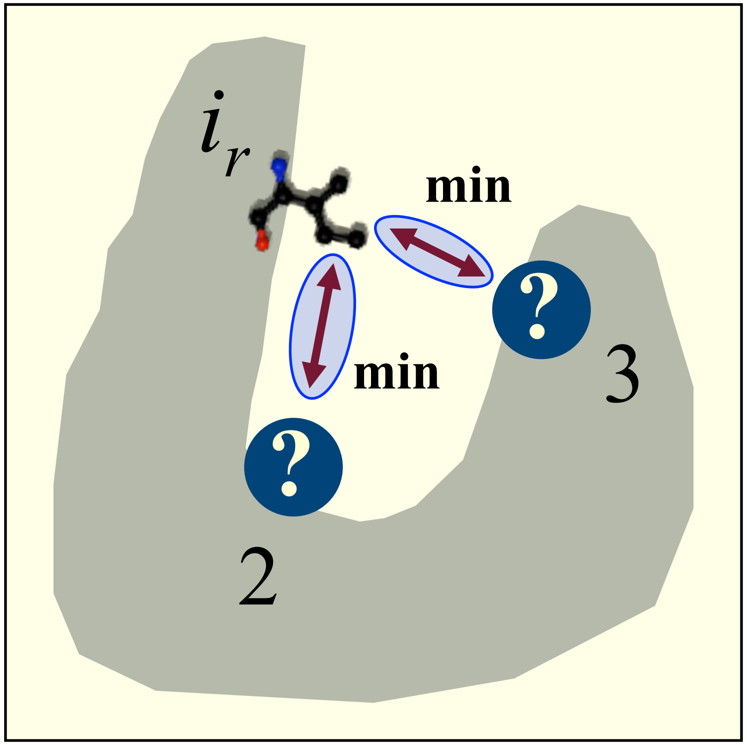


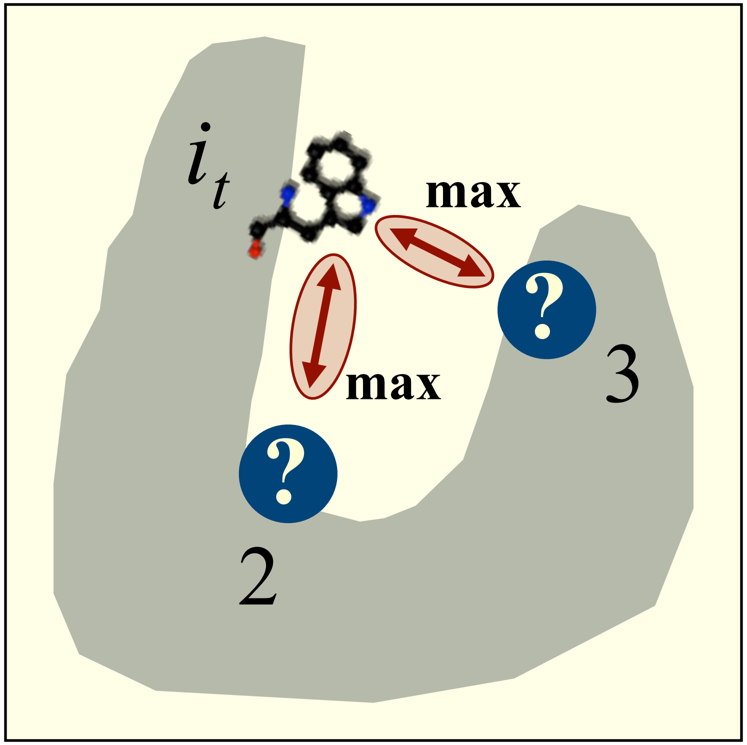


Dead End Elimination

Original DEE (Simplified)

$$\sum_{j \neq i}^{N} \min_{s} E(i_r, j_s) > \sum_{j \neq i}^{N} \max_{s} E(i_t, j_s)$$

