



COMPUTATIONAL CRYSTALLOGRAPHY INITIATIVE

Crystallographic Structure Refinement



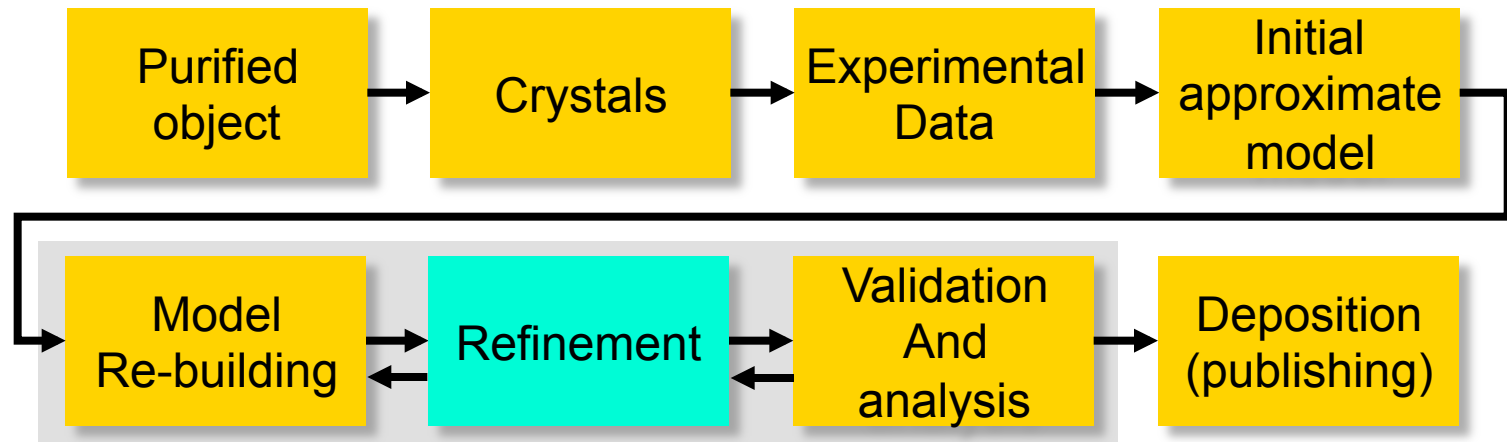
Pavel Afonine

Computation Crystallography Initiative
Physical Biosciences Division

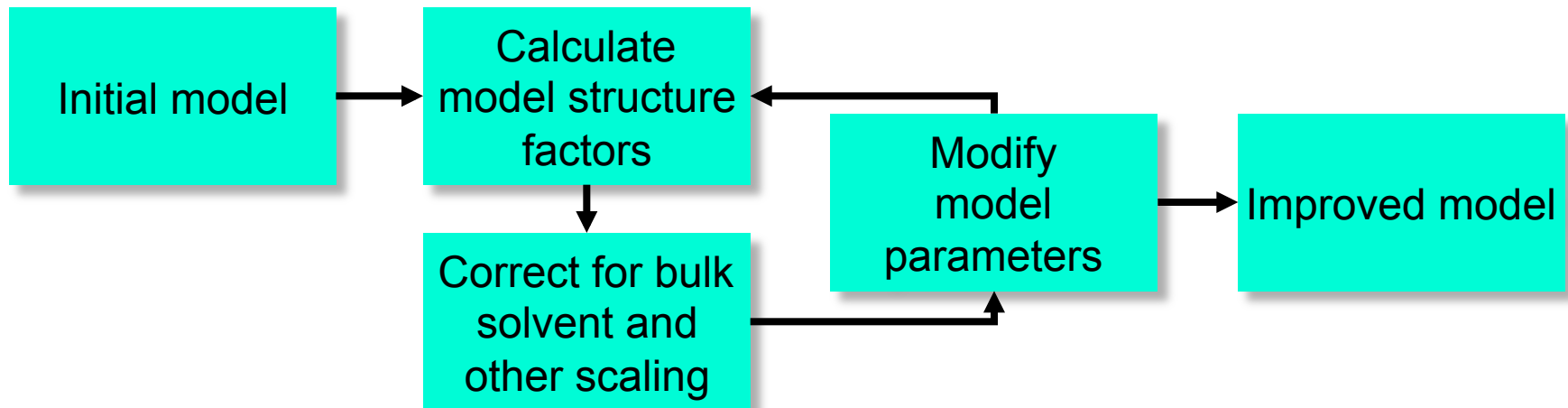
Lawrence Berkeley National Laboratory, Berkeley CA, USA

Structure refinement

Crystallographic structure determination workflow

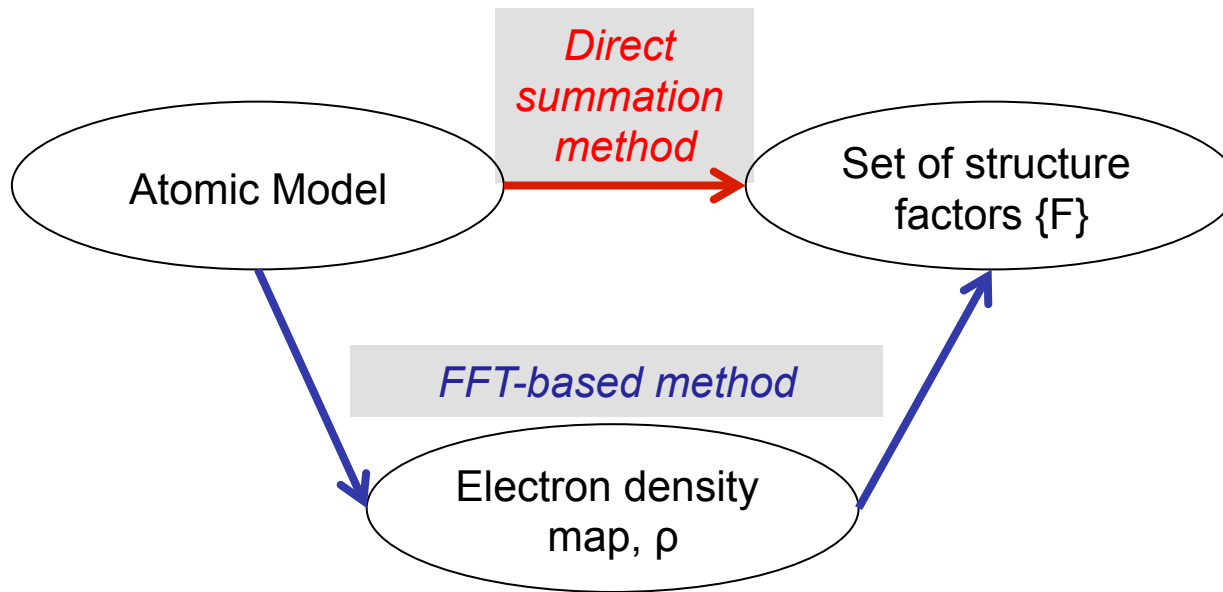


Structure refinement: modify model parameters to describe the experimental data as good as possible



Note about structure factors calculation

- Two ways of computing structure factor from atomic model



- For macromolecules the *FFT-based method* is much faster than the direct summation method
- Most of macromolecular refinement programs use *FFT-based method*
- FFT-based method* is based on a number of approximations and therefore it is less accurate than direct summation; however, inaccuracies introduced by these approximations are negligible in most of practical cases
- At ultra-high resolution (1 Å and higher) you may still want to use direct summation (if there is a reason to do so and it is not prohibitively expensive time-wise)

Note about structure factors calculation

- **Structure factor formula (direct summation method)**

$$\mathbf{F}(h,k,l) = \sum_{n=1}^{N_{atoms}} q_n f_n(s) \exp\left(-\frac{B_n s^2}{4}\right) \exp(2i\pi \mathbf{r}_n \cdot \mathbf{s})$$

$$f(s) = \sum_{k=1}^P a_k \exp\left(-\frac{b_k s^2}{4}\right) \quad \text{Gaussian approximation for atomic form-factor}$$

q_n , B_n and $\mathbf{r}_n = (x_n, y_n, z_n)$ – atomic occupancy, isotropic B-factor and coordinates

$P \sim 5$ (depends on approximation), a_k and b_k – parameters of approximation specific to atom type

$s^2 = \mathbf{h}^t \mathbf{G}^* \mathbf{h}$, \mathbf{h} – column-vector of Miller indices, \mathbf{G}^* - reciprocal-space metric tensor

- ✓ Calculation time \sim number of reflections * number of atoms
- ✓ Formula above yields exact values for F

Note about structure factors calculation

- **Structure factor formula (FFT-based summation)**

Fundamental formula
$$\mathbf{F}(h,k,l) = \int_{V_{cell}} \rho(\mathbf{r}) \exp\{2\pi i \mathbf{s} \cdot \mathbf{r}\} dV$$

Approximate way to compute this integral numerically:

$$\mathbf{F}(h,k,l) = \frac{V_{cell}}{N_X N_Y N_Z} \sum_{j_X}^{N_X-1} \sum_{j_Y}^{N_Y-1} \sum_{j_Z}^{N_Z-1} \rho(j_X, j_Y, j_Z) \exp\{2\pi i (h j_X + k j_Y + l j_Z)\}$$

which is discrete Fourier transform of electron density:

$$\rho(r) = \sum_{n=1}^{N_{atoms}} q_n \sum_{k=1}^P a_k \left(\frac{4\pi}{b_k + B_n} \right)^{3/2} \exp\left(-\frac{4\pi^2 |\mathbf{r} - \mathbf{r}_n|^2}{b_k + B_n} \right)$$

sampled at grid N_X, N_Y, N_Z in a sphere of radius R ($\sim 2 \text{ \AA}$) around each atom.

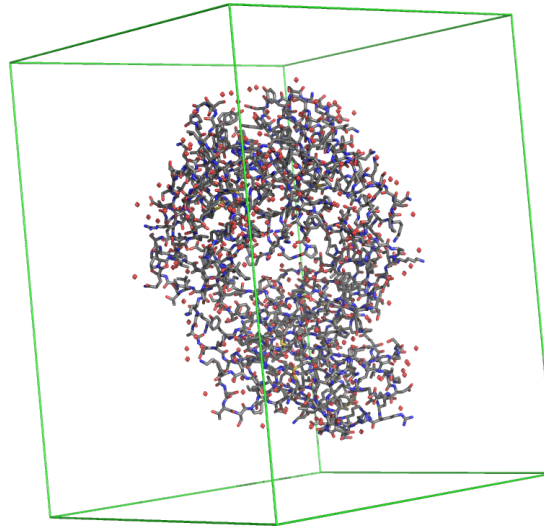
- ✓ Source of inaccuracy: replacement of continuous integral with discrete summation and truncation of atomic density within a sphere R .
- ✓ Calculation time \sim density calculation + FFT $\sim K_{grid} * (V_{atom}/V_{crystal}) + K_{grid} * \ln(K_{grid})$, where $K_{grid} = N_X N_Y N_Z$

Structure refinement

- 1. Model parameters**
- 2. Optimization goal**
- 3. Optimization method**

Model parameters or how we parameterize the crystal content

Crystal (unit cell)

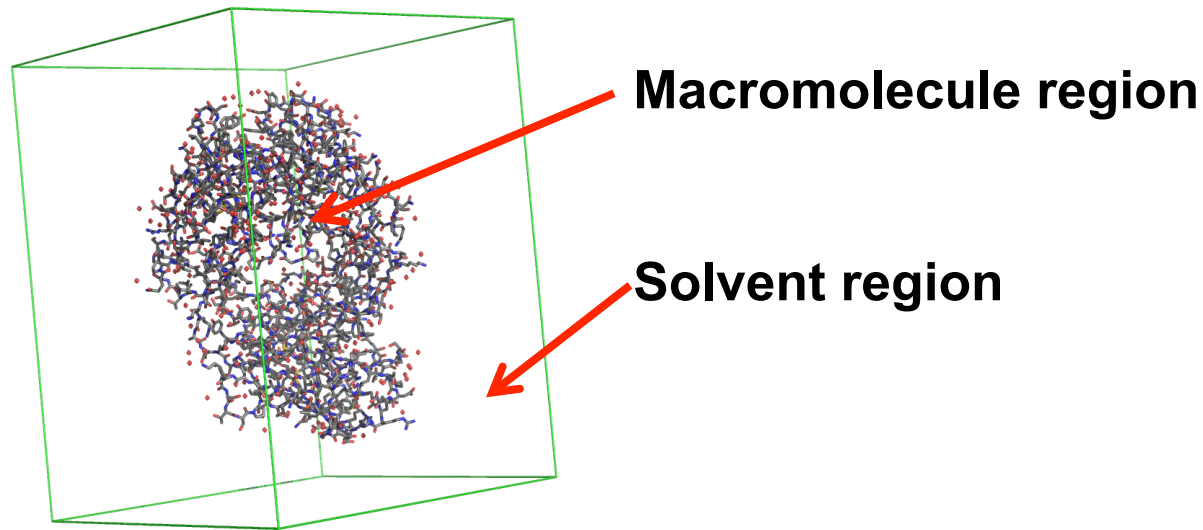


Non-atomic model parameters (Bulk solvent, anisotropy, twinning)

- Macromolecular crystals contain ~20-80% of solvent (mostly disordered)
- Crystal-specific: description of anisotropy or twinning

Atomic model parameters

Model parameters – Bulk solvent and anisotropy



Flat Bulk Solvent model (currently best available and most popular model):

- Electron density in solvent region is flat with some average value k_{SOL} ($\text{e}/\text{\AA}^3$)
- Solvent mask: a binary function: 0 in *Macromolecular* and 1 in *Solvent* region
- \mathbf{F}_{MASK} are structure factors calculated from Bulk solvent mask
- Contribution to the model structure factor:

$$\mathbf{F}_{\text{BULK}} = k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}}$$

- B_{SOL} is another bulk solvent parameter defining how deeply bulk solvent penetrates into a Macromolecular region

Non-atomic model (Bulk solvent and anisotropy)

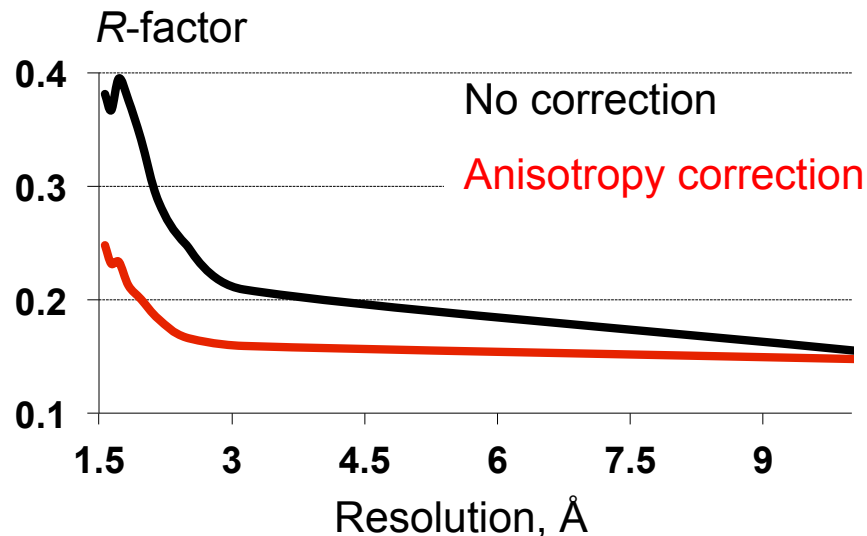
- Total model structure factor used in refinement, R -factor and map calculation:

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} e^{-sU_{\text{CRYSTAL}} s^t} \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

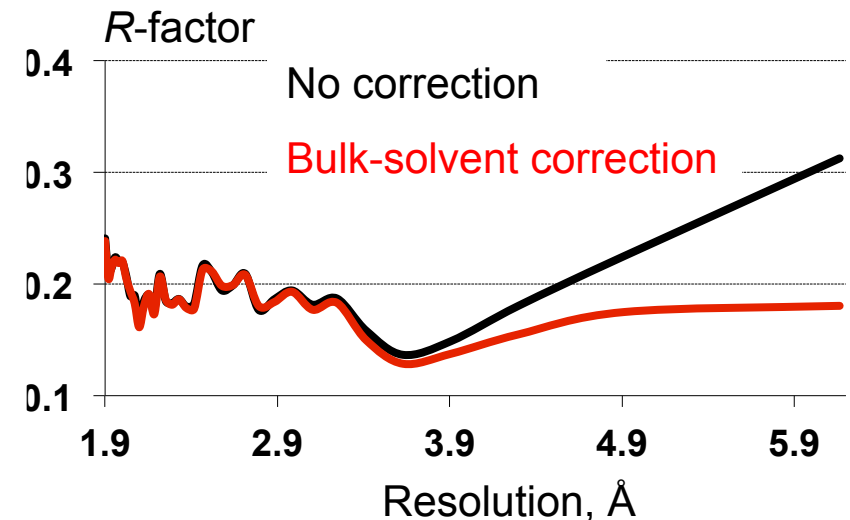
Anisotropy **Bulk-solvent contribution**

- Contribution to R -factor :

Effect of Anisotropic scaling (PDB: 2mhr)

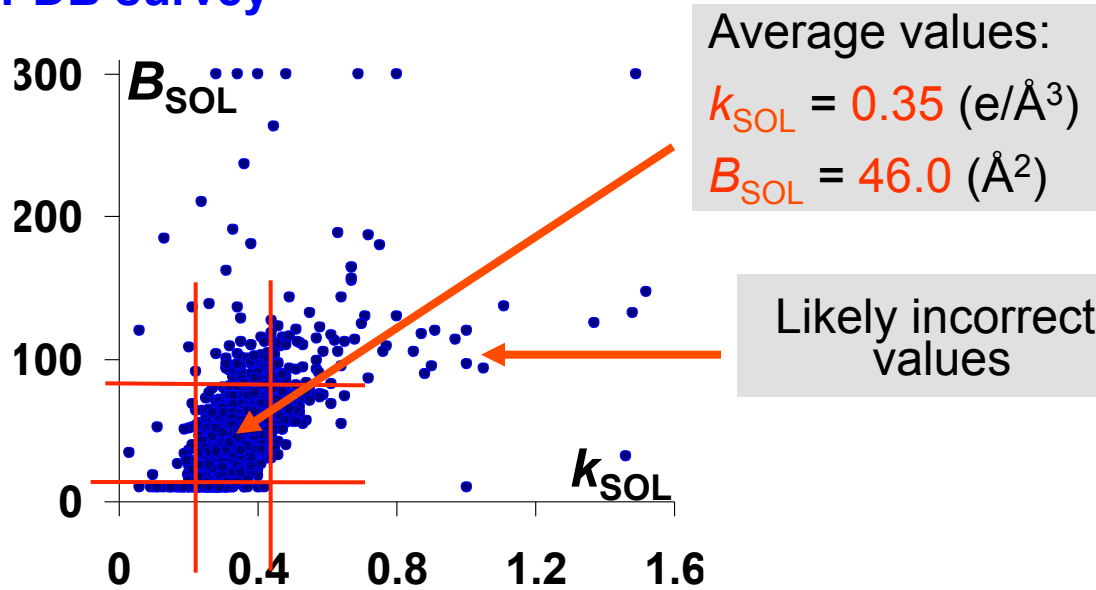


Effect of Bulk Solvent

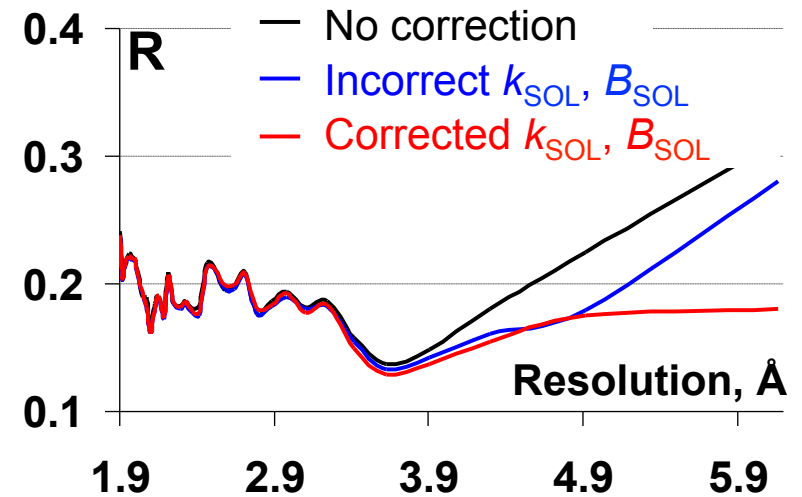


Bulk-solvent parameters: k_{SOL} and B_{SOL}

PDB survey



Bulk-solvent contributes to low resolution



Non-atomic model (Bulk solvent and anisotropy)

- Total model structure factor used in refinement, R -factor and map calculation:

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} e^{-\mathbf{s} \mathbf{U}_{\text{CRYSTAL}} \mathbf{s}^t} \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

Anisotropy
Bulk-solvent contribution

$\mathbf{U}_{\text{CRYSTAL}}$ is 3x3 symmetric anisotropy scale matrix with 6 refinable parameters:

$$\begin{pmatrix} U_{11} & U_{12} & U_{13} \\ & U_{22} & U_{23} \\ & & U_{33} \end{pmatrix}$$

- symmetry constraints apply

Crystal System	Restrictions on U
Triclinic 1-2	None
Monoclinic 3-15	$U_{13}=U_{23}=0$ when $\beta=\alpha=90^\circ$ $U_{12}=U_{23}=0$ when $\gamma=\alpha=90^\circ$ $U_{12}=U_{13}=0$ when $\gamma=\beta=90^\circ$
Orthorhombic 16-74	$U_{12}=U_{13}=U_{23}=0$
Tetragonal 75-142	$U_{11}=U_{22}$ and $U_{12}=U_{13}=U_{23}=0$
Rhombohedral (trigonal) 143-167	$U_{11}=U_{22}=U_{33}$ and $U_{12}=U_{13}=U_{23}$
Hexagonal 168-194	$U_{11}=U_{22}$ and $U_{13}=U_{23}=0$
Cubic 195-230	$U_{11}=U_{22}=U_{33}$ and $U_{12}=U_{13}=U_{23}=0$ (=isotropic)

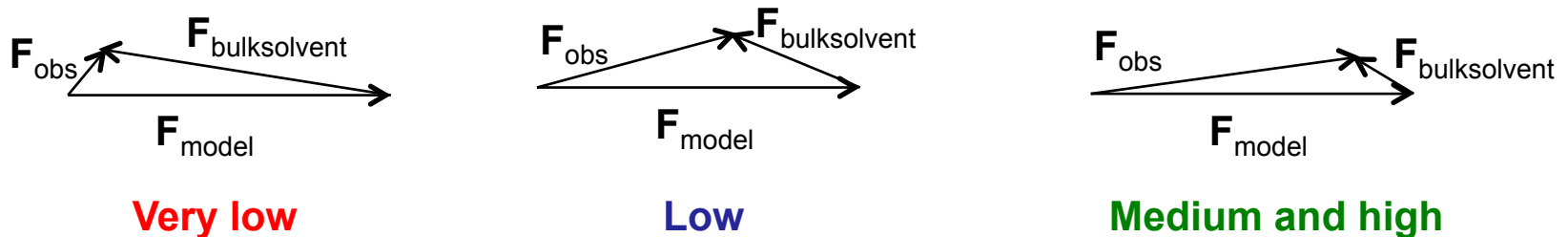
Other bulk-solvent model

▪ Bulk-solvent model based on Babinet principle:

- Assume $\rho_{\text{model}} = \rho_{\text{macromolecule}} + \rho_{\text{bulksolvent}}$
- $F_{\text{model}} = F_{\text{macromolecule}} + F_{\text{bulksolvent}}$
- Babinet principle (the Fourier transform of the solvent mask is related to the Fourier transform of the protein mask by a 180° phase shift):

$$F_{\text{macromolecule}} \approx -F_{\text{bulksolvent}}$$
- $F_{\text{bulksolvent}} = -k_{\text{sol}} \cdot \exp(-B_{\text{sol}} \cdot s^2) \cdot F_{\text{macromolecule}}$
- $F_{\text{model}} = F_{\text{macromolecule}} - k_{\text{sol}} \cdot \exp(-B_{\text{sol}} \cdot s^2) \cdot F_{\text{macromolecule}} = F_{\text{macromolecule}} \cdot (1 - k_{\text{sol}} \cdot \exp(-B_{\text{sol}} \cdot s^2))$

This is only correct at resolutions lower than 15-20Å, and brakes at higher resolutions (Podjarny, A. D. & Urzhumtsev, A.G. (1997). Methods Enzymol. 276, 641-658):



- ✓ Since a better model is available to account for bulk-solvent, the Babinet principle based model should not be used.