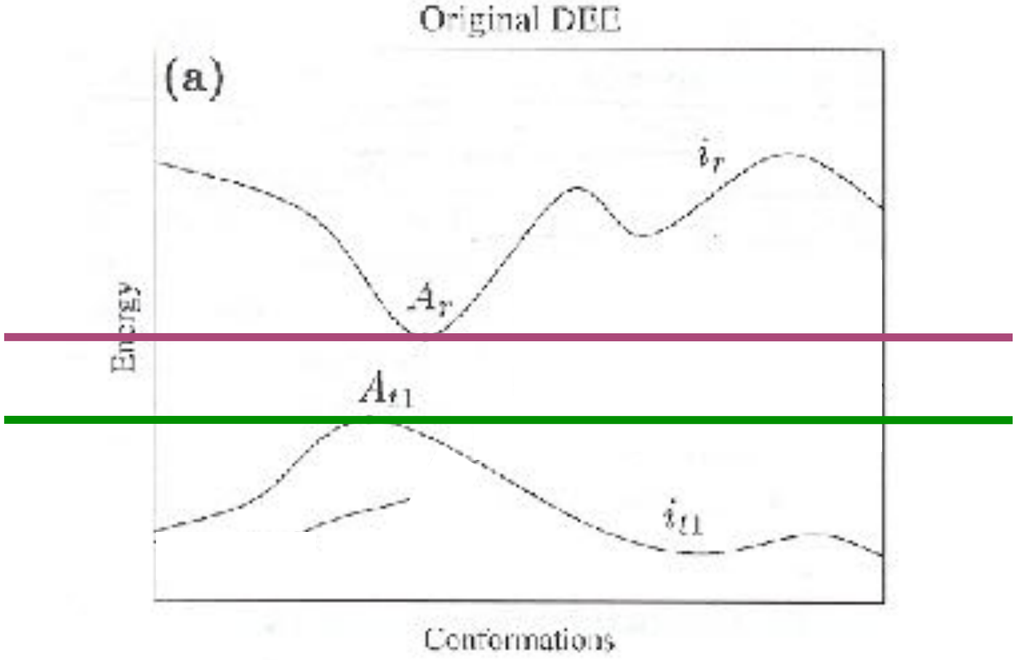




Dead End Elimination

Original DEE (Simplified)

$$\sum_{j \neq i}^{N} \min_{s} E(i_r, j_s) > \sum_{j \neq i}^{N} \max_{s} E(i_t, j_s)$$