Other anisotropy correction model

- Polynomial model with 12 parameters as implemented in SHELXL (Usón et al., 1999; Parkin et al., 1995):

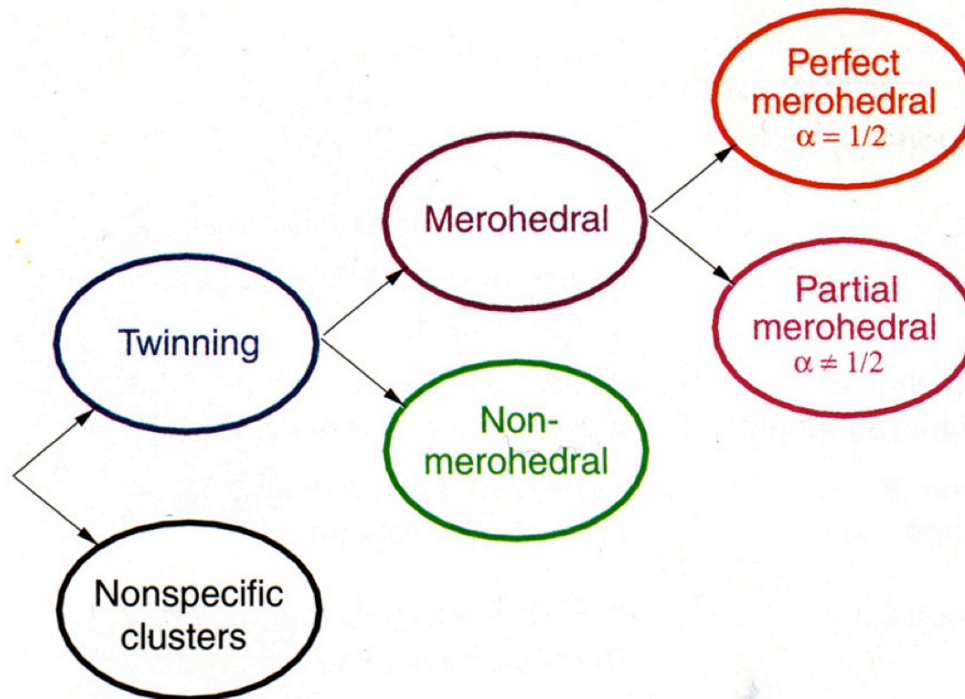
$$\begin{aligned} F_{\text{corr}}^2 = F_{\text{calc}}^2 [& h^2 a^{*2} (a_1 s + a_7) + k^2 b^{*2} (a_2 s + a_8) \\ & + l^2 c^{*2} (a_3 s + a_9) + 2klb^* c^* (a_4 s + a_{10}) \\ & + 2hla^* c^* (a_5 s + a_{11}) \\ & + 2hka^* b^* (a_6 s + a_{12})], \end{aligned}$$

where $s = \sin^{-2} \theta$.

θ – diffraction angle

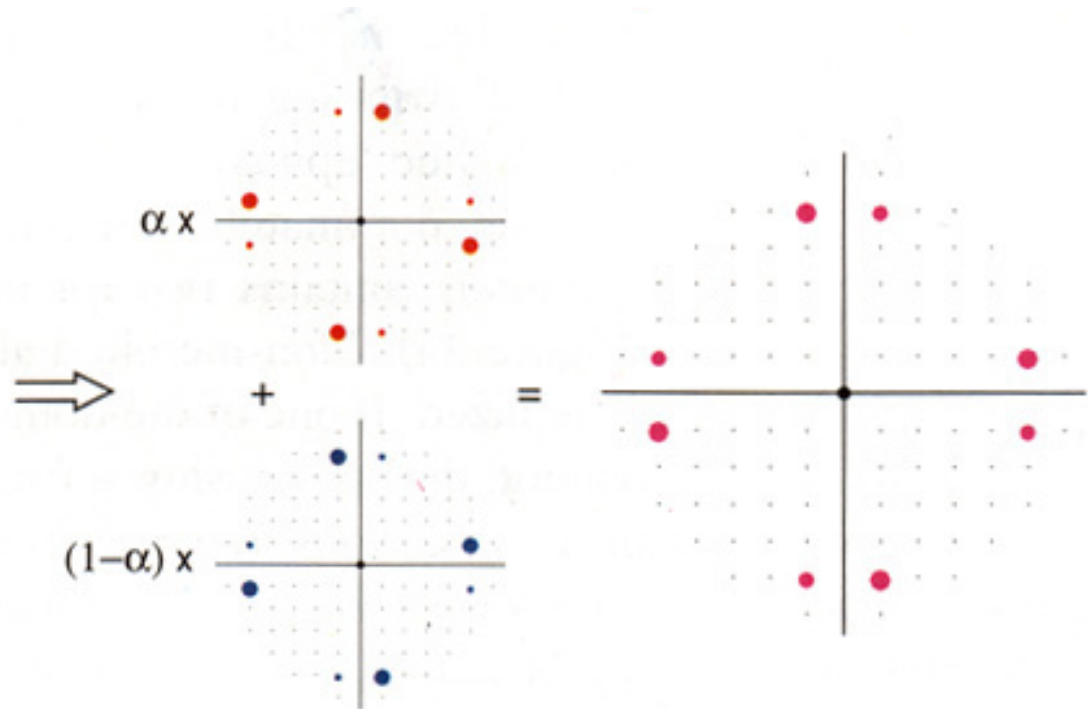
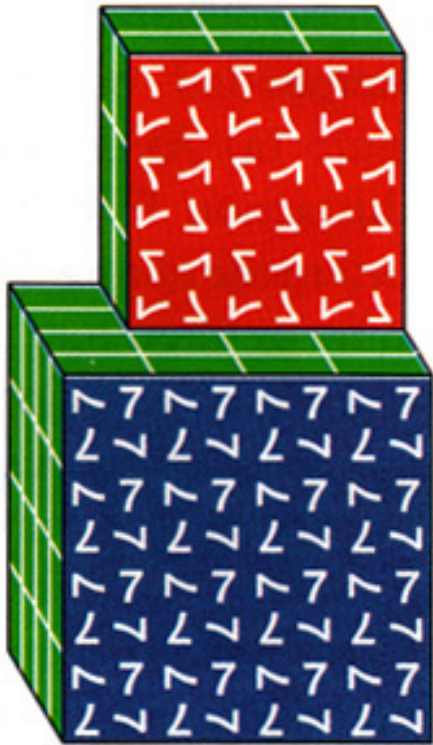
Non-atomic model parameters: Twinning

- Twinning is a kind of a crystal growth disorder.
- "Twins are regular aggregates consisting of crystals of the same species joined together in some definite mutual orientation" (Giacovazzo, 1992).
- A twinned crystal contains two or more identical single crystals (with identical packing) in different orientations. They are intergrown in such a way that at least some of their lattice directions are parallel.
- Only crystals that are intergrown in an ordered way are called twinned.



Non-atomic model parameters: Twinning

- Merohedral twinned crystals



- Hemihedral twinning:

- A special case of merohedral twinning: only two distinct orientations are assumed;
- Typically only merohedral twin form is reported for macromolecules

Non-atomic model parameters: Twinning

- Twinning parameterization:
 - *Twin law*: a description of the orientation of the different species relative to each other. This is an operator (matrix \mathbf{T}) that transforms the hkl indices of one species into the other.
 - *Twin fraction* (α): the fractional contribution of each component.
 - $\alpha=0$: no twinning; $\alpha<0.5$: partial hemihedral twinning; $\alpha=0.5$: perfect hemihedral twinning.
- In hemihedral case, the observed intensity is a weighted sum of the intensities of two reflections, \mathbf{h} and \mathbf{Th} (its twin mate):

$$I_{\text{OBS}}(\mathbf{h}) = (1 - \alpha)I(\mathbf{h}) + \alpha I(\mathbf{Th})$$

$$\mathbf{F}_{\text{M}}(\mathbf{h}) = e^{-s\mathbf{U}_{\text{CRYSTAL}}} s^t \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

$$F_{\text{MODEL}} = |\mathbf{F}_{\text{MODEL}}| = k_{\text{OVERALL}} \sqrt{\alpha |\mathbf{F}_{\text{M}}(\mathbf{h})|^2 + (1 - \alpha)^2 |\mathbf{F}_{\text{M}}(\mathbf{Th})|^2}$$

Atomic model parameters

Example of a PDB atom descriptors:

					<i>Position</i>			<i>Larger-scale disorder</i>				
ATOM	25	CA	PRO	A	4	31.309	29.489	26.044	1.00	57.79	C	
ANISOU	25	CA	PRO	A	4	8443	7405	6110	2093	-24	-80	C

Local mobility (small harmonic vibration)

Atomic model parameters

- **Position** (coordinates)
- **Local mobility** (ADP; Atomic Displacement Parameters or *B*-factors):

Diffraction data represents time- and space-averaged images of the crystal structure: time-averaged because atoms are in continuous thermal motions around mean positions, and space-averaged because there are often small differences between symmetry copies of the asymmetric unit in a crystal. ADP is to model the *small* dynamic displacements as isotropic or anisotropic *harmonic* displacements.

- **Larger-scale disorder** (occupancies)

Larger displacements (beyond harmonic approximation) can be modeled using occupancies (“alternative conformations/locations”).

Atomic model parameters

Atomic model parameterization is defined by:

- quality of experimental data (resolution, completeness, ...)
- quality of current model (initial with large errors, almost final, ...)
- data-to-parameters ratio (restraints have to be accounted for)

Can you see it in the map (= does the data amount and quality support the model parameterization)?

- $\sim 0.01\text{\AA}$ deviations from ideal bond lengths at resolutions $\sim 2\text{\AA}$?
- anisotropy of individual atoms at resolutions 2\AA and lower?

Model parameters

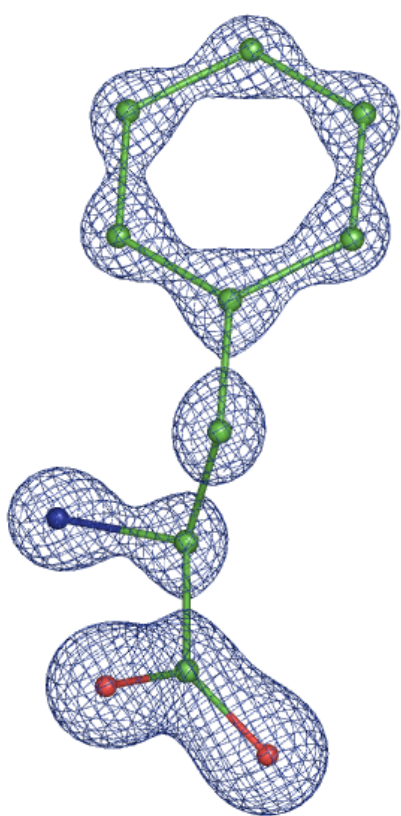
- Choice for model parameterization depends on amount of available data and its resolution

Key resolution limits and corresponding features

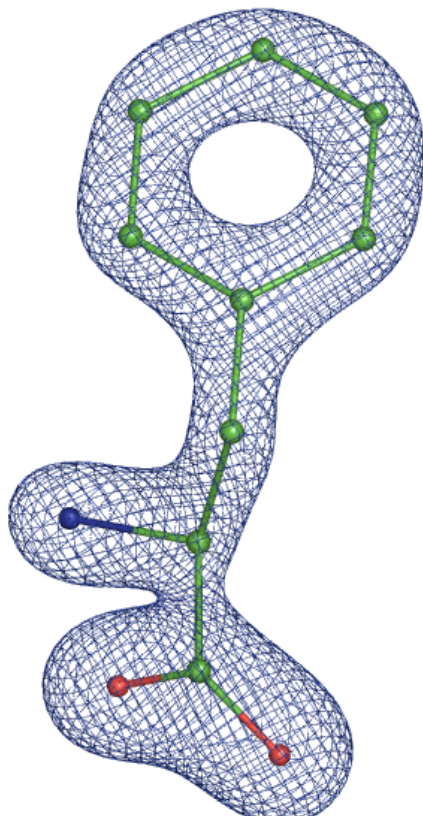
High res., Å	Size of details	Features of images	Mean Nref/at
~0.7	half of a X-X bond	deformation density	140
~0.9	X-H bond	some deformation density	75
~1.0		H atoms	60
~1.2	shortest covalent bond	individual atoms	40
~1.5	C _α -C		25
~2.0		most of ordered solvent	12
~2.5	distance N-C _β , C-C _β , N-O, C _α -O	clear side chains	7
~3.5	inter-C _α -distance	side chains may be guessed	3
~4.5	distance between chains	main chain	1.5
~6		α-helices	0.9
~12-15			0.1
~20	small domains	molecular envelopes	0.05

Data quality (resolution, completeness) defines how detailed the model is

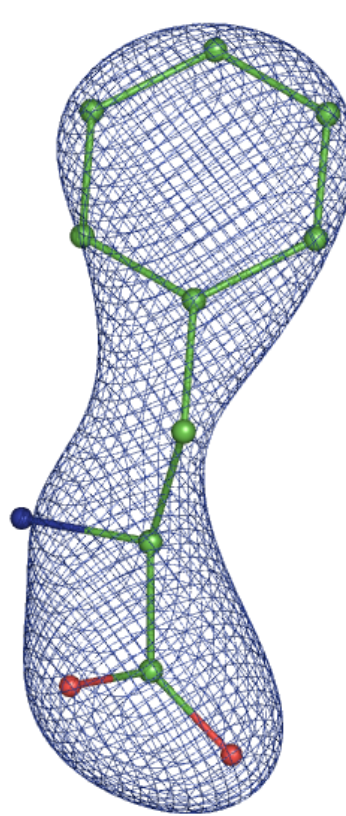
High Resolution Low



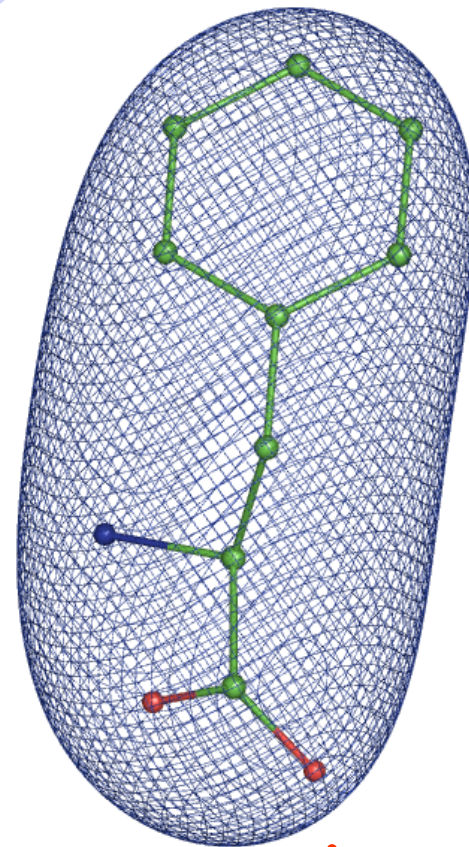
0.7Å



2Å



3Å

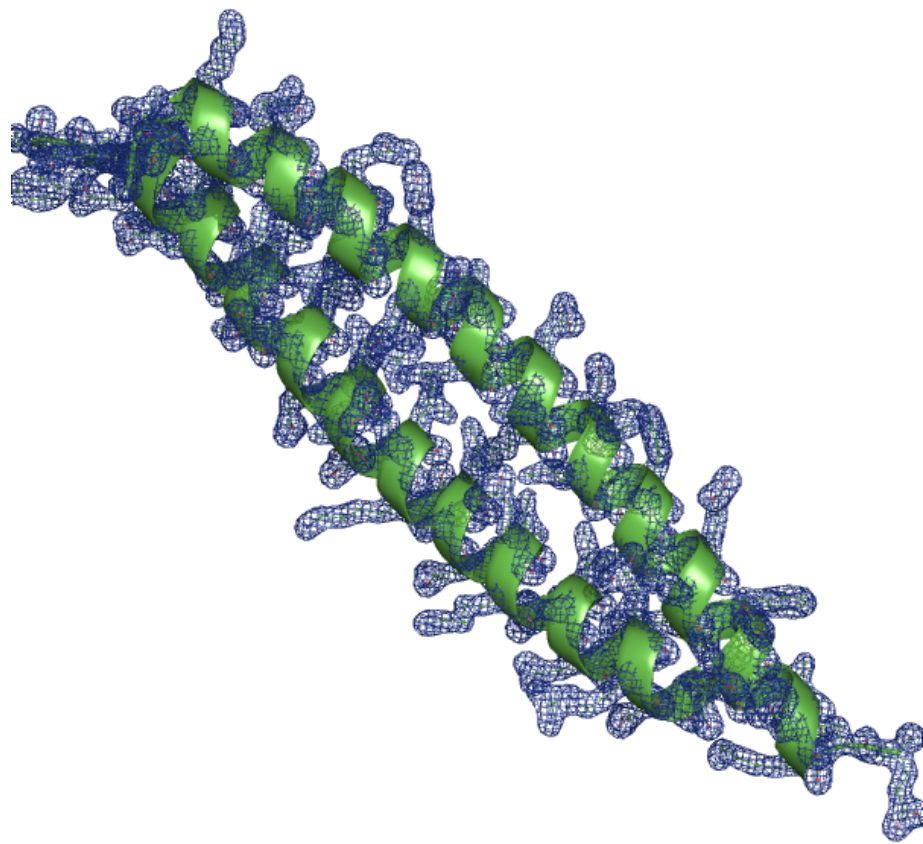


6.0Å

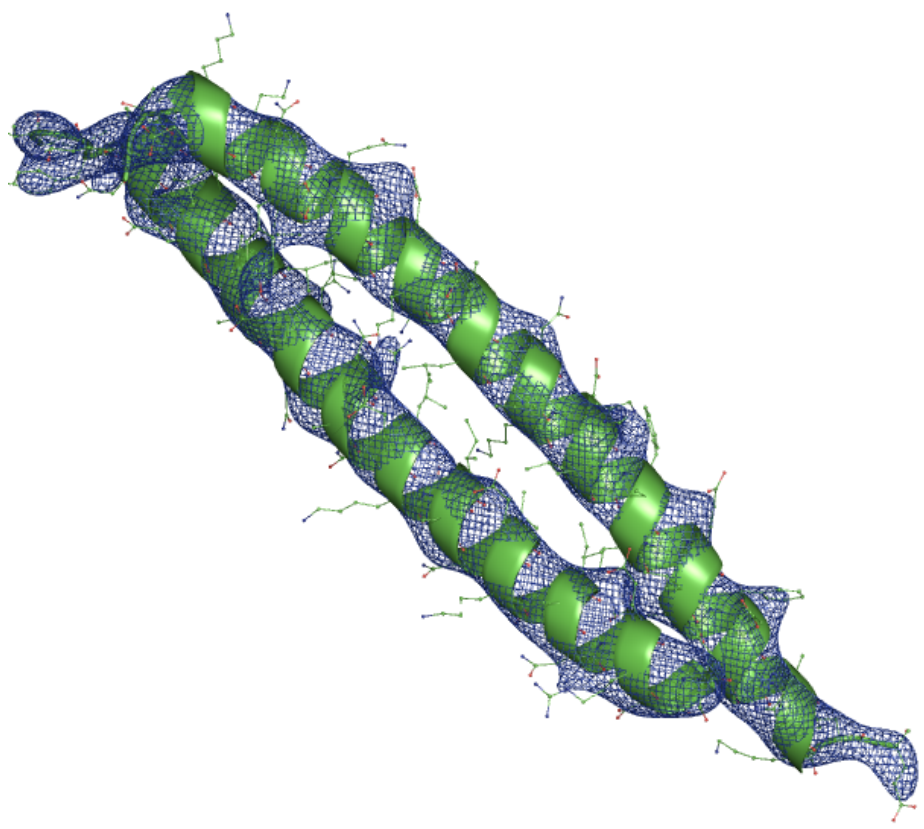
- **High resolution:** atomic models with or without restraints
- **Medium resolution:** atomic models with restraints or constraints
- **Low resolution:** atomic models with constraints or non-atomic models (cylinders for secondary-structure elements such as helices)

Data quality (resolution, completeness) defines how detailed the model is

High Resolution Low



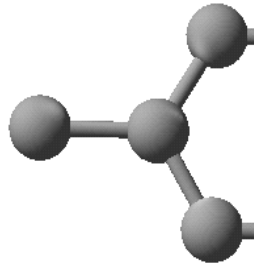
2Å



6.0Å

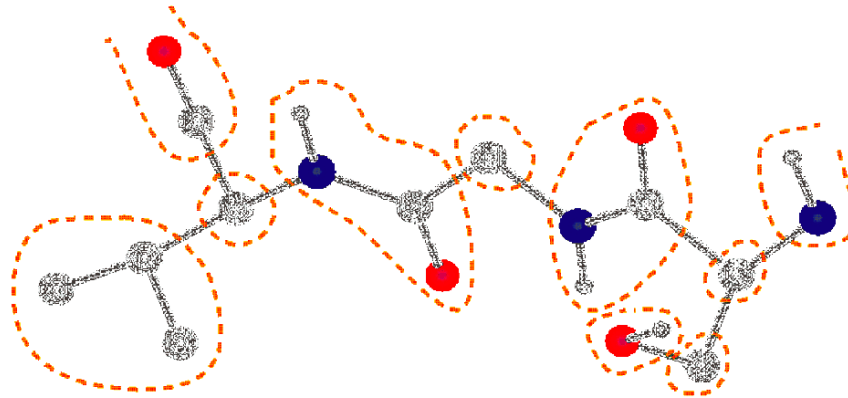
Model parameterization: coordinates

Individual atoms



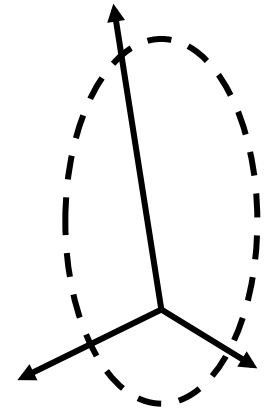
$3 * N_{atoms}$

Constrained rigid bodies (torsion angle parameterization)



$3 * N_{atoms} / (7 \dots 10)$

Rigid body



$6 * N_{groups}$

High

Resolution

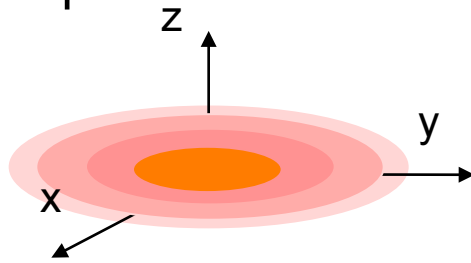
Low

Some *a priori* information may be needed:

- Stereochemistry restraints
- NCS restraints or constraints

Atomic Displacement Parameters (ADP or “B-factors”)

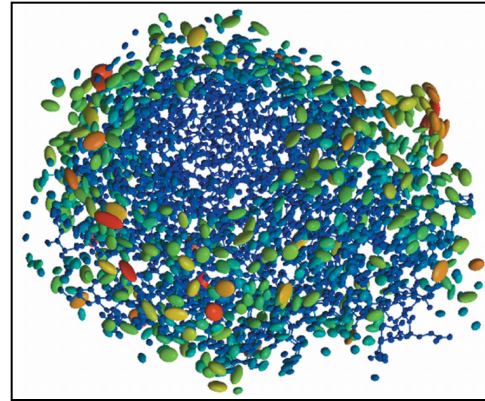
- Atomic displacements are anisotropic



$$\rho(\Delta\mathbf{r}) \sim \exp\{-\Delta\mathbf{r} \cdot \mathbf{U}^{-1} \Delta\mathbf{r}\}$$

$$\mathbf{U} = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix}$$

- Hierarchy of atomic displacements

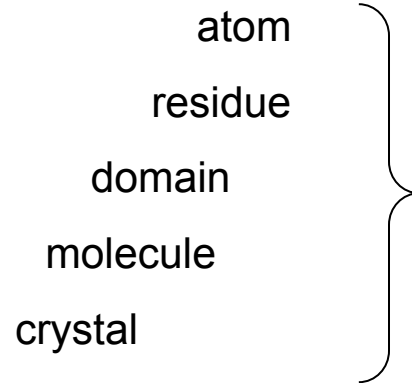
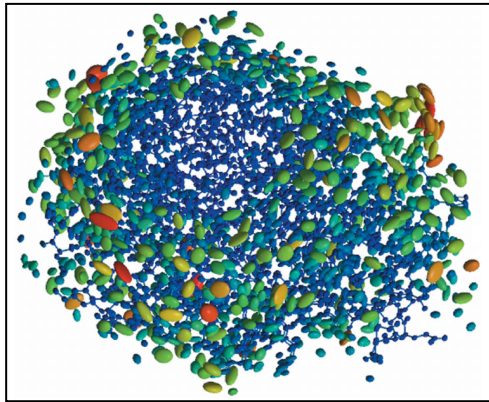


atom
residue
domain
molecule
crystal

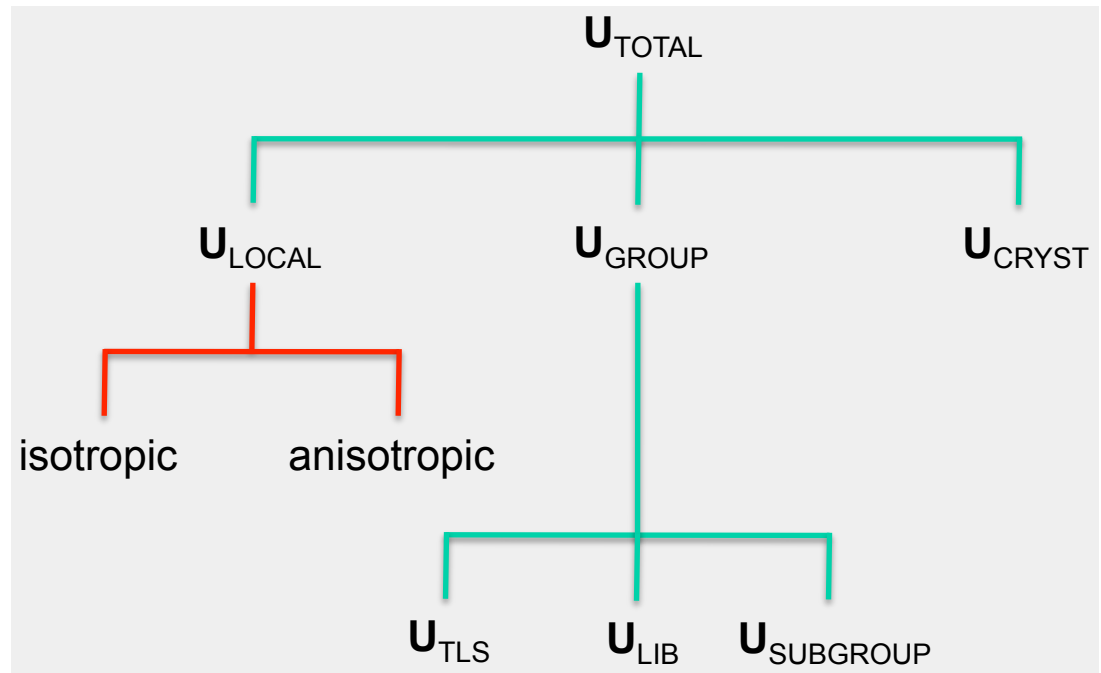


Atomic Displacement Parameters (ADP or “B-factors”)

- Hierarchy of atomic displacements

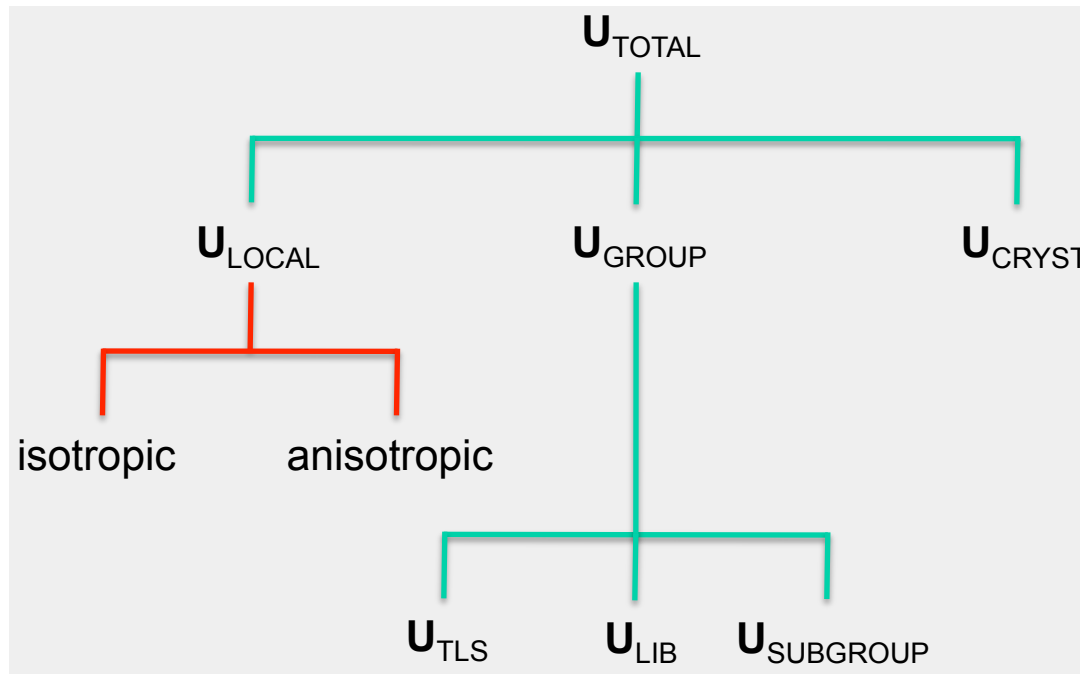


Total ADP: $\mathbf{U}_{\text{TOTAL}} = \mathbf{U}_{\text{CRYST}} + \mathbf{U}_{\text{GROUP}} + \mathbf{U}_{\text{LOCAL}}$



Atomic Displacement Parameters (ADP or “B-factors”)

- Total ADP $\mathbf{U}_{\text{TOTAL}} = \mathbf{U}_{\text{CRYST}} + \mathbf{U}_{\text{GROUP}} + \mathbf{U}_{\text{LOCAL}}$

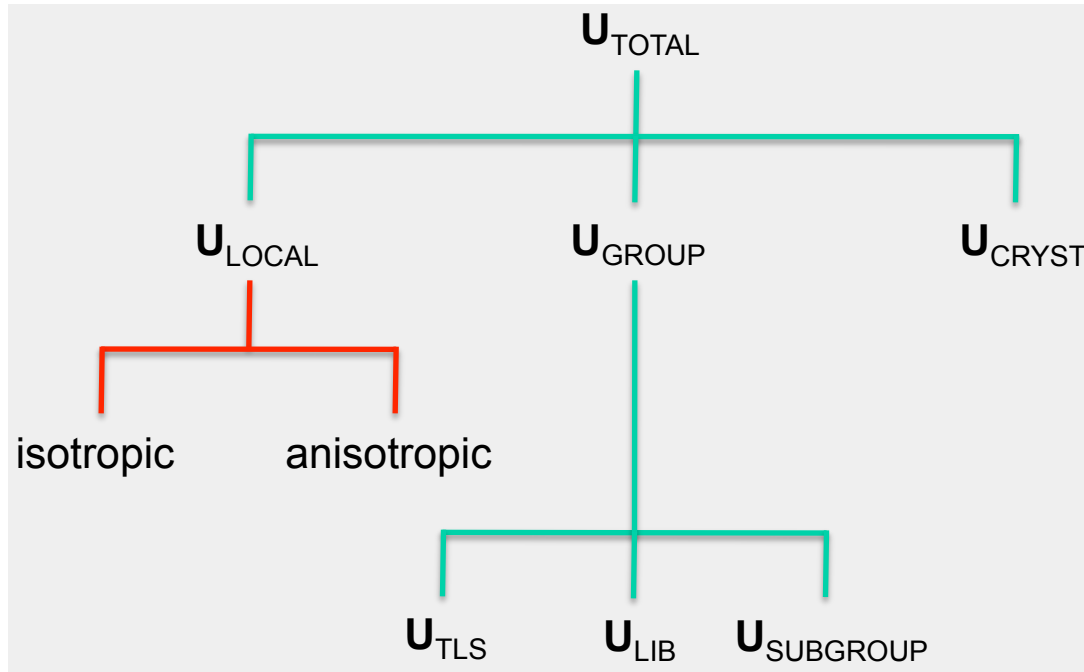


- $\mathbf{U}_{\text{CRYST}}$ – lattice vibrations; accounted for by overall anisotropic scale (6 parameters).

$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} e^{-s\mathbf{U}_{\text{CRYSTAL}} s^t} \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

Atomic Displacement Parameters: TLS

▪ Total ADP $\mathbf{U}_{\text{TOTAL}} = \mathbf{U}_{\text{CRYST}} + \mathbf{U}_{\text{GROUP}} + \mathbf{U}_{\text{LOCAL}}$



\mathbf{U}_{TLS} – rigid body collective displacements of whole molecules, domains, secondary structure elements.

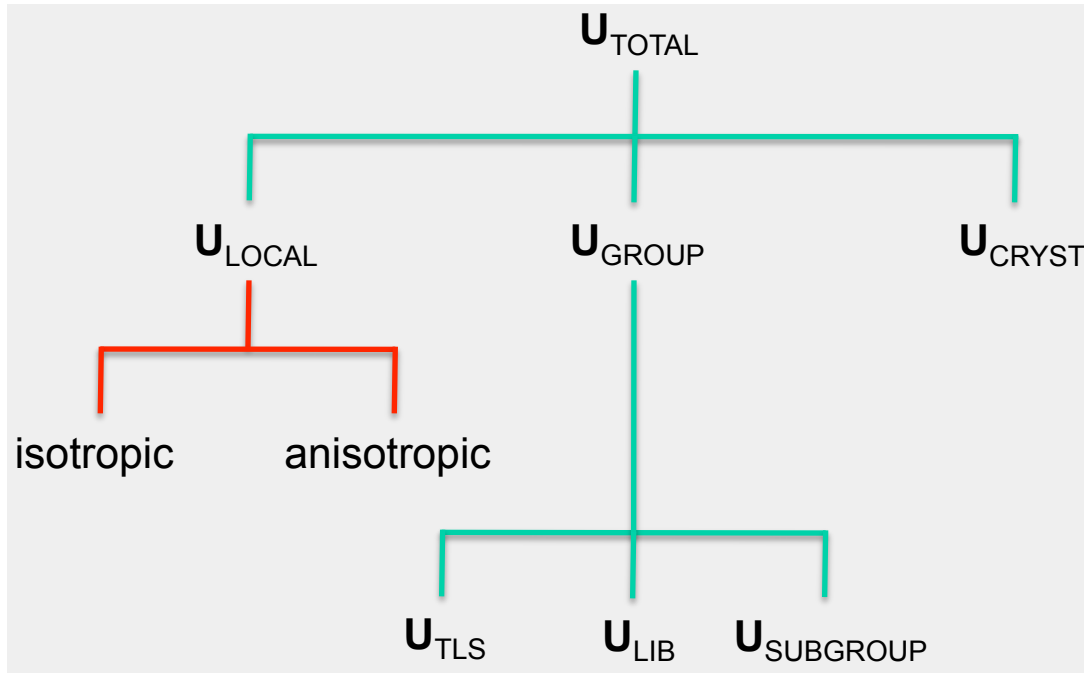
$$\mathbf{U}_{\text{TLS}} = \mathbf{T} + \mathbf{A}\mathbf{L}\mathbf{A}^t + \mathbf{A}\mathbf{S} + \mathbf{S}^t\mathbf{A}^t$$

(20 TLS parameters per group);
 \mathbf{T} , \mathbf{L} and \mathbf{S} are 3x3 tensors. \mathbf{T} and \mathbf{L} are symmetric, \mathbf{S} is not.

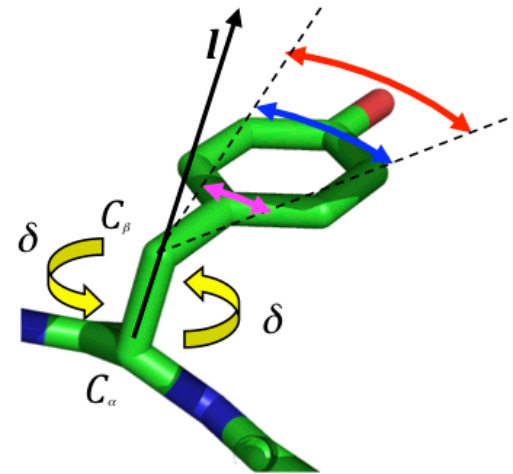
- \mathbf{T} describes anisotropic translational displacement (units: \AA^2).
- \mathbf{L} describes rotational displacement (libration) of the rigid group (units: rad^2).
- \mathbf{S} describes the correlation between the rotation and translation of a rigid body that undergoes rotation about three orthogonal axes that do not intersect at a common point.
- \mathbf{A} is anti-symmetric tensor; a function of atomic coordinates and TLS origin.

Atomic Displacement Parameters: U_{LIB}

- Total ADP $U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$



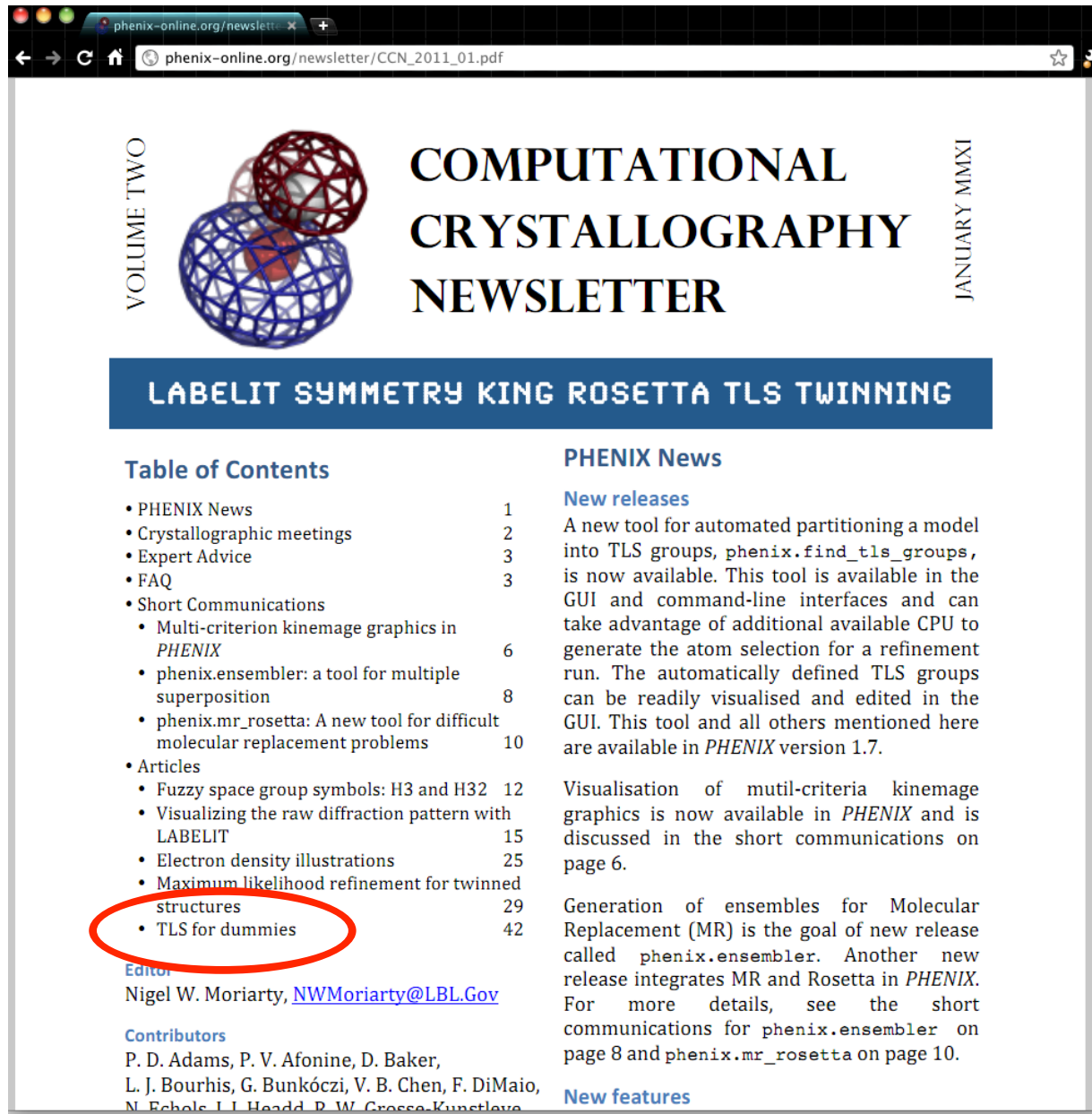
U_{LIB} – librational motion of side chain around bond vector.



- U_{LIB} is simplified TLS model with one refinable parameter per libration axis:

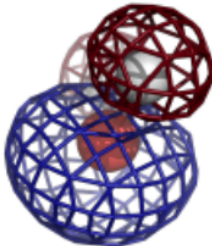
$$U_{LIB} = \delta \mathbf{A} \mathbf{L} \mathbf{A}^t$$

where \mathbf{A} and \mathbf{L} are completely determined by the coordinates of involved atoms



phenix-online.org/newslet...
phenix-online.org/newsletter/CCN_2011_01.pdf

VOLUME TWO



COMPUTATIONAL
CRYSTALLOGRAPHY
NEWSLETTER

JANUARY MMXI

LABELIT SYMMETRY KING ROSETTA TLS TWINNING

Table of Contents

- PHENIX News 1
- Crystallographic meetings 2
- Expert Advice 3
- FAQ 3
- Short Communications
 - Multi-criterion kinemage graphics in *PHENIX* 6
 - phenix.ensembl: a tool for multiple superposition 8
 - phenix.mr_rosetta: A new tool for difficult molecular replacement problems 10
- Articles
 - Fuzzy space group symbols: H3 and H32 12
 - Visualizing the raw diffraction pattern with LABELIT 15
 - Electron density illustrations 25
 - Maximum likelihood refinement for twinned structures 29
 - TLS for dummies 42

Editor
Nigel W. Moriarty, NWMoriarty@LBL.Gov

Contributors
P. D. Adams, P. V. Afonine, D. Baker,
L. J. Bourhis, G. Bunkóczi, V. B. Chen, F. DiMaio,
N. Echols, J. L. Headd, P. W. Grosse-Kunstleve

PHENIX News

New releases

A new tool for automated partitioning a model into TLS groups, `phenix.find_tls_groups`, is now available. This tool is available in the GUI and command-line interfaces and can take advantage of additional available CPU to generate the atom selection for a refinement run. The automatically defined TLS groups can be readily visualised and edited in the GUI. This tool and all others mentioned here are available in *PHENIX* version 1.7.