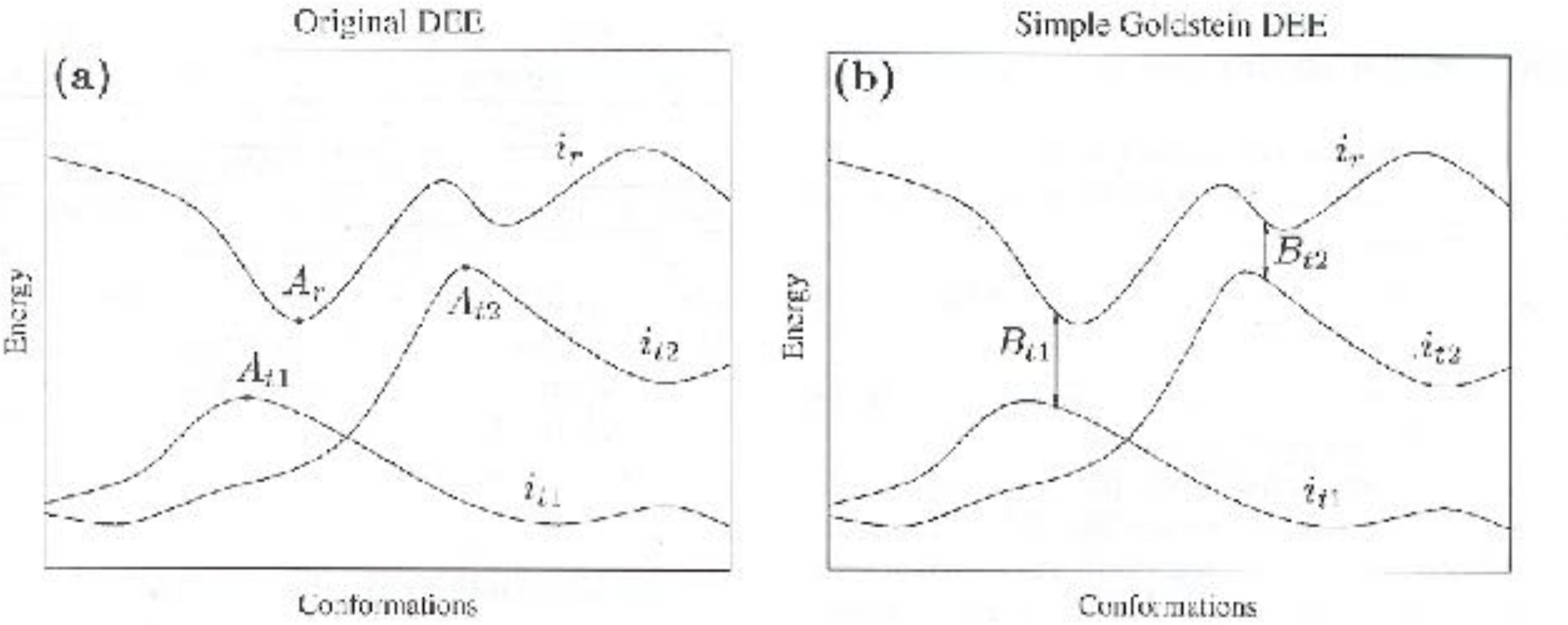


Pierce, Spriet, Desmet, Mayo, JCC, 2000

Dead End Elimination - Extensions

Original DEE (Simplified)

$$E(i_r) + \sum_{j \neq i}^{N} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max_s E(i_t, j_s)$$