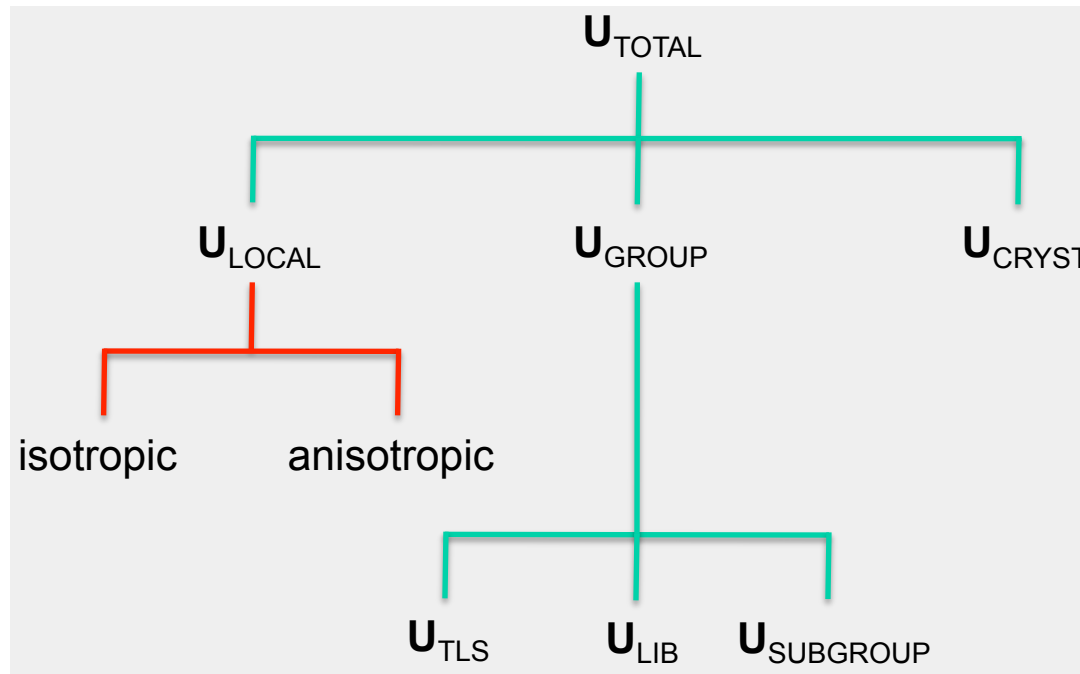
Visualisation of mutil-criteria kinemage graphics is now available in *PHENIX* and is discussed in the short communications on page 6.

Generation of ensembles for Molecular Replacement (MR) is the goal of new release called `phenix.ensembl`. Another new release integrates MR and Rosetta in *PHENIX*. For more details, see the short communications for `phenix.ensembl` on page 8 and `phenix.mr_rosetta` on page 10.

New features

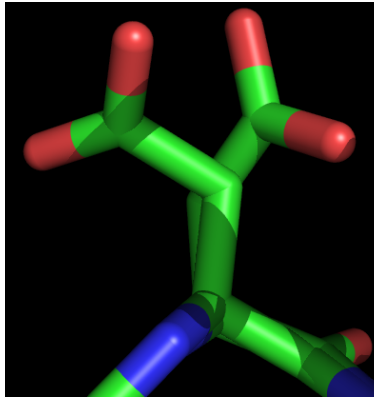
Atomic Displacement Parameters (ADP or “B-factors”)

▪ Total ADP $U_{\text{TOTAL}} = U_{\text{CRYST}} + U_{\text{GROUP}} + U_{\text{LOCAL}}$



- U_{LOCAL} – local vibration of individual atoms.
- Depending on data amount and quality, it can be less precise (isotropic) or more precise (anisotropic).
 - These vibrations are expected to be very small due to assumption of rigidity of interatomic bonds (vibrating atoms cannot stretch the bond much).

Occupancy: large-scale disorder that cannot be modeled with harmonic model (ADP)



- Occupancy is the fraction of molecules in the crystal in which a given atom occupies the position specified in the model.
- If all molecules in the crystal are identical, then occupancies for all atoms are 1.00.

- We may refine occupancy because sometimes a region of the molecules may have several distinct conformations.
- Refining occupancies provides estimates of the frequency of alternative conformations.

ATOM	1	N	AARG	A	192	-5.782	17.932	11.414	0.72	8.38	N
ATOM	2	CA	AARG	A	192	-6.979	17.425	10.929	0.72	10.12	C
ATOM	3	C	AARG	A	192	-6.762	16.088	10.271	0.72	7.90	C
ATOM	7	N	BARG	A	192	-11.719	17.007	9.061	0.28	9.89	N
ATOM	8	CA	BARG	A	192	-10.495	17.679	9.569	0.28	11.66	C
ATOM	9	C	BARG	A	192	-9.259	17.590	8.718	0.28	12.76	C

Structure refinement

1. **Model parameters**
2. **Optimization goal**
3. **Optimization method**

Refinement target function

- **Structure refinement** is a process of changing a model parameters in order to optimize a goal (target) function:

$$T = F(\text{Experimental data}, \text{Model parameters}, \text{A priori knowledge})$$

- **Experimental data** – a set of diffraction amplitudes F_{obs} (and phases, if available).
 - **Model parameters**: coordinates, ADP, occupancies, bulk-solvent, ...
 - **A priori knowledge (restraints or constraints)** – additional information that may be introduced to compensate for the insufficiency of experimental data (finite resolution, poor data-to-parameters ratio)
- Typically: $T = T_{\text{DATA}} + w * T_{\text{RESTRAINTS}}$
 - E_{DATA} relates model to experimental data
 - $E_{\text{RESTRAINTS}}$ represents *a priori* knowledge
 - w is a weight to balance the relative contribution of E_{DATA} and $E_{\text{RESTRAINTS}}$
 - A priori knowledge can be imposed in the form of constraints so

$$T = E_{\text{DATA}}$$

Target function

A function that relates model parameters to experimental data. Typically looks like this:

$$T = T_{\text{DATA}}(F_{\text{OBS}}, F_{\text{MODEL}}) + wT_{\text{RESTRAINTS}}$$

▪ Least-Squares (reciprocal space)

$$T_{\text{DATA}} = \sum_s \mathbf{w}_s (F_s^{\text{OBS}} - kF_s^{\text{MODEL}})^2$$

- Widely used in small molecule crystallography
- Used in macromolecular crystallography in the past

▪ Maximum-Likelihood (reciprocal space; much better option for macromolecules)

$$T_{\text{DATA}} = \sum_s (1 - K_s^{cs}) \left(-\frac{\alpha_s^2 (F_s^{\text{MODEL}})^2}{\varepsilon_s \beta_s} + \ln \left(I_0 \left(\frac{2\alpha_s F_s^{\text{MODEL}} F_s^{\text{OBS}}}{\varepsilon_s \beta_s} \right) \right) \right) +$$
$$+ K_s^{cs} \left(-\frac{\alpha_s^2 (F_s^{\text{MODEL}})^2}{2\varepsilon_s \beta_s} + \ln \left(\cosh \left(\frac{\alpha_s F_s^{\text{MODEL}} F_s^{\text{OBS}}}{\varepsilon_s \beta_s} \right) \right) \right)$$

▪ Real space target

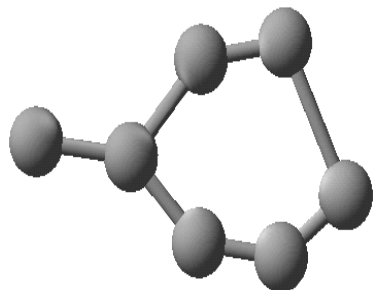
$$T_{\text{DATA}} = \sum_{\text{grid points}} (\rho_{\text{best}} - k\rho_{\text{calc}})^2$$

ρ_{best} - best available map: experimental, 2mFo-DFc
 ρ_{calc} - calculated map from current atomic model

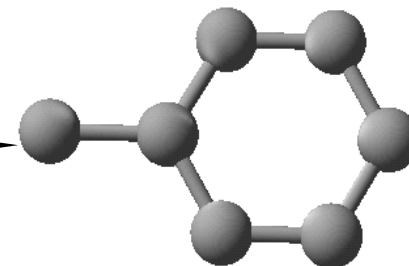
T_{DATA} : Least-Squares vs Maximum-Likelihood

- **Removable Errors** (never the case for macromolecular model, common for small molecules)

Complete model *before* refinement



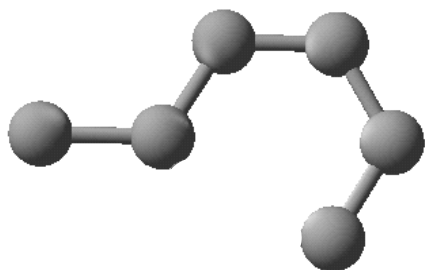
Least-Squares Target



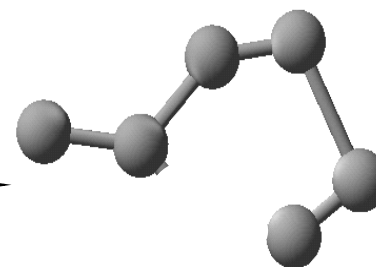
Complete model *after* refinement

- **Irremovable Errors** (always the case for macromolecular models)

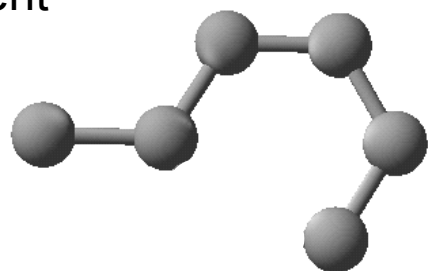
Partial model *before* refinement



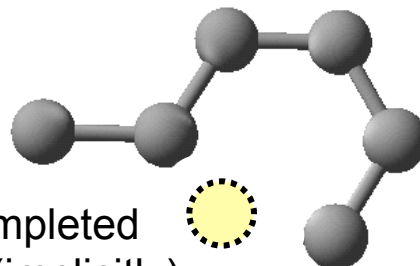
Least-Squares Target



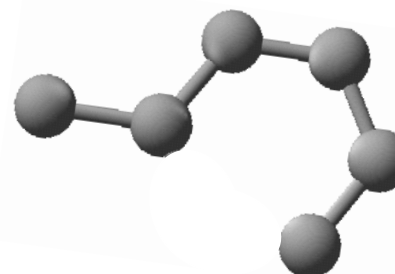
Partial model *after* refinement



Maximum-Likelihood Target



Model is completed statistically (implicitly)



Final model is less affected by incompleteness (by missing atoms)

Target function

- Maximum-Likelihood (reciprocal space; option of choice for macromolecules)

$$ML = T_{\text{DATA}} = \sum_s (1 - K_s^{cs}) \left(-\frac{\alpha_s^2 (F_s^{\text{MODEL}})^2}{\varepsilon_s \beta_s} + \ln \left(I_0 \left(\frac{2\alpha_s F_s^{\text{MODEL}} F_s^{\text{OBS}}}{\varepsilon_s \beta_s} \right) \right) \right) + K_s^{cs} \left(-\frac{\alpha_s^2 (F_s^{\text{MODEL}})^2}{2\varepsilon_s \beta_s} + \ln \left(\cosh \left(\frac{\alpha_s F_s^{\text{MODEL}} F_s^{\text{OBS}}}{\varepsilon_s \beta_s} \right) \right) \right)$$

- α and β account for model imperfection:
 - α is proportional to the error in atomic parameters and square of overall scale factor;
 - β is proportional to the amount of missing (unmodeled) atoms.
- α and β are estimated using test reflections by minimization of ML function w.r.t. α and β in each relatively thin resolution bin where α and β can be assumed constant.
 - This is why ML-based refinement requires *test set reflections*^(*) that should be defined sensibly:
 - Each resolution bin should contain at least 50 randomly distributed test reflections.

(*) *Test reflections* – a fraction of reflections (5-10%) put aside for cross-validation.

T_{DATA} : Least-Squares vs Maximum-Likelihood

- **Why Maximum-Likelihood target is better than Least-Squares** (in a nutshell):
 - ML accounts for model incompleteness (missing, unmodeled atoms) while LS doesn't;
 - ML automatically downweights the terms corresponding to reflections with the poor fit (poorly measured inaccurate F_{OBS} , high resolution reflections at the beginning of refinement, etc.)
- **R -factors in LS and ML refinement:**
 - R -factor is expected to decrease during LS based refinement, since the LS target and R -factor formula are very similar:

$$R = \frac{\sum |F_{\text{OBS}} - F_{\text{MODEL}}|}{\sum F_{\text{MODEL}}} \quad LS = \sum_s (F_{\text{OBS}} - F_{\text{MODEL}})^2$$

- In ML based refinement the R -factor may eventually decrease (and this is what typically happens in practice) but this is not implied by the ML target function

Real-space refinement – long history

- Booth, A. D. (1946). Proc. Roy. Soc. London Ser. A, 188, 77-92.
- Booth, A. D. (1947). Proc. Roy. Soc. London Ser A, 190, 482-489.
- Cochran, W. (1948). Acta Cryst. 1, 138-142.
- W. Cochran Acta Cryst. (1951). 4, 408-411
- Cruickshank, D. W. (1952). Acta Cryst. 5, 511-518.
- Cruickshank, D. W. (1956). Acta Cryst. 9, 747-753.
- R. Diamond Acta Cryst. (1971). A27, 436
- R. Diamond. J. Mol. Biol. (1974) 82, 371-391
- R.J. Fleterick and H.W. Wyckof Acta Cryst. (1975). A31,
- J.C. Hanson and B.P. Schoenborn J. Mol. Biol. (1981) 15
- S. Fitzwater and H. A. Scheraga Proc. NatL Acad. Sci. U
- R.Diamond. (1985). Methods Enzymol. 115, 237-252.
- S.Freer. (1985). Methods Enzymol. 115, 235-237.

R.J. Read and J. Moulton Acta Cryst. (1992). A48, 104-113
Fitting Electron Density by Systematic Search

- Deisenhofer, J., Remington, S. J. & Staigemann, W. (1985). Met
- A.G. Urzhumtsev, V. YU. Lunin and E. A. Vernoslava J. Appl. Cry
- T.A. Jones, J.-Y. Zou and S.W. Cowan, M. Kjeldgaard Acta Cryst
- R.J. Read and J. Moulton Acta Cryst. (1992). A48, 104-113
- V.S. Lamzin and K.S. Wilson Acta Cryst. (1993). D49, 129-147
- D.G. Levit, L.J. Banasza J. Appl. Cryst. (1993). 26, 736-745
- V.Y. Lunin, M.M. Woolfson Acta Cryst. (1993). D49, 530-533
- J.-Y. Zou, S.L. Mowbray Acta Cryst. (1994). D50, 237-249

T.J. Oldfield Acta Cryst. (2001). D57, 82-94
A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent

- M.S. Chapman Acta Cryst. (1995). A51, 69-80
- M.S. Chapman and M.G. Rossmann Acta Cryst. (1996). D52, 129-142
- V.Yu. Lunin and N.L. Lunina Acta Cryst. (1996). A52, 365-368
- J.-Y. Zou and T. A. Jones Acta Cryst. (1996). D52, 833-841
- E. Blanc and M.S. Chapman J. Appl Cryst (1997). 30, 566-567
- M.S. Chapman and E. Blanc Acta Cryst. (1997). D53, 203-206
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). Acta Cryst. D53, 1
- E. Blanc, Z. Chen and M.S. Chapman, Direct Methods for solving macromol
- Z. Chen, E. Blanc and M.S. Chapman. Acta Cryst. (1999). D55, 219-224
- Z. Chen, E. Blanc and M.S. Chapman. Acta Cryst. (1999). D55, 464-468

J.J. Headd, R. M. Immormino, D.A. Keedy, P. Emsley, D.C. Richardson, J.S. Richardson J Struct Funct Genomics (2009) 10:83–93
Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place

- G. Zhou, T. Somasundaram, E. Blanc, Z. Chena and M.S. Chapman, Acta Cryst. (1999). D55, 835-845
- T. Holton, T.R. Ioerger, J.A. Christopher and J.C. Sacchettini Acta Cryst. (2000). D56, 722-734
- G.J. Kleywegt Acta Cryst. (2000). D56, 249-265
- C.E. Wang Acta Cryst. (2000). D56, 1591-1611
- Z. Chen and M.S. Chapman Biophysical Journal Volume 80 March 2001 1466-1472
- L.F. Chen, E. Blanc, M.S. Chapman and K.A. Taylor Journal of Structural Biology 133, 221–232 (2001)
- D.G. Levitt Acta Cryst. (2001). D57, 1013-1019

- T.J. Oldfield Acta Cryst. (2001). D57, 82-94
- A. Korostelev, R. Bertramc, and M.S. Chapmana, Acta Cryst. (2002). D58, 71
- J.Z. Chen, J. Furst, M.S. Chapman and N. Grigorieff Journal of Structural Bi
- J.F. Hunt and J. Deisenhofer Acta Cryst. (2003). D59, 214-224
- Gao, H., Sengupta, J., Valle, M., Korostelev, A., Eswar, N., Stagg, S
- F. Pavelcik Acta Cryst. (2003). A59, 487-494
- S.X. Cohen, R.J. Morris, F.J. Fernandez, M. Ben Jelloul, M. Kakaris, V. Parth
- L. Potterton, S. McNicholas, E. Krissinel, J. Gruber, K. Cowtan, P. Emsley, G
- J. Aishima, D.S. Russel, L.J. Guibas, P.D. Adams and A.T. Brungera Acta Cr
- H. van den Bedem, I. Lotan, J.-C. Latombe and A.M. Deacon Acta Cryst. (20
- M.S. Chapman, F. Fabiola, A. Korostelev, M. Fenley Acta Cryst. (2005). A61
- B. DeLaBarre and A.T. Brunger Acta Cryst. (2006). D62, 923–932

A. G. Urzhumtsev, V. Yu. Lunin and E. A. Vernoslava J. Appl. Cryst. (1989). 22, 500-506
FROG - high-speed restraint-constraint refinement program for macromolecular structure
Real + reciprocal space refinement of coordinates, B-factors, occupancies, rigid groups

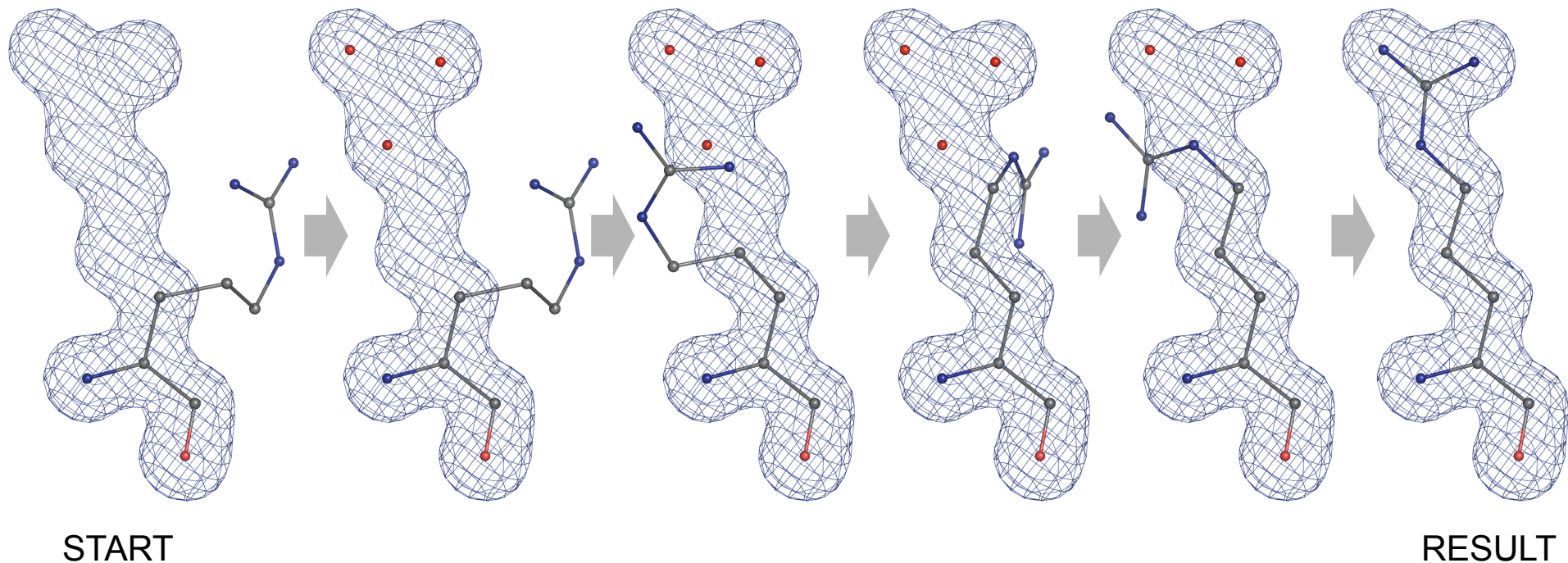
- N. Furnham, A.S. Dore, D.Y. Chirgadze, P.I.W. de Bakker, M.A. DePristo, and T. L. Blundell Structure 14, 1313–1320, August 2006
- S. Wlodek, A. G. Skillman and A. Nicholls Acta Cryst. (2006). D62, 741–749
- K. Joosten, S.X. Cohen, P. Emsley, W. Mooij, V.S. Lamzin and A. Perrakis Acta Cryst. (2008). D64, 416–424
- J. L. Knight, Z. Zhou, E. Gallicchio, D.M. Himmel, R.A. Friesner, E. Arnold and R. M. Levy Acta Cryst. (2008). D64, 383–396
- D. Turk Acta Cryst. (2008). A64, C23
- H. van den Bedem, A. Dhanik, J.-C. Latombe and A.M. Deacon Acta Cryst. (2009). D65, 1107–1117
- J.J. Headd, R. M. Immormino, D.A. Keedy, P. Emsley, D.C. Richardson, J.S. Richardson J Struct Funct Genomics (2009) 10:83–93

Dual-space refinement: combining real and reciprocal space refinement

Why real-space refinement ?

- Can be done locally (for example, for a residue or ligand)
- Grid search can be used -> Convergence radius can be dramatically increased compared to gradient driven-refinement or SA
- Ordered solvent update can be enabled at earlier stage

✓ **Eliminate the tedium of manual work on fixing side chains on graphics**



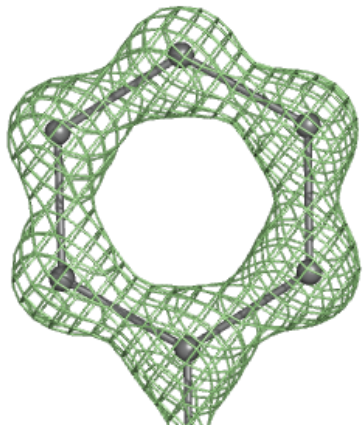
Real-space refinement

$$\text{Real space refinement target } T = w \sum_{\text{grid points}} (\rho_{\text{best}} - k\rho_{\text{calc}})^2 + T_{\text{RESTRAINTS}}$$

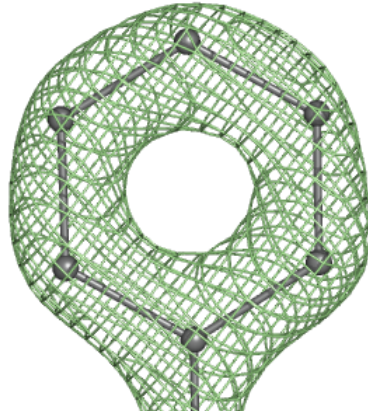
ρ_{best} is best available map: experimental, 2mFo-DFc, ...

ρ_{calc} is calculated map from current atomic model

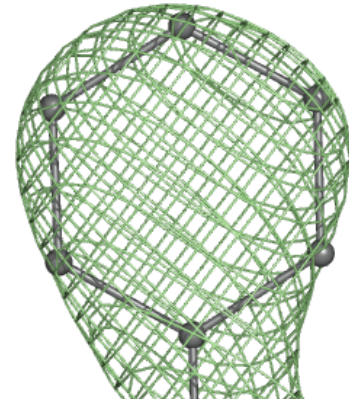
Fitting ρ_{calc} to ρ_{best} may be problematic because the exact ρ_{calc} computed from atomic model and its Fourier image (ρ_{best}) may look very different depending on resolution and data completeness:



exact



2 Å



3 Å

... so using the exact ρ_{calc} directly computed from atomic model may not be a good idea.

Real-space refinement

Solutions:

- Resolution and completeness dependent analytical functions for ρ_{calc} (M.S. Chapman; used in RSRef – a real-space refinement extension of CNS)
- Compute $\rho_{\text{calc}} = \text{FT}(\mathbf{F}_{\text{MODEL}})$ that naturally accounts for resolution and completeness.

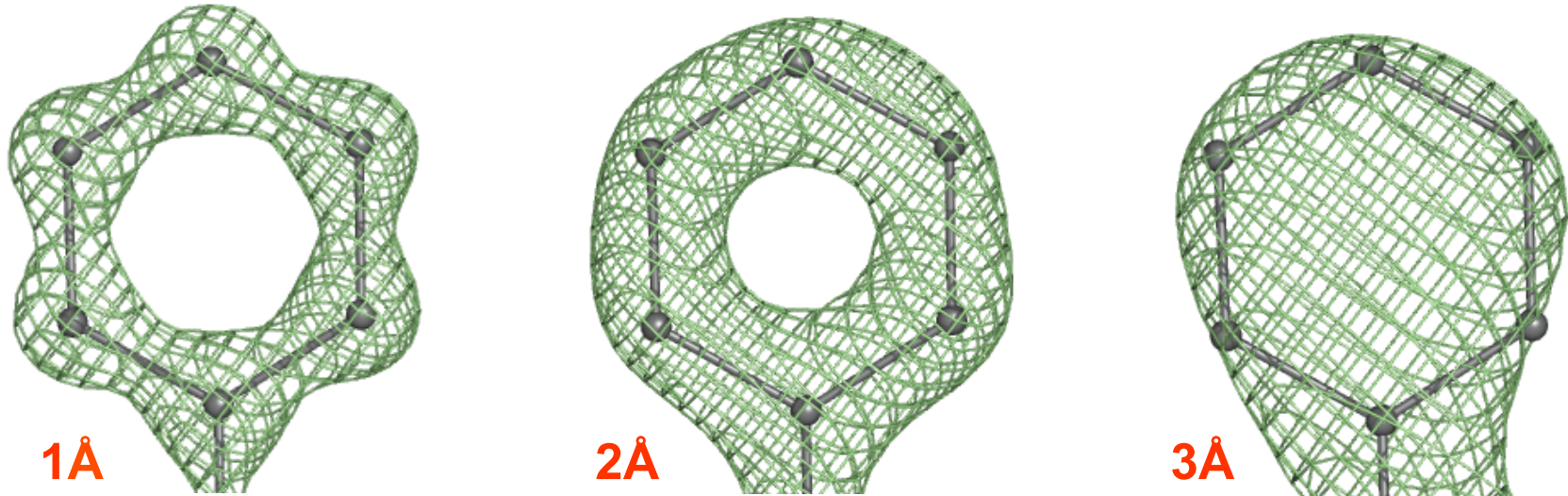
Alternative – more simplistic target that moves atoms to the closest density peak:

$$T = -w \sum_{\text{atoms}} \rho_{\text{best}} \Big|_{\text{computed at atom center}} + T_{\text{RESTRAINTS}}$$

- Fast (no need to re-compute ρ_{calc} each time an atom moved)
- Uses only one map (no issues related to dissimilarity of ρ_{calc} and ρ_{best} due to resolution)
- Less accurate since doesn't use shape of electron density (moves atoms to the closest density peak without considering how similar that peak is to expected density. May not be applicable at low resolution where the atomicity of the map is lost (no distinct peaks corresponding to atoms, but rather sphere and tube-like shapes).

Restraints in refinement of individual coordinates

Fourier images at different data resolution:



- At lower resolution the electron density is not informative enough to keep the molecule geometry sensible
- Therefore there is a need to bring in some additional a priori knowledge that we may have about the molecules in order to keep the geometry ...
- This knowledge is typically expressed *either* as an additional term to the refinement target (**restraints** term):

$$E_{\text{TOTAL}} = w * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$$

or strict requirement that the model parameter must exactly match the prescribed value and never change during refinement (**constraints**).

Restraints in refinement of individual coordinates

- A *a priori* chemical knowledge (restraints) is introduced to keep the model chemically correct while fitting it to the experimental data at lower resolution (less resolution, stronger the weight W):

$$E_{\text{TOTAL}} = w * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$$

$$E_{\text{RESTRAINTS}} = E_{\text{BOND}} + E_{\text{ANGLE}} + E_{\text{DIHEDRAL}} + E_{\text{PLANARITY}} + E_{\text{NONBONDED}} + E_{\text{CHIRALITY}} + E_{\text{NCS}} + E_{\text{RAMACHANDRAN}} + E_{\text{REFERENCE}} + \dots$$

- Higher resolution – less restraints contribution (can be completely unrestrained for well ordered parts at subatomic resolution).
- Typically, each term in $E_{\text{RESTRAINTS}}$ is a harmonic (quadratic) function:
 $E = \sum \textit{weight} * (X_{\text{model}} - X_{\text{ideal}})^2$
- $\textit{weight} = 1/\sigma(X)^2$ is the inverse variance, in least-squares methods (e.g. 0.02 Å for a bond length)
- Making $\sigma(X)$ too small is NOT equivalent to constraints, but will make weight infinitely large, which in turn will stall the refinement.

Restraints: bonds and angles

- Bond distances:

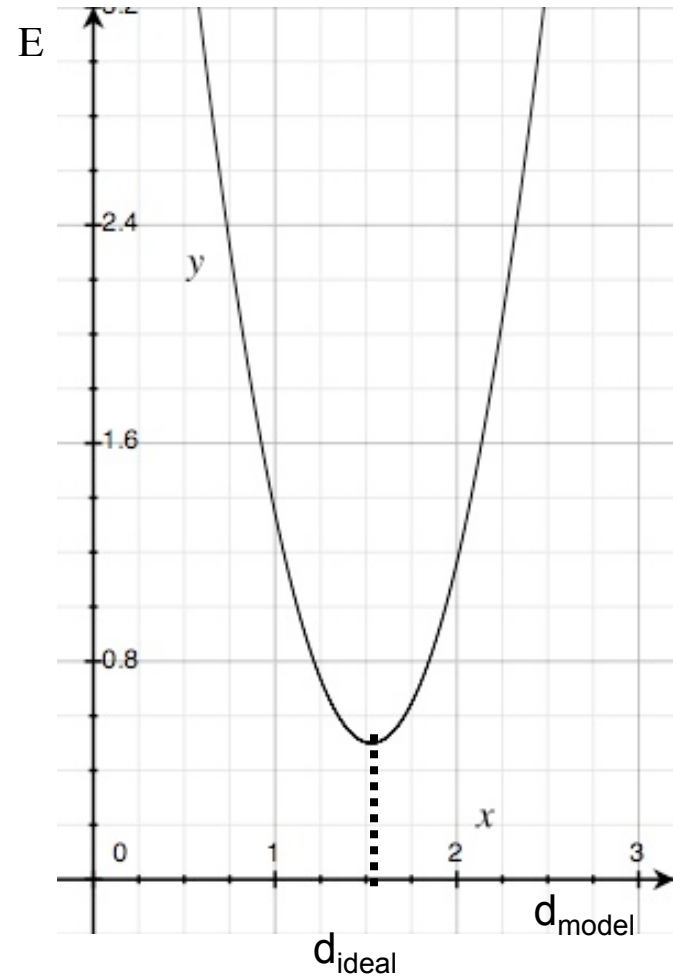
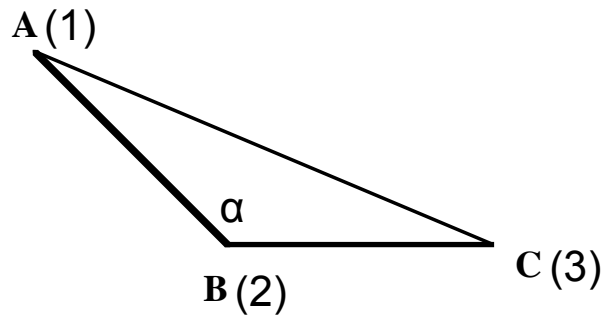
$$E = \sum_{\text{bonds}} \text{weight} * (d_{\text{model}} - d_{\text{ideal}})^2$$

- Bond angles:

$$E = \sum_{\text{angles}} \text{weight} * (\alpha_{\text{model}} - \alpha_{\text{ideal}})^2$$

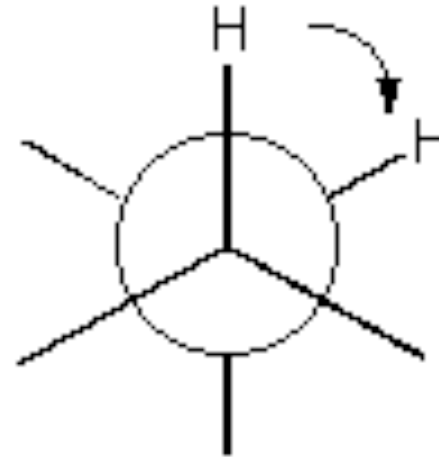
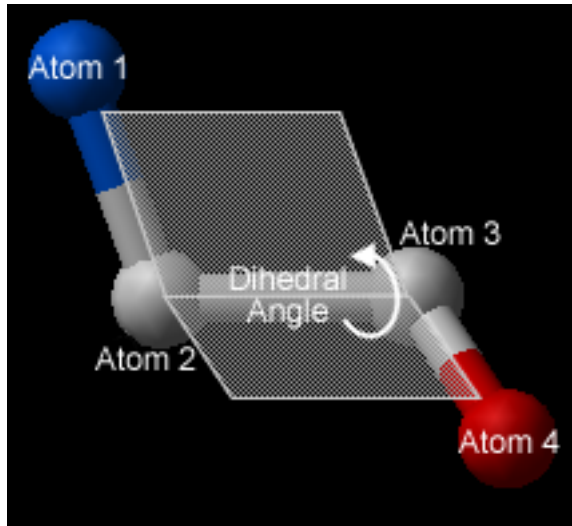
Alternatively, one can restrain 1-3 distances:

$$E = \sum_{\text{1-3-pairs}} \text{weight} * (d_{\text{model}} - d_{\text{ideal}})^2$$



Restraints: dihedral (torsion) angles

- Dihedral or torsion angle is defined by 4 sequential bonded atoms 1-2-3-4
 - Dihedral = angle between the planes 123 and 234
 - Torsion = looking at the projection along bond B-C, the angle over which one has to rotate A to bring it on top of D (clockwise = positive)



- Three possible ways to restraining dihedrals:
 - $E = \sum_{\text{dihedrals}} \text{weight} * (\chi_{\text{ideal}} - \chi_{\text{model}})^2$ (if only one target value for the dihedral)
 - $E = \sum_{\text{dihedrals}} \text{weight} * (1 + \cos(n \chi_{\text{model}} + \chi_{\text{shift}}))$ (n = periodicity)
 - $E = \sum_{1-4\text{-pairs}} \text{weight} * (d_{\text{model}} - d_{\text{ideal}})^2$
(sign ambiguity unless $\chi = 0^\circ$ or 180° , *i.e.* both χ and $-\chi$ give rise to the same 1-4 distances)

Restraints: chirality

- A chiral molecule has a non-superposable mirror image
- Chirality restraints (example: for C_α atoms) defined through chiral volume:

$$V = (\mathbf{r}_N - \mathbf{r}_{CA}) \cdot [(\mathbf{r}_C - \mathbf{r}_{CA}) \times (\mathbf{r}_{CB} - \mathbf{r}_{CA})]$$

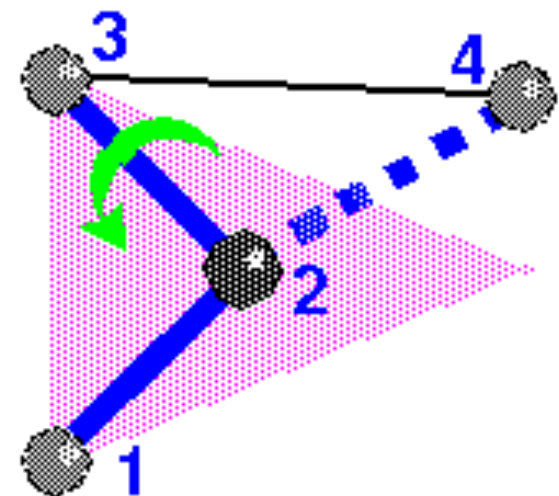
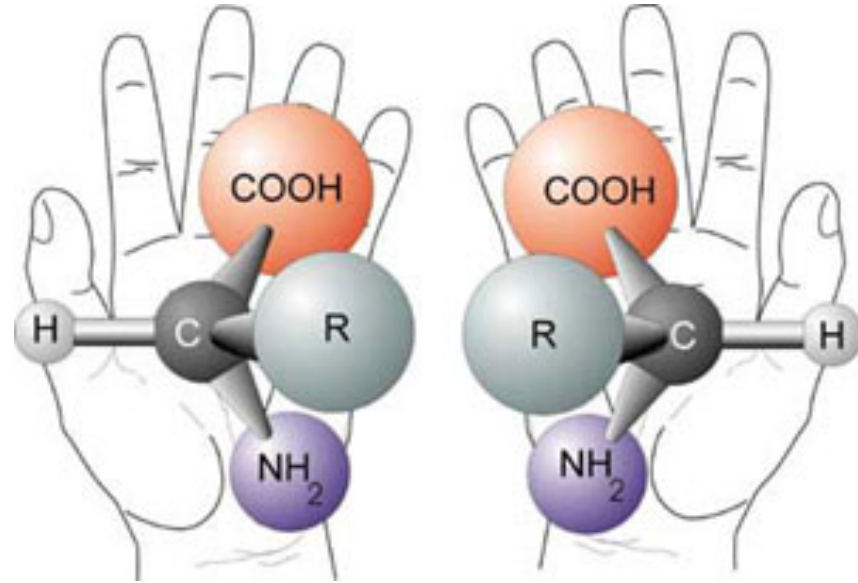
sign depends on handedness ($V_D = -V_L$)

$$E = \sum_{\text{chiral}} \textit{weight} * (V_{\text{model}} - V_{\text{ideal}})^2$$

- Alternatively, chirality restraints can be defined by an “improper torsion” (“improper”, because it is not a torsion around a chemical bond)

Example: for C_α : torsion (C_α -N-C- C_β)
= $+35^\circ$ for L-aa, -35° for D-aa

$$E = \sum_{\text{chiral}} \textit{weight} * (\chi_{\text{ideal}} - \chi_{\text{model}})^2$$

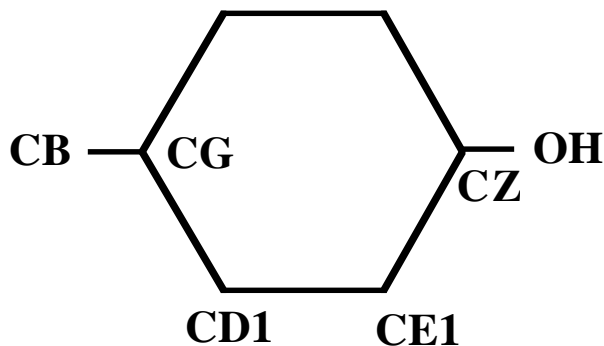


Restraints: planarity

- Planarity (double bonds, aromatic rings):
 - Identify a set of atoms that has to be in plane, and then for each set, minimise sum of distances to the best-fitting plane through the atoms

$$E = \sum_{\text{planes}} \sum_{\text{atoms_in_plane}} \text{weight} * (\underline{\mathbf{m}} \cdot \underline{\mathbf{r}} - d)^2$$

- Restrain the distances of all atoms in the plane to a dummy atom that lies removed from the plane
- Define a set of (“fixed”, “non-conformational”) dihedral angles (or improper torsions) with target values of 0° or 180°:



$$(\text{CB-CG-CD1-CE1}) = 180$$

$$(\text{CG-CD1-CE1-CZ}) = 0$$

$$(\text{CD1-CE1-CZ-OH}) = 180$$

$$(\text{CD1-CE1-CZ-CE2}) = 0$$

$$(\text{CE1-CZ-CE2-CD2}) = 0$$

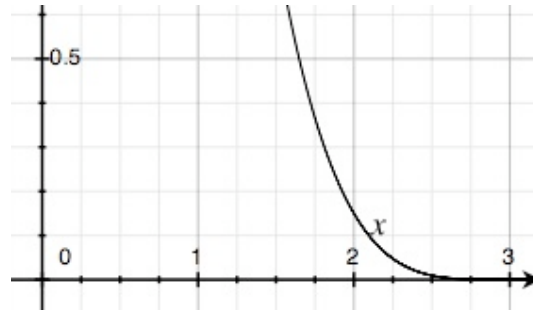
$$(\text{CZ-CE2-CD2-CG}) = 0$$

$$(\text{CE2-CD2-CG-CD1}) = 0$$

$$(\text{CD2-CG-CD1-CE1}) = 0$$

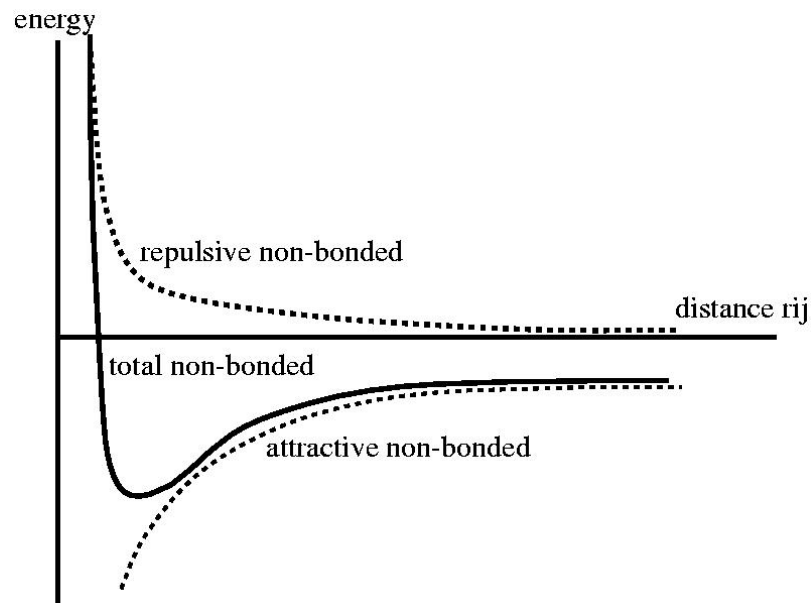
Restraints: non-bonded

- Simple repulsive term: $E = \sum_{nb} weight * (d_{model} - d_{min})^4$ (only if $d_{model} < d_{min}$)



- Combined function: Van der Waals and electrostatics terms

$$E = E_{attractive} + E_{repulsive} + E_{electrostatic} = \sum_{nb} (A d_{model}^{-12} - B d_{model}^{-6} + C q_1 q_2 / d_{model})$$

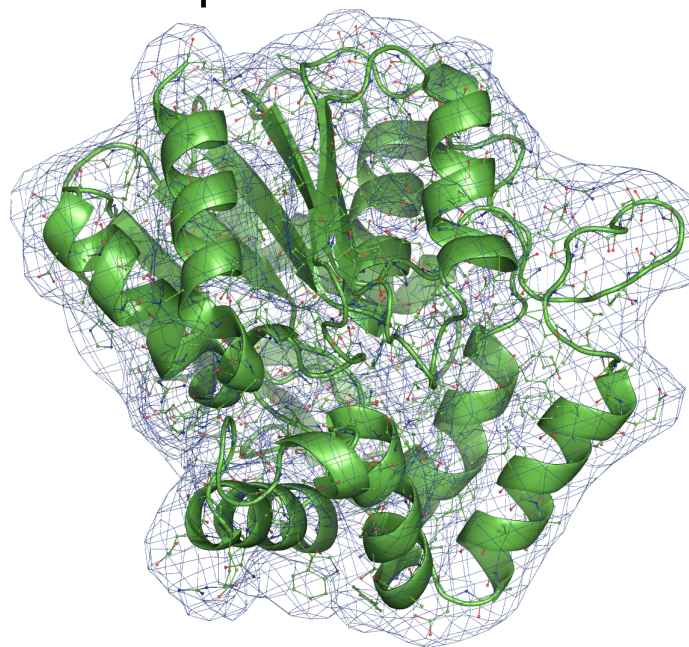


Sources of target (“ideal”) values for constraints and restraints

- Libraries (for example, Engh & Huber) created out of small molecules that are typically determined at much higher resolution, use of alternative physical methods (spectroscopies, etc).
- Analysis of macromolecular structures solved at ultra-high resolution
- Pure conformational considerations (Ramachandran plot), tabulated secondary structure parameters
- QM (quantum-chemical) calculations

Specific restraints for refinement at low and very low resolution

- At low(ish) resolution the electron density map is not informative enough and a set of local restraints are insufficient to maintain known higher order structure (secondary structure), and the amount of data is too small compared to refinable model parameters ...