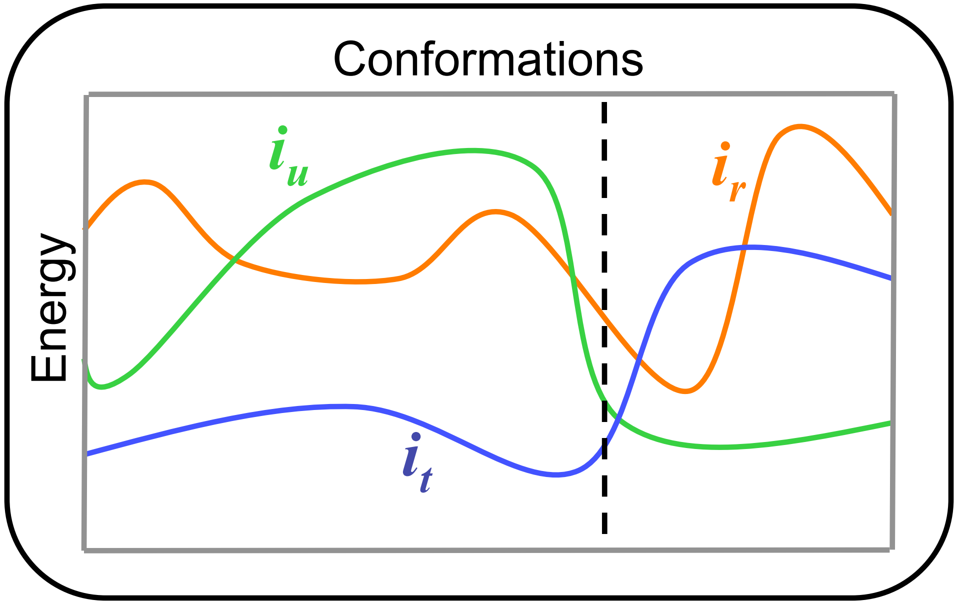


Pierce, Spriet, Desmet, Mayo, JCC, 2000

Dead End Elimination - Extensions

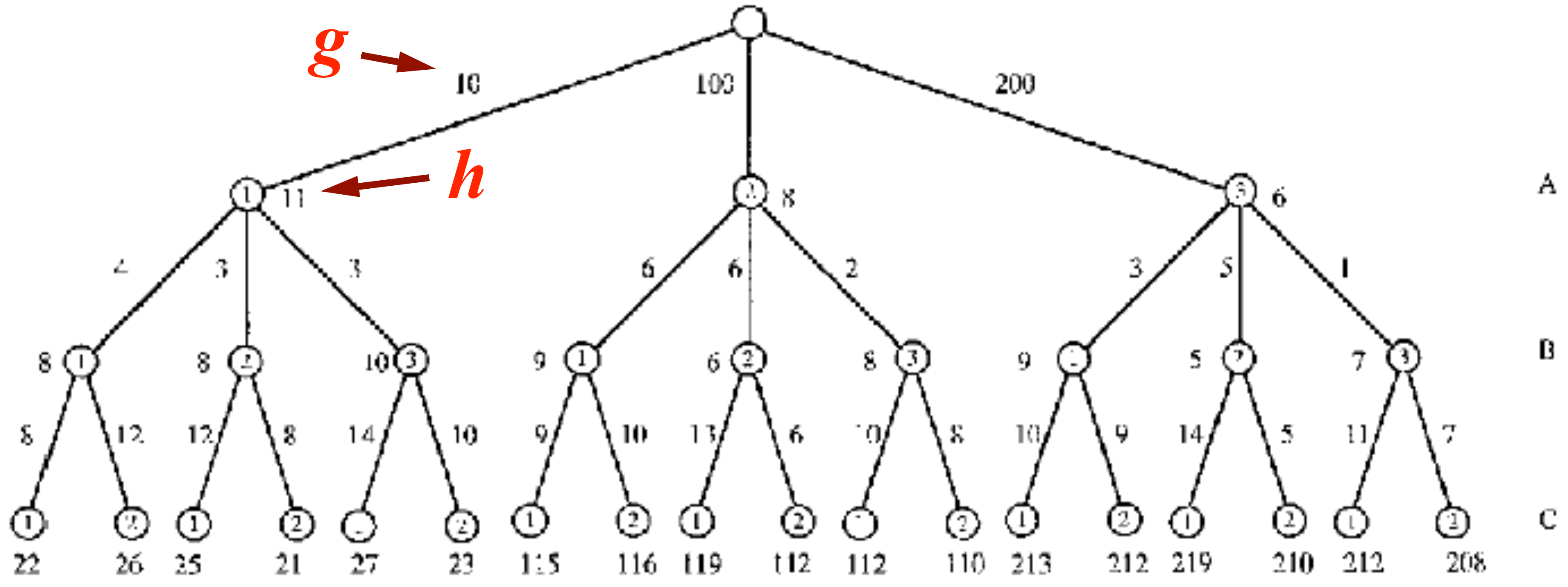
Original DEE (Simplified)

$$E(i_r) + \sum_{j \neq i}^{N} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max_s E(i_t, j_s)$$



i_r cannot be pruned by i_t or i_u *but it can be by i_t AND i_u*

A* Search - Conformation Tree



Leach, Lemon. Proteins 33(2):227-39 (1998)