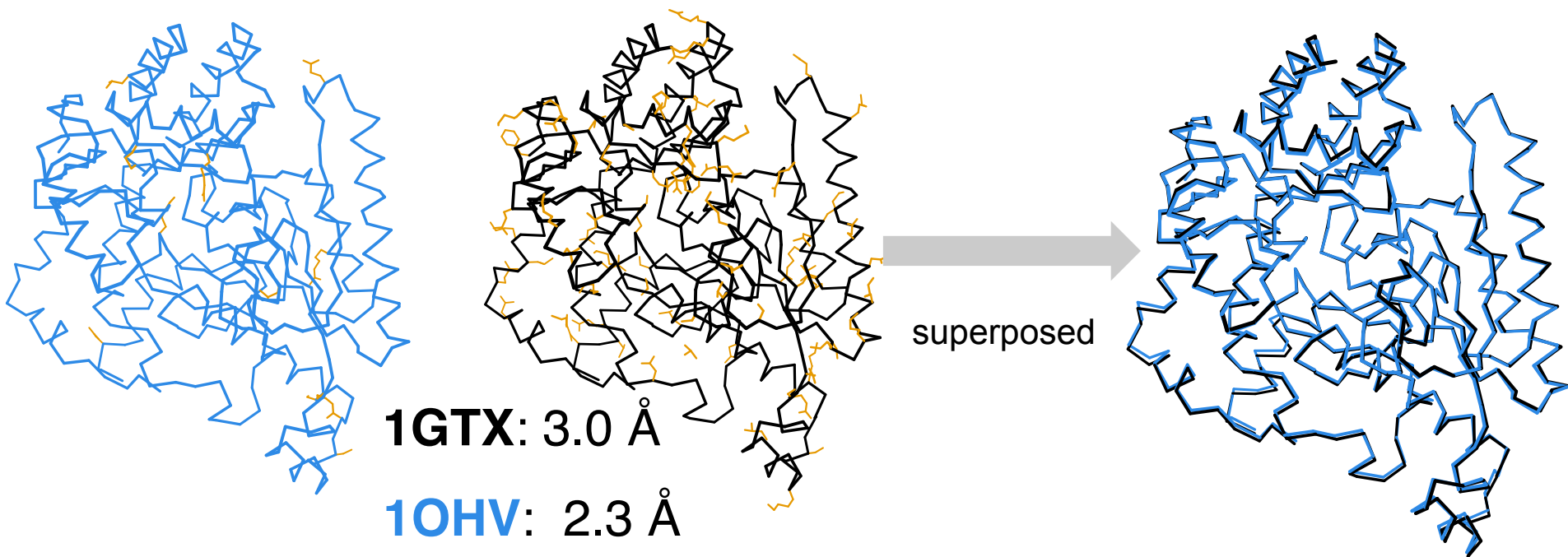


- ... therefore one needs to bring in more information in order to assure the overall correctness of the model:
 - Reference model or point
 - Secondary structure restraints
 - Ramachandran restraints
 - NCS restraints/constraints

Specific restraints for refinement at low and very low resolution

- **Reference model:**

- If you are lucky, there may be a higher resolution structure available that is similar to low resolution structure
- Use higher resolution information to direct low-resolution refinement



- Reference point restraint for isolated atoms (water / ions): sometime density peak may not be strong enough to keep an atom in place (due to low resolution or low site occupancy, for example), so it can drift away from it. Use harmonic restraint to peak position.

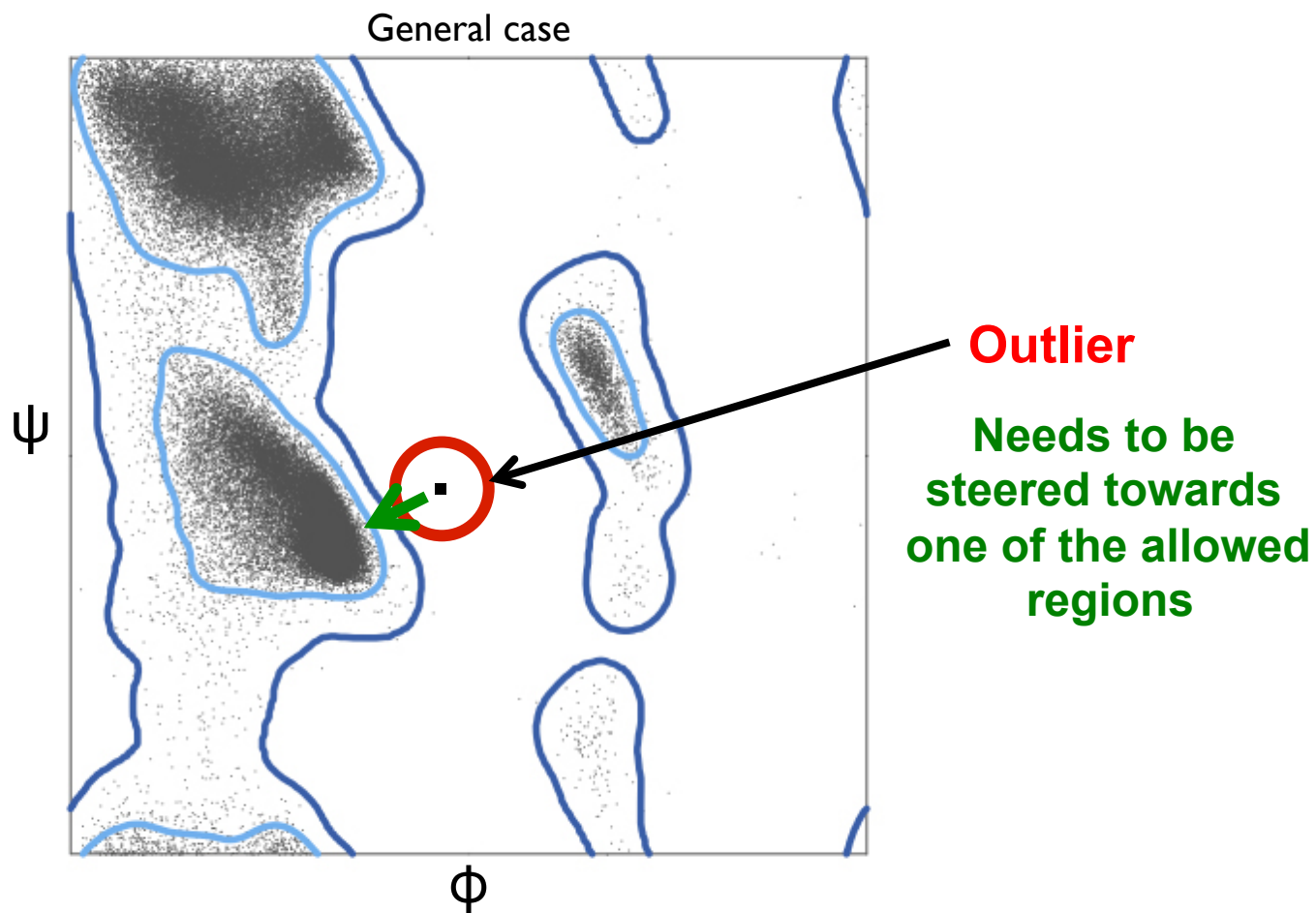
Specific restraints for refinement at low and very low resolution

- **Secondary structure restraints**
 - H-bond restraints for alpha helices, beta sheets, RNA/DNA base pairs
 - This requires correct annotation of secondary structure elements:
 - It can be done automatically using programs like DSSP / KSDSSP
 - Or... manually (quite an exercise for a Ribosome structure!)

Specific restraints for refinement at low and very low resolution

- **Ramachandran restraints**

- steer outliers towards favored region
- should only be used at low resolution
- should never be used at higher resolution, since it is one of the few precious validation tools (sometimes compare to “real-space analog of Rfree”)



Ramachandran plot restraints

research papers

Acta Crystallographica Section D
**Biological
Crystallography**
ISSN 0907-4449

Features and development of *Coot*

P. Emsley,^{a*} B. Lohkamp,^b
W. G. Scott^c and K. Cowtan^d

Coot is a molecular-graphics application for model building and validation of biological macromolecules. The program displays electron-density maps and atomic models and allows model manipulations such as idealization, real-space refine-

Restraints on Ramachandran plot distribution w/ added weak gradient for flat regions

research papers

Acta Crystallographica Section D
**Biological
Crystallography**
ISSN 0907-4449

A number of real-space torsion-angle refinement techniques for proteins, nucleic acids, ligands and solvent

Thomas J. Oldfield

This paper describes the implementation of real-space torsion-angle refinement as a tool for model (re)building. The algorithmic details and parameterization for a number of

Received 3 August 2000
Accepted 9 October 2000

Simple restraint that drives outliers toward nearest allowed point in Ramachandran plot

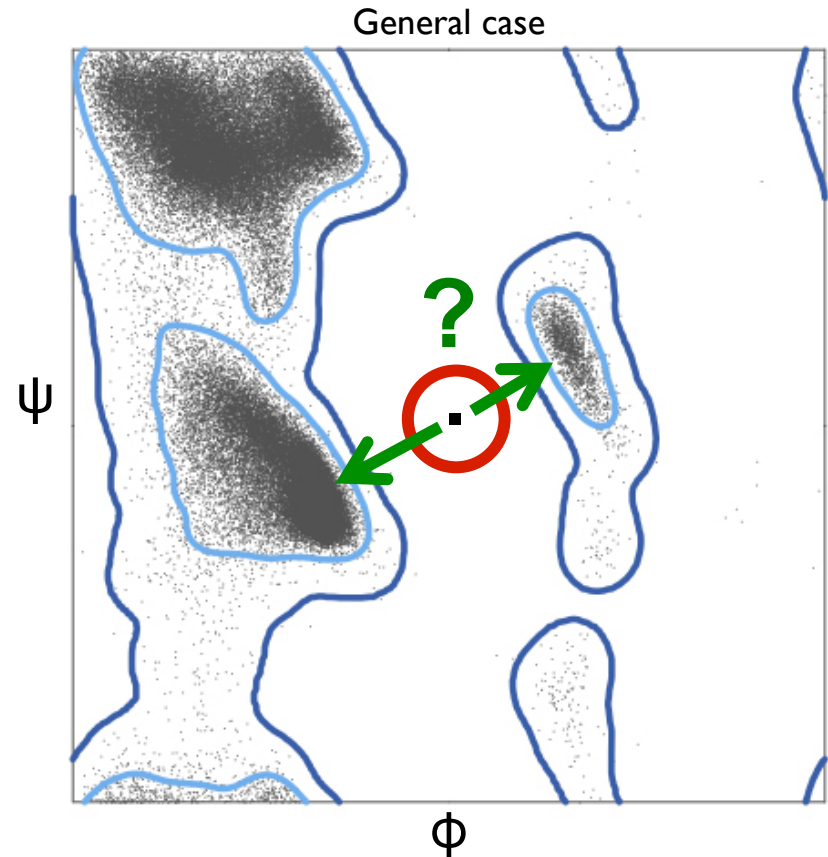
Ramachandran plot restraints: Oldfield

- Simple harmonic potential:

$$E = \sum weight * (\varphi_{model} - \varphi_{target})^2 + \sum weight * (\psi_{model} - \psi_{target})^2$$

- φ_{target} and ψ_{target} are determined based on the distance of the outlier to the closest allowed region, and updated during refinement every time the model coordinates are changed.
- Potential problem: in ambiguous cases a residue can be locked in a wrong region:

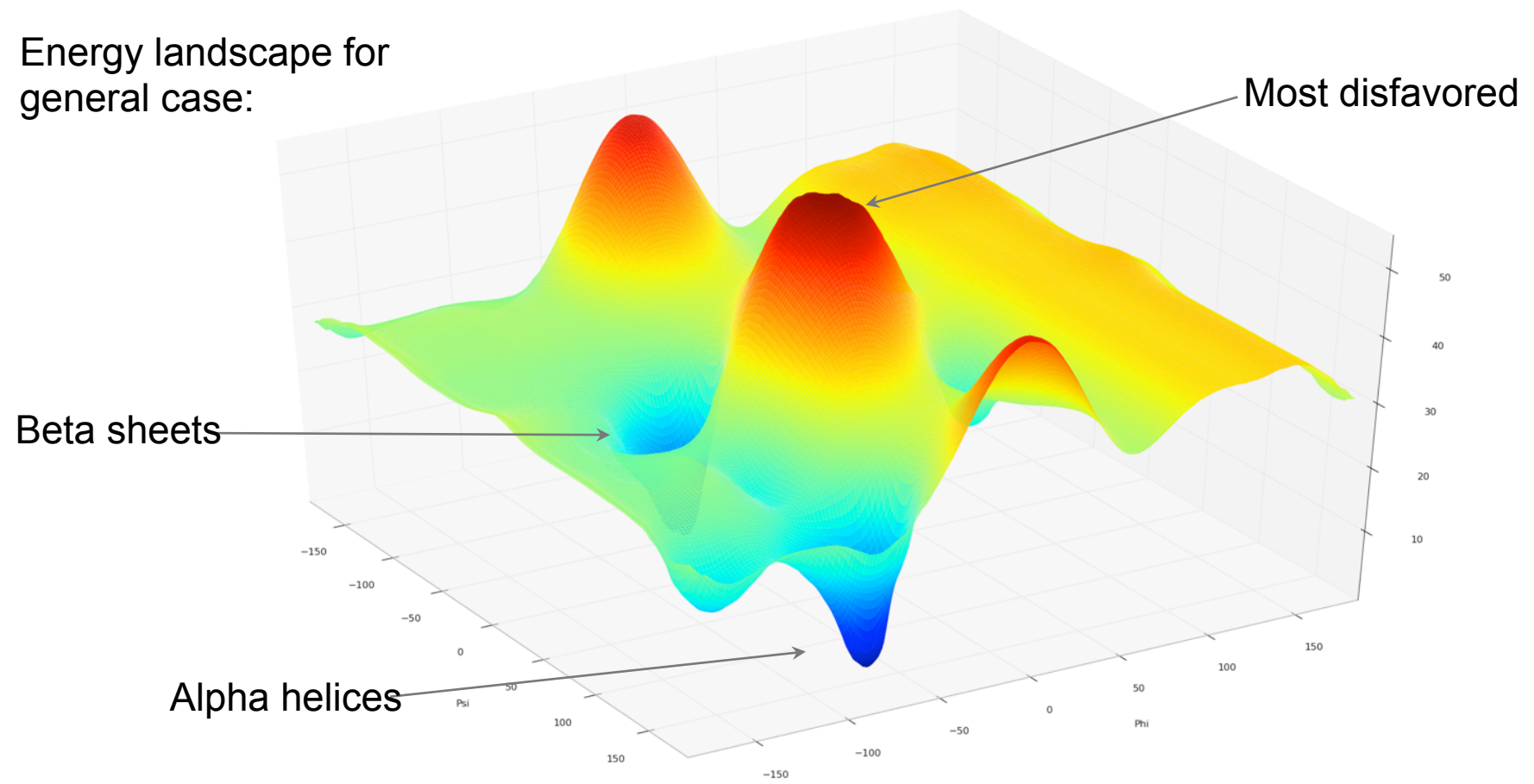
- Every time the Ramachandran restraints are used on a structure with Ramachandran outliers, check what happened to these outliers after refinement (Ramachandran restraints will eliminate the outliers, but where it will put them – is a big question!)



Ramachandran plot restraints (Enhanced Ramachandran pseudo-energy)

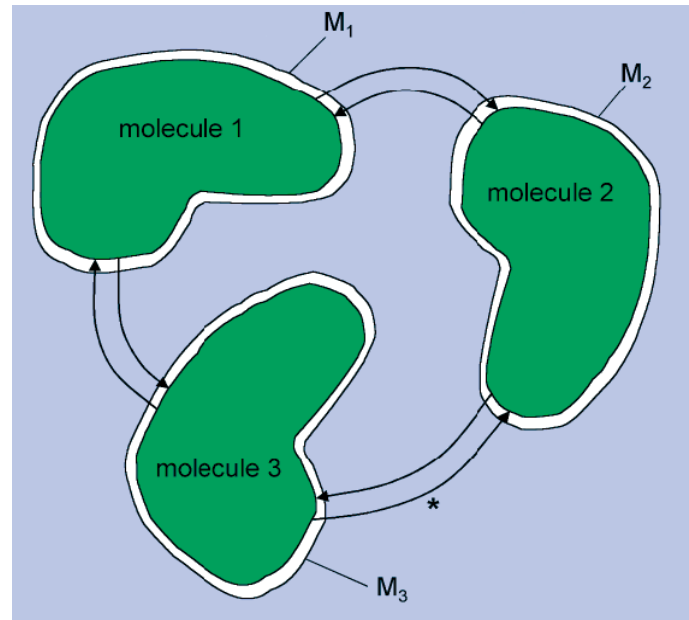
- Similar to what Coot has, but uses MolProbity clashscore for dipeptides to amplify disallowed regions
- Ramachandran plot is not a binary function anymore, but is a “continuous” function with small gradient in disallowed regions towards the allowed ones

Energy landscape for general case:



Specific restraints for refinement at low and very low resolution: NCS

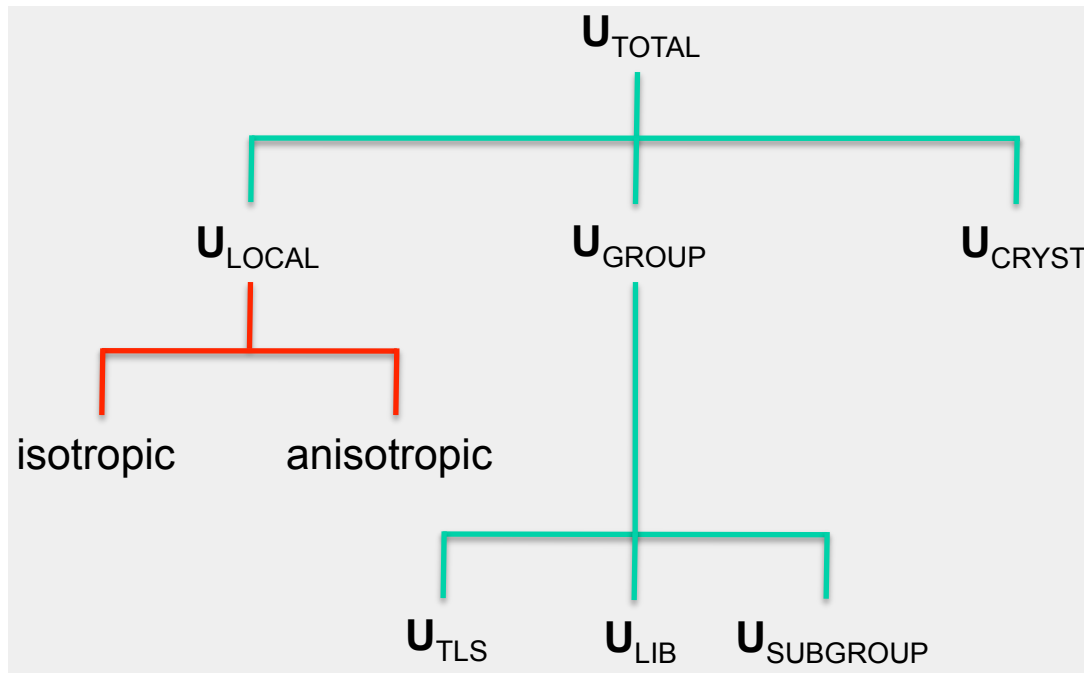
- **NCS (non-crystallographic symmetry) restraints/constraints**
 - Multiple copies of a molecule/domain in the asymmetric unit that are assumed to have similar conformations (and sometimes B-factors)
 - Restrain positional deviations from the average structure
$$E = \sum_{\text{atoms}} \textit{weight} * \sum_{\text{NCS}} |\mathbf{r} - \langle \mathbf{r} \rangle|^2$$
 - Different weights for different parts of the model possible



NCS restraints and B-factors

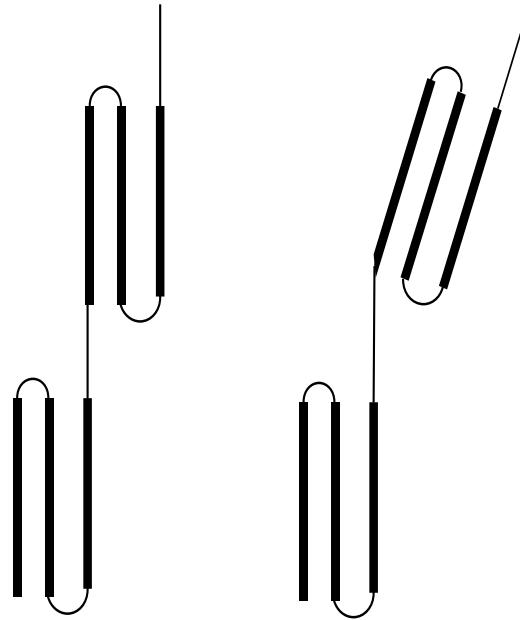
- **NCS (non-crystallographic symmetry) restraints/constraints**
 - Similarly for B-factors: $E = \sum_{\text{atoms}} \text{weight} * \sum_{\text{NCS}} (B - \langle B \rangle)^2$
 - In case when TLS is used, the NCS is applied to $\mathbf{U}_{\text{LOCAL}}$

Total ADP: $\mathbf{U}_{\text{TOTAL}} = \mathbf{U}_{\text{CRYST}} + \mathbf{U}_{\text{GROUP}} + \mathbf{U}_{\text{LOCAL}}$



Specific restraints for refinement at low and very low resolution: NCS

- Potential problem when using position-based NCS restraints:
 - Restraining whole will introduce substantial errors (hinge does not obey NCS)



- Solution:
 - Need to use finer-grained NCS groups (in this example treat each domain separately), OR
 - Instead of restraining atomic positions, restrain the orientation of atom with respect to its neighbours → construct restraint target in torsion angle space.

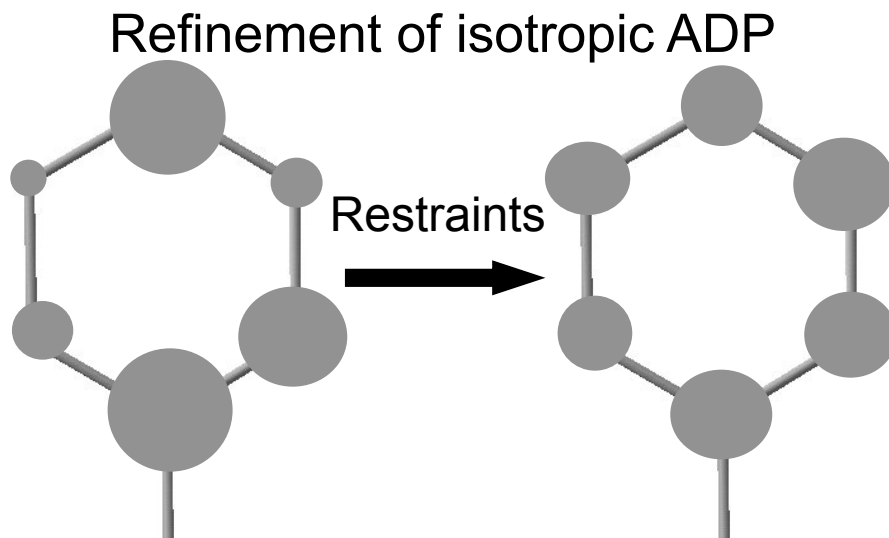
Ramachandran, secondary structure and NCS restraints: when to use ?

- Ramachandran and secondary structure restraints should be used only at very low resolution^(*), when you essentially should use it to assure correctness of your structure (~3-3.5Å or even lower, depends on data and model quality)
- NCS restraints:
 - Unlike Ramachandran and secondary-structure, NCS restraints should be used at higher resolution (2Å and lower)
 - Some big crystallography names state that NCS should always be used in refinement (if available)
 - This is not quite true: at higher resolution, say lower than 2Å, using NCS may rather harm than help, because it may wipe out the naturally occurring differences between NCS-related copies visible at that resolutions
 - Suggestion: simply try refining with and without NCS restraints and see what works better – this is the most robust way to find out!

() Urzhumtsev, A., Afonine, P.V. & Adams P.D. (2009). On the use of logarithmic scales for analysis of diffraction data. Acta Cryst. D65, 1283-1291.*

Restraints in refinement of individual isotropic ADP

$$E_{\text{TOTAL}} = W * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$$



- Similarity restraints: $E = \sum_{\text{all pairs of bonded atoms}} \textit{weight} * (B_i - B_j)^2$
- Knowledge-based restraints: $E = \sum_{\text{all pairs of bonded atoms}} \textit{weight} * (|B_i - B_j| - \Delta_{ij})^2$
where Δ_{ij} comes from a library of values collected from well-trusted structures for given type of atoms.

Restraints in refinement of individual isotropic ADP

$$E_{\text{TOTAL}} = W * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$$

- A better way of defining restraints for isotropic ADPs is based on the following facts:
 - A bond is almost rigid, therefore the ADPs of bonded atoms are similar (Hirshfeld, 1976);
 - ADPs of spatially close (non-bonded) atoms are similar (Schneider, 1996);
 - The difference between the ADPs of bonded atoms, is related to the absolute values of ADPs. Atoms with higher ADPs can have larger differences (Ian Tickle, CCP4 BB, March 14, 2003).

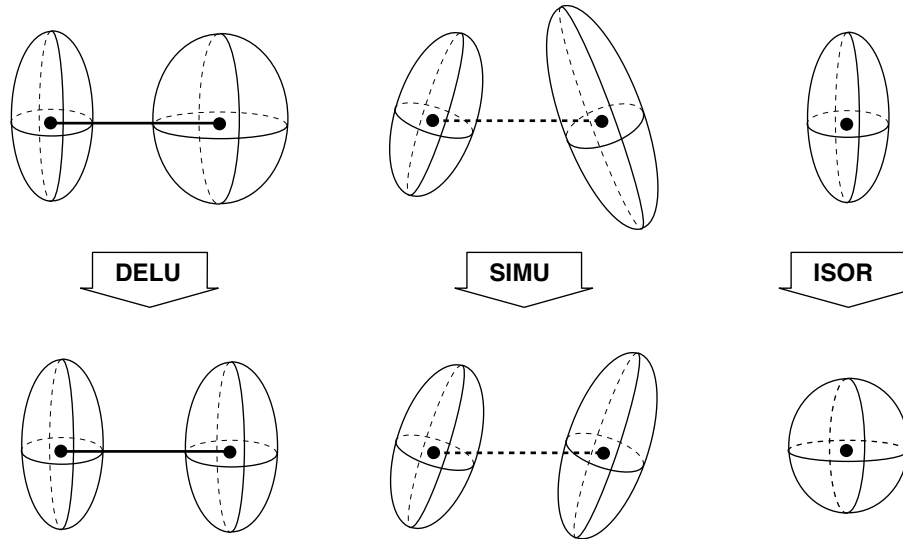
$$E_{\text{RESTRAINTS}} = \sum_{i=1}^{N_{\text{ALL ATOMS}}} \left[\sum_{j=1}^{M_{\text{ATOMS IN SPEHERE}}} \frac{1}{r_{ij}^{\text{distance_power}}} \frac{(U_i - U_j)^2}{\left(\frac{U_i + U_j}{2}\right)^{\text{average_power}}} \Big|_{\text{sphereR}} \right]$$

- Distance power, average power and sphere radius are some empirical parameters

Restraints in refinement of individual anisotropic ADP

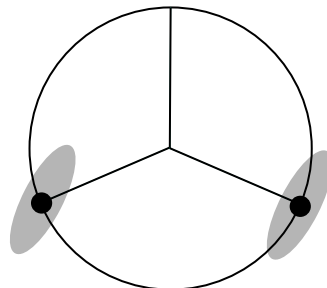
$$E_{\text{TOTAL}} = w * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$$

- Restraints for anisotropic ADP

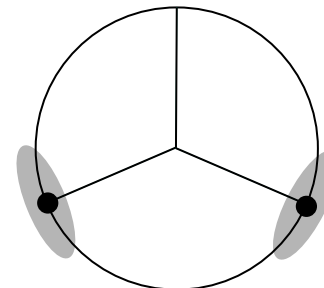


Picture stolen from Tom Schneider

- Caveat: none of the above restraints will do good in this case



Wrong

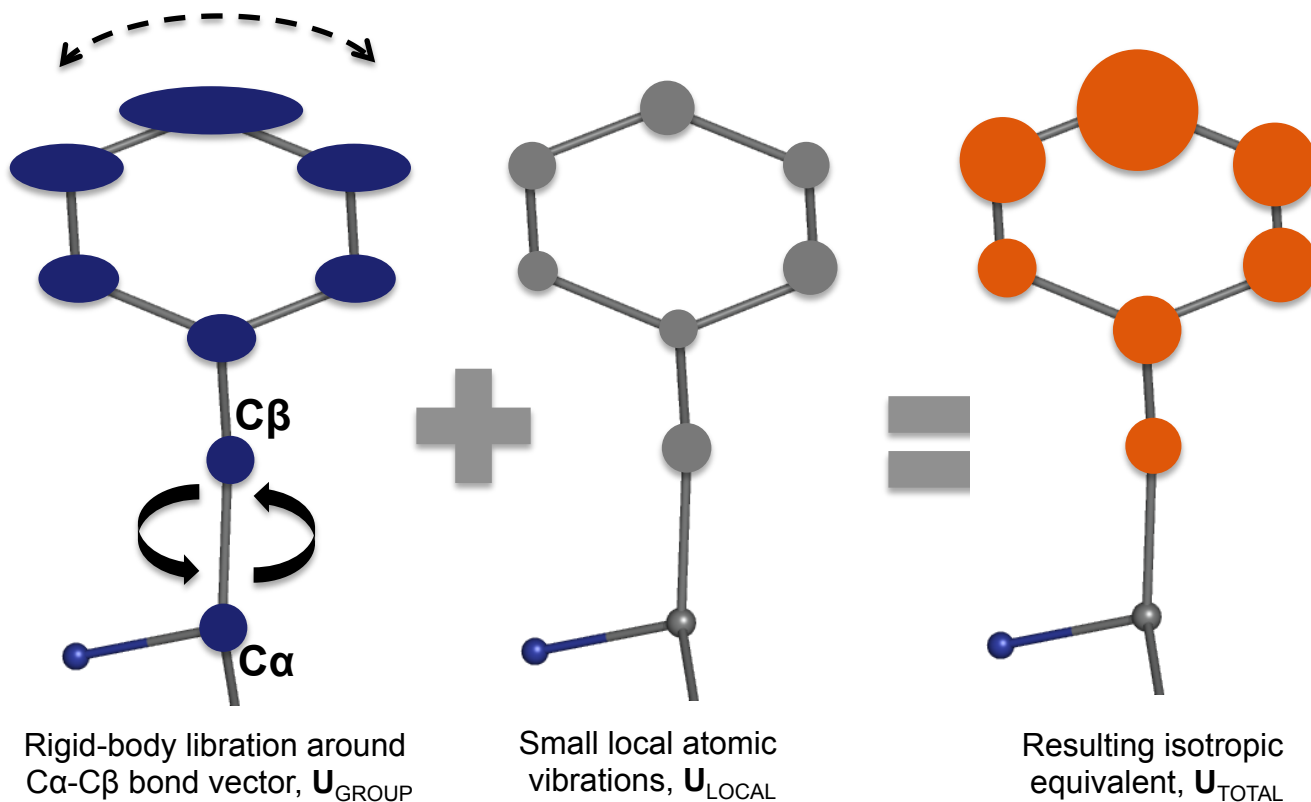


Correct

Restraints in refinement of individual ADP

▪ A nuance about using similarity restraints

- Total ADP is: $\mathbf{U}_{\text{TOTAL}} = \mathbf{U}_{\text{CRYST}} + \mathbf{U}_{\text{GROUP}} + \mathbf{U}_{\text{LOCAL}}$
- Similarity restraints should be applied to $\mathbf{U}_{\text{LOCAL}}$
- Applying it to $\mathbf{U}_{\text{TOTAL}}$ is much less justified

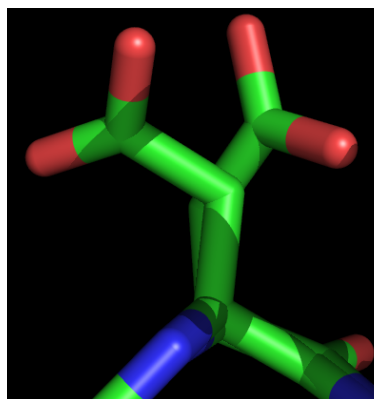


Example of constraints

- Rigid body refinement: mutual positions of atoms within a rigid groups are forced to remain the same, while the rigid group can move as a whole. 6 refinable parameters per rigid group (3 translations + 3 rotations).
- Constrained rigid groups: torsion angle parameterization. Reduction of refinable parameters by a factor between 7 and 10.
- Occupancies of atoms in alternative conformations: occupancies of alternate conformers must add up to 1.
- Group ADP refinement: mutual distribution of all B-factors within the group must remain the same. One refinable B-factor per group.
- Constrained NCS refinement: a number of N NCS related molecules or domains are assumed to be identical. Reduction of refinable parameters by a factor N .

- Do not confuse restraints and constraints
 - Constraints: model property = ideal value
 - Restraints: model property \sim ideal value

Constraints in occupancy refinement



- As it stands, occupancy refinement is always a constrained refinement...
- When we do not refine occupancy we essentially constrain its value to whatever value comes from input model (typically 1)

- Refining occupancies of alternative conformations we apply two constraints:
 - Occupancies of atoms within each conformer must be equal
 - Sum of occupancies for each set of matching atoms taken over all conformers must add to 1. Ideally, it should be less than or equal to 1, since we may not be including all existing conformers; however inequality constraints are very hard to handle in refinement.

ATOM	1	N	AARG	A	192	-5.782	17.932	11.414	0.72	8.38	N
ATOM	2	CA	AARG	A	192	-6.979	17.425	10.929	0.72	10.12	C
ATOM	3	C	AARG	A	192	-6.762	16.088	10.271	0.72	7.90	C
ATOM	7	N	BARG	A	192	-11.719	17.007	9.061	0.28	9.89	N
ATOM	8	CA	BARG	A	192	-10.495	17.679	9.569	0.28	11.66	C
ATOM	9	C	BARG	A	192	-9.259	17.590	8.718	0.28	12.76	C

Refinement target weight (MORE DETAILS)

- Refinement target $E_{\text{TOTAL}} = w * E_{\text{DATA}} + E_{\text{RESTRAINTS}}$
 - the weight w is determined automatically
 - in most of cases the automatic choice is good
- If automatic choice is not optimal there are two possible refinement outcomes:
 - structure is over-refined: *Rfree-Rwork* is too large. This means the weight w is too small making the contribution of E_{DATA} too large.
 - weight w is too large making the contribution of restraints too strong. This results increase of *Rfree* and/or *Rwork*.
 - A possible approach to address this problem is to perform a grid search over an array of w values and choose the one w that gives the best *Rfree* and *Rfree-Rwork*.
- A random component is involved in w calculation. Therefore an ensemble of identical refinement runs each done using different random seed will result in slightly different structures. The *R*-factor spread depends on resolution and may be as large as 1...2%.

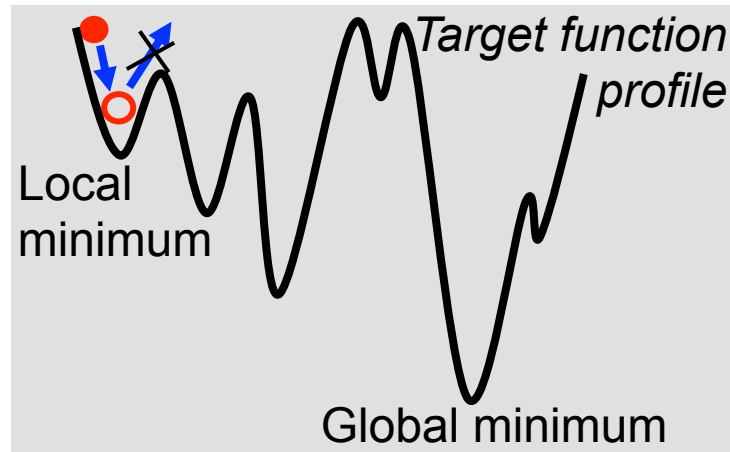
Structure refinement

1. **Model parameters**
2. **Optimization goal**
3. **Optimization method**

Refinement target optimization methods

▪ Gradient-driven minimization

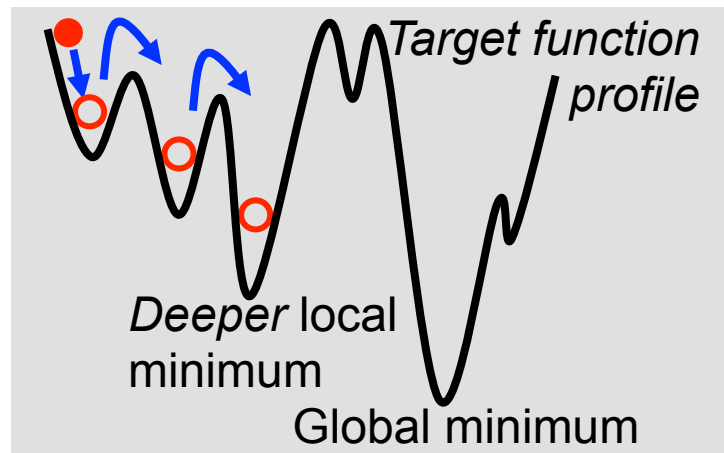
- Follows the local gradient.
- The target function depends on many parameters – many local minima.



Refinement target optimization methods

▪ Simulated annealing (SA)

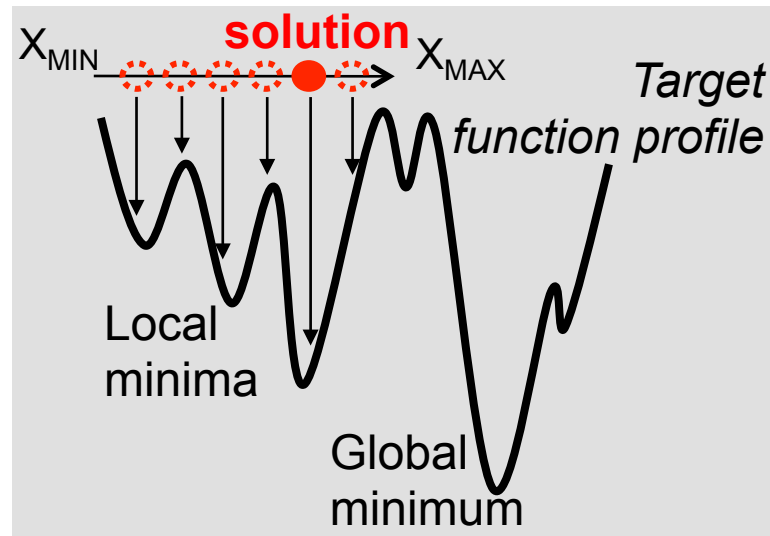
- SA is an optimization method which is good at escaping local minima.
- Annealing is a physical process where a solid is heated until all particles are in a liquid phase, followed by cooling which allows the particles to move to the lowest energy state.
- Simulated annealing is the simulation of the annealing process.
 - Increased probability of finding a better solution because motion against the gradient is allowed.
 - Probability of uphill motion is determined by the temperature.



Refinement target optimization methods

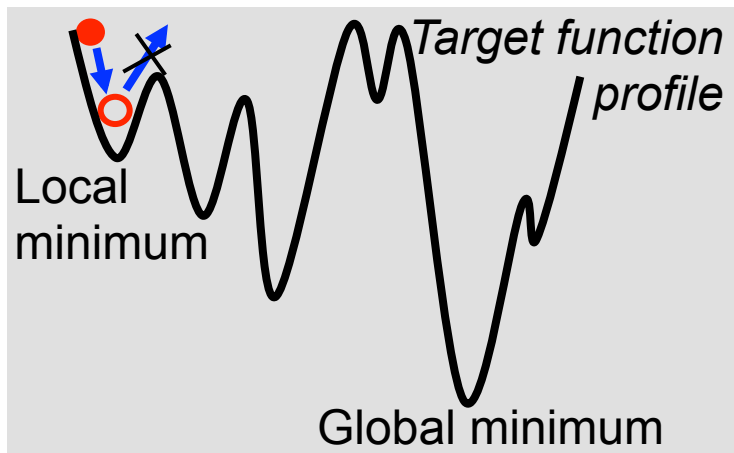
- **Grid search** (Sample parameter space within known range $[X_{\text{MIN}}, X_{\text{MAX}}]$)

Robust but may be time inefficient for many parameter systems, and not as accurate as gradient-driven. Good for small number of parameters (1-3 or so), and impractical for larger number of parameters.