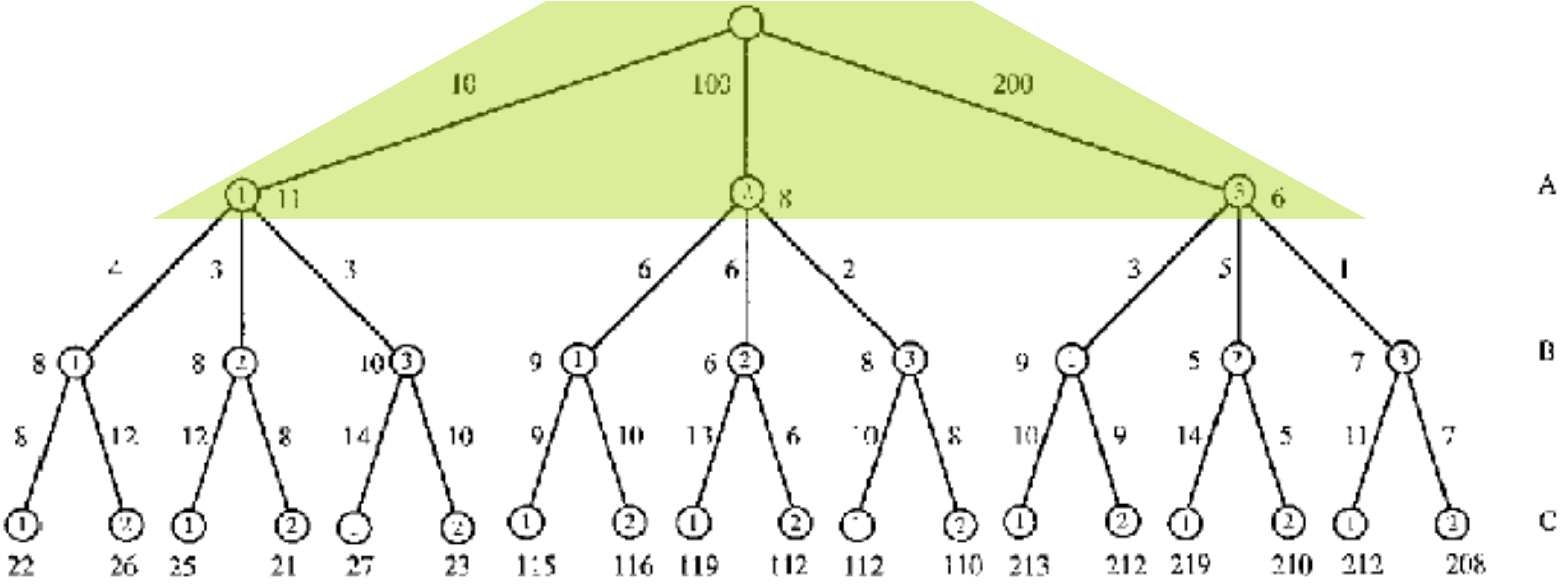
Let $f\,(x)$ be the score of node x

$$f\,(x) = g(x) + h(x)$$

$g(x)$ = cost of path from root to node x

$h(x)$ = lower bound on cost of path from x to leaf

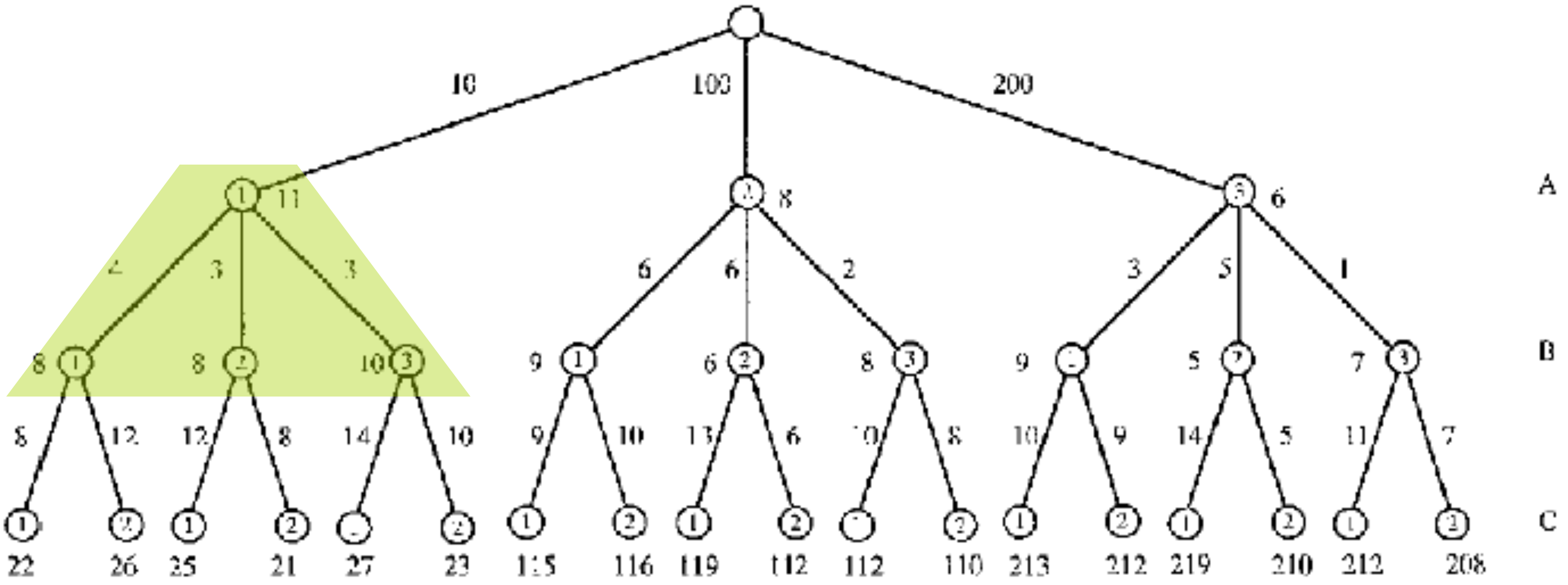
A* Search - Conformation Tree



Leach, Lemon. Proteins 33(2):227-39 (1998)