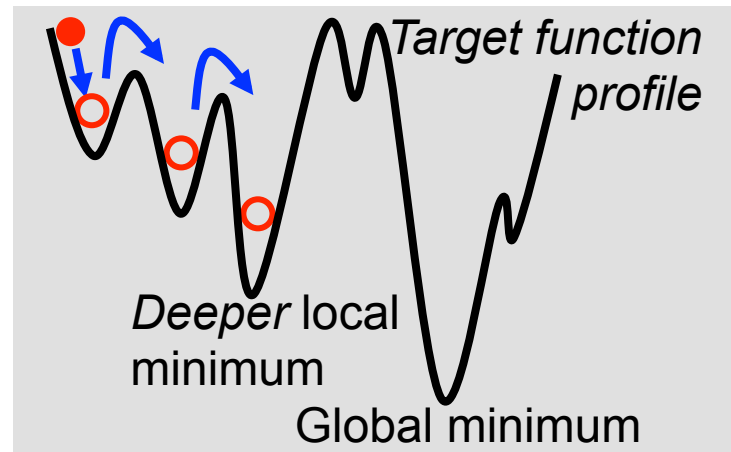


Refinement target optimization methods

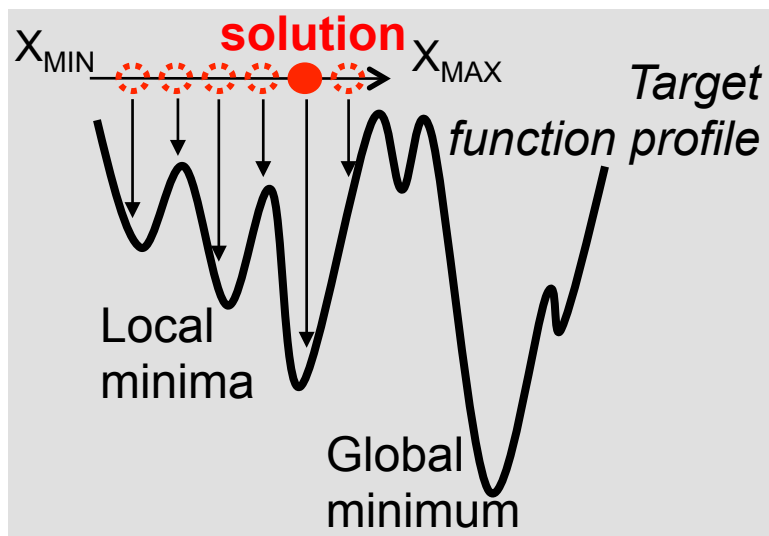
▪ Gradient-driven minimization



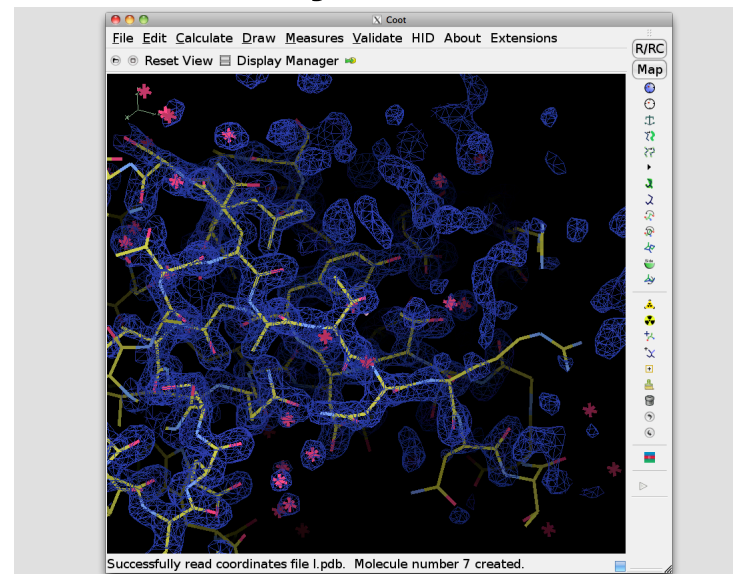
▪ Simulated annealing (SA)



▪ Grid search (Sample parameter space within known range $[X_{MIN}, X_{MAX}]$)

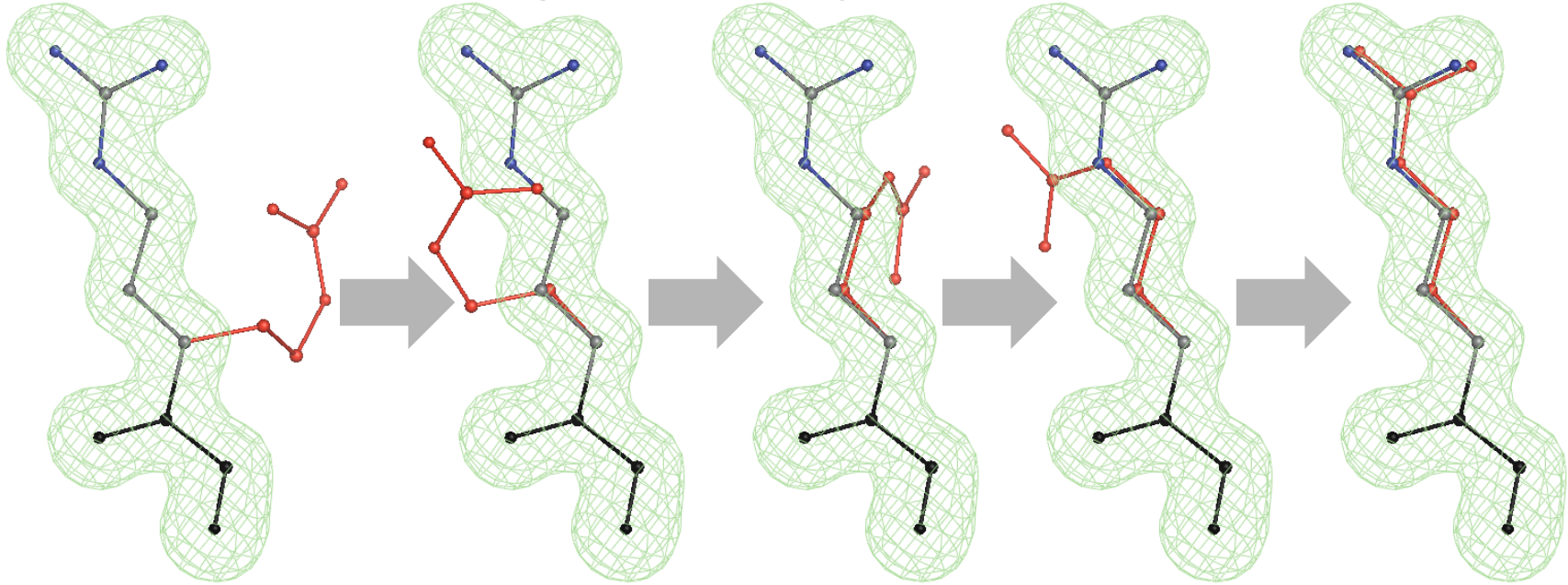


▪ Hands & eyes (Via Coot)



Grid search examples

- Real-space sampling to fit density



- Finding bulk-solvent k_{SOL} and B_{SOL}

– $k_{\text{SOL}} : [0.2, 0.6]$

– $B_{\text{SOL}} : [10, 100]$

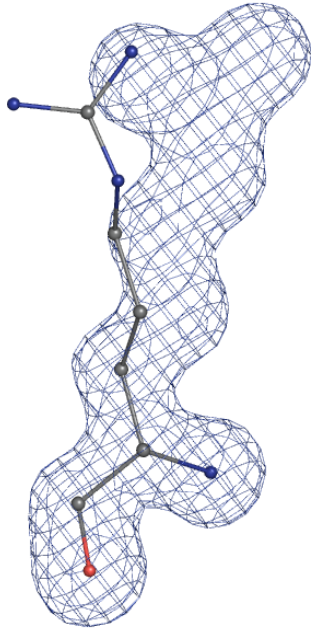
$$\mathbf{F}_{\text{MODEL}} = k_{\text{OVERALL}} e^{-sU_{\text{CRYSTAL}}} s^t \left(\mathbf{F}_{\text{CALC_ATOMS}} + k_{\text{SOL}} e^{-\frac{B_{\text{SOL}} s^2}{4}} \mathbf{F}_{\text{MASK}} \right)$$

- Twin fraction refinement

$$F_{\text{MODEL}} = |\mathbf{F}_{\text{MODEL}}| = k_{\text{OVERALL}} \sqrt{\alpha |\mathbf{F}_{\text{M}}(\mathbf{h})|^2 + (1 - \alpha)^2 |\mathbf{F}_{\text{M}}(\mathbf{T}\mathbf{h})|^2}$$

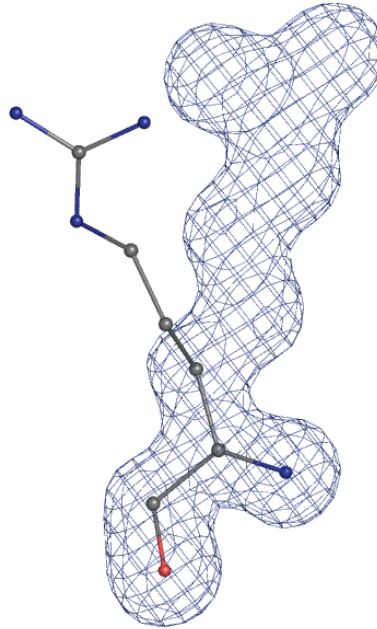
Refinement convergence

Minimization



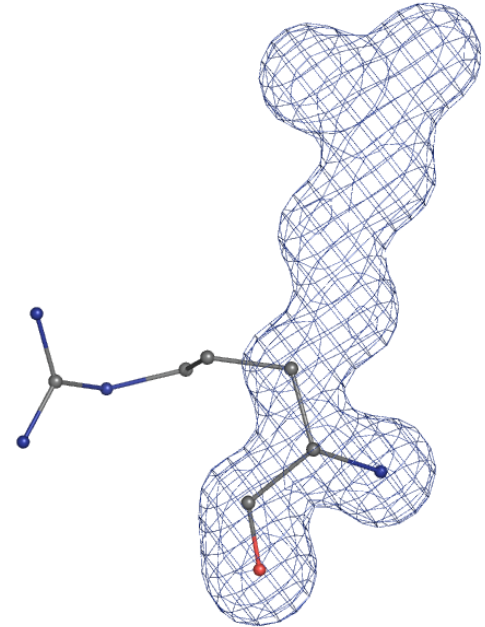
Both minimization and
SA can fix it

Simulated Annealing



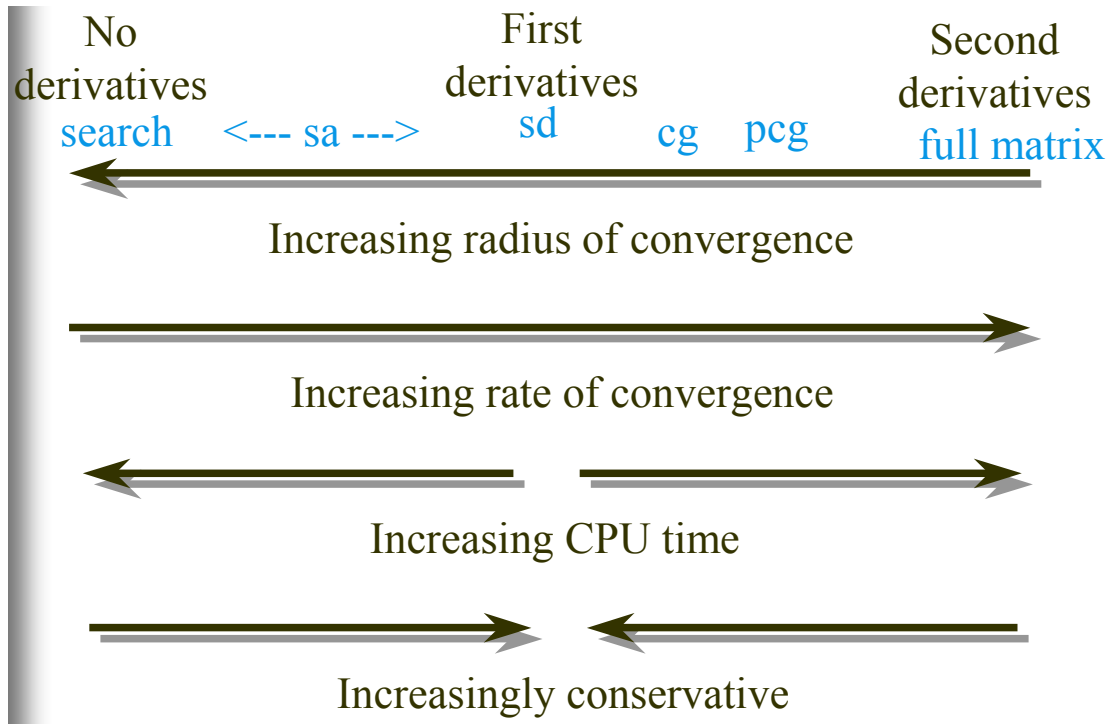
This is beyond the
convergence radius
for minimization

Real-space grid search



This is beyond the
convergence radius for
minimization and SA

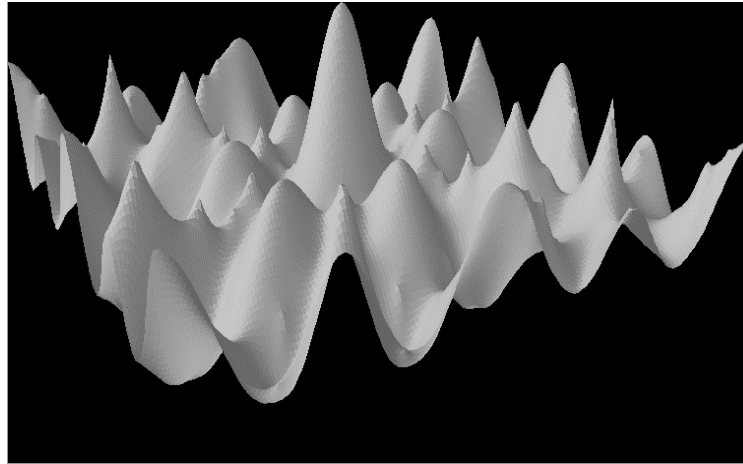
Summary on optimization tools



Picture stolen from Dale Tronrud

Refinement convergence

- Landscape of a refinement function is very complex

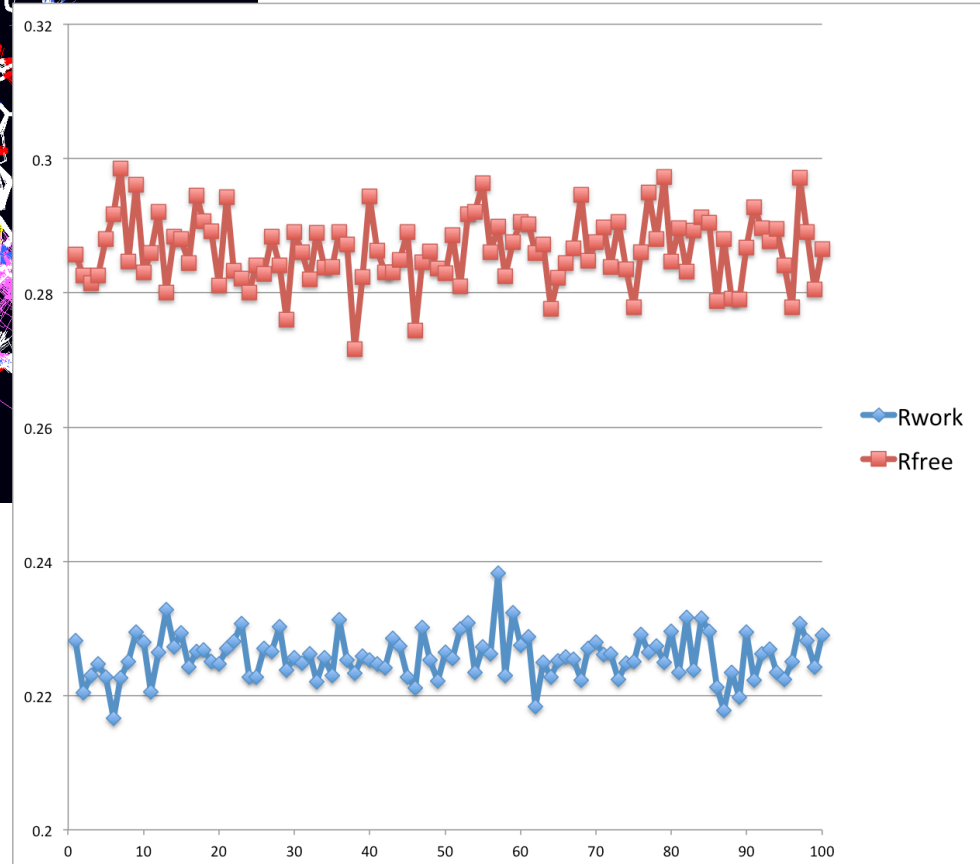
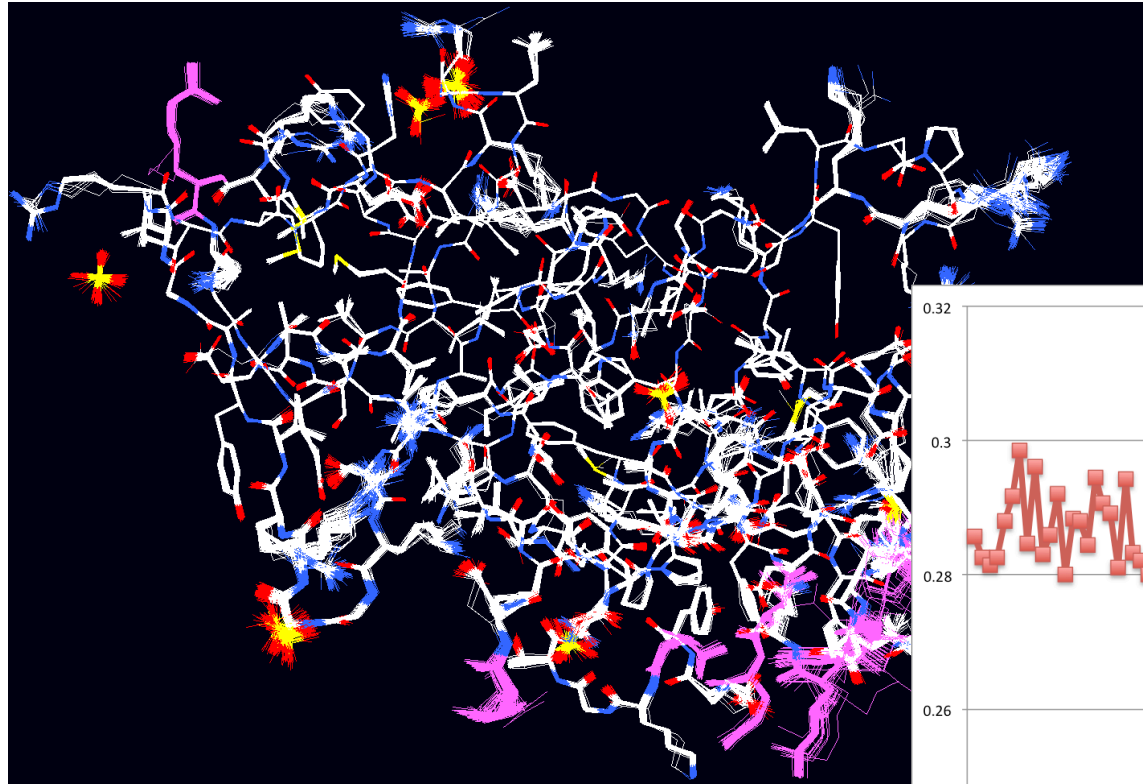


Picture stolen from Dale Tronrud

- Refinement programs have very small convergence radii compared to the size of the function profile
 - Depending where you start, the refinement engine will bring the structure to one of the closest local minimum
- What does it mean in practice ? Let's do the following experiment: run 100 identical Simulate Annealing refinement jobs, each starting with different random seed...

Refinement convergence

- As result we get an ensemble of slightly different structures having small deviations in atomic positions, B-factors, etc... R-factors deviate too.



Refinement convergence

- Interpretation of the ensemble:
 - The variation of the structures in the ensemble reflects:
 - Refinement artifacts (limited convergence radius and speed)
 - Some structural variations
 - Spread between the refined structures is the function of resolution (lower the resolution – higher the spread), and the differences between initial structures
 - Obtaining such ensemble is very useful in order to assess the degree of uncertainty that comes from refinement alone

Refinement summary

- **Model parameterization:**
 - quality of experimental data (resolution, completeness, ...)
 - quality of current model (initial with large errors, almost final, ...)
 - data-to-parameters ratio (restraints have to be accounted)
 - individual vs grouped parameters
 - knowledge based restraints/constraints (NCS, reference higher resolution model, etc...)
- **Refinement target:**
 - ML target is the option of choice for macromolecules
 - Real-space vs reciprocal space
 - Use experimental phase information if available
- **Optimization method:**
 - Choice depends on the size of the task, refinable parameters, desired convergence radius

Refinement - summary

▪ Refinement is:

- Process of changing model parameters to optimize a target function
- Various tricks are used (restraints, different model parameterizations) to compensate for imperfect experimental data

▪ Refinement is NOT :

- Getting a 'low enough' R-value (to satisfy supervisors or referees)
- Getting 'low enough' B-values (to satisfy supervisors or referees)
- Completing the sequence in the absence of density

Typical refinement steps

▪ **Input data and model processing:**

- Read in and process PDB file
- Read in and process library files (for non-standard molecules, ligands)
- Read in and process reflection data file
- Check correctness of input parameters
- Create objects that will be reused in refinement later on (geometry restraints,...)

▪ **Main refinement loop (macro-cycle; repeated several times):**

- Bulk solvent correction, anisotropic scaling, twinning parameters estimation
- Update ordered solvent (water) (add or remove)
- Target weights calculation
- Refinement of coordinates (rigid body, individual) (minimization or Simulated Annealing)
- ADP refinement (TLS, group, individual isotropic or anisotropic)
- Occupancy refinement (individual, group, constrained)

▪ **Output results:**

- PDB file with refined model
- Various maps (2mFo-DFc, mFo-DFc) in various formats (CNS, MTZ)
- Complete statistics
- Structure factors

Phenix
http://www.phenix-online.org/

Phenix NEW [Development release of PHENIX version 1.4 now available](#)
Python-based Hierarchical ENvironment for Integrated Xtallography

PHENIX is a new software suite for the automated determination of macromolecular structures using X-ray crystallography and other methods.

Citing PHENIX:
PHENIX: building new software for automated crystallographic structure determination P.D. Adams, R.W. Grosse-Kunstleve, L.-W. Hung, T.R. Ioerger, A.J. McCoy, N.W. Moriarty, R.J. Read, J.C. Sacchettini, N.K. Sauter and T.C. Terwilliger. *Acta Cryst.* D58, 1948-1954 (2002)

Download the latest development release (1.4-3) [First request download password]

Help: [FAQ](#) [Mailing List Subscription](#) [List Archives](#) [Report a Bug](#) [Email for Help](#)

Using PHENIX (release 1.4-3): [Full Documentation](#) [PDF](#)

- Assessing data quality with [phenix.xtriage](#)
- Automated structure solution with [AutoSol](#)
- Automated molecular replacement with [AutoMR](#)
- Automated model building and rebuilding with [AutoBuild](#)
- Automated ligand fitting with [LigandFit](#)
- Structure refinement with [phenix.refine](#)
- Generation of ligand coordinates and restraints with [elbow](#)
- The [PHENIX Graphical User Interface](#)

[Documentation for 1.3-final](#)





The PHENIX system also includes SOLVE/RESOLVE, Phaser, Textal, the CCI Applications (phenix.xtriage, phenix.refine, elbow and many more), components from Molprobit, and the Computational Crystallography Toolbox in a Python framework.

Funding for PHENIX: [Protein Structure Initiative \(NIH General Medical Sciences\)](#)

The PHENIX Industrial Consortium

For-profit groups can obtain access to PHENIX through a Consortium agreement. This provides a license to use PHENIX and research funds to develop new features in PHENIX tailored to the needs of commercial users.

Groups developing PHENIX:

Paul Adams	Randy Read	Jane & Dave Richardson	Tom Terwilliger	Tom Ioerger & Jim Sacchettini
				

[Privacy and Security Notice](#) [About this website](#)

Introduction to PHENIX
Using PHENIX
Platforms
Licensing
Download
Recent Changes
Publications
Presentations
Computational Crystallography Toolbox
Contact Us
The PHENIX Team
Acknowledgments
Intranet

Information
Members
Download
Contact Us

This presentation (PDF file) and much more