$A_1(21)\ A_2(108)\ A_3(206)$

A* Search - Conformation Tree

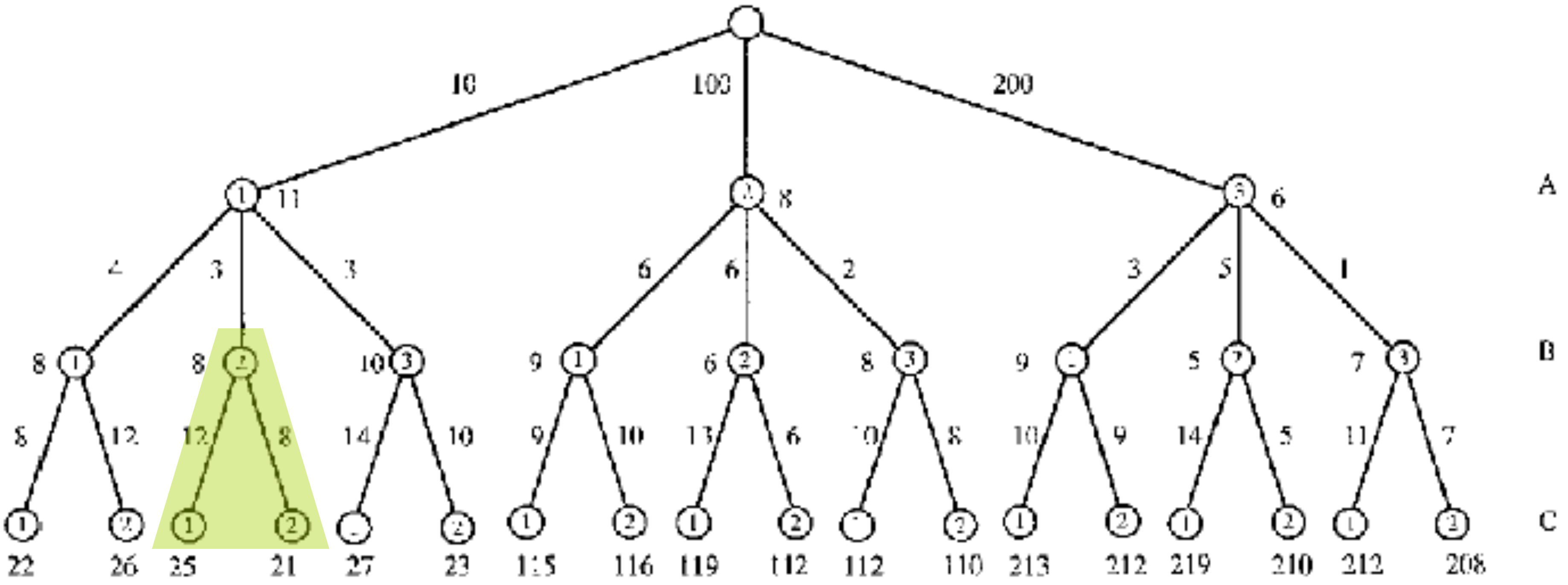


Leach, Lemon. Proteins 33(2):227-39 (1998)

A_1(21) A_2(108) A_3(206)

A_1B_2(21) A_1B_1(22) A_1B_3(22) A_2(108) A_3(206)

A* Search - Conformation Tree



Leach, Lemon. Proteins 33(2):227-39 (1998)

$A_1(21)$ $A_2(108)$ $A_3(206)$

$A_1B_2(21)$ $A_1B_1(22)$ $A_1B_3(22)$ $A_2(108)$ $A_3(206)$

$\mathbf{A_1B_2C_2}(21)$ $A_1B_1(22)$ $A_1B_3(23)$ $A_1B_2C_1(25)$ $A_2(108)$ $A_3(206)$