Fourier map modification by Maximum Entropy Method (MEM) and its implementation in Phenix

Pavel Afonine, Vladimir Lunin

24-JAN-2013

Objectives

- Unveil mystery about Maxim Entropy Method as it's applied in Crystallography
  - Answer questions "what it is for?" "why?" "what to expect?"
- Describe MEM algorithm and its implementation in Phenix (phenix.maximum_entropy_map)
- Show examples

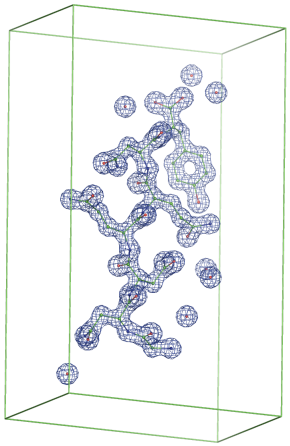
Crystallographic structure determination as an example of inverse problem

- **Inverse problem** is a task of converting observed measurements into information about a physical object. Associated framework provides methods to overcome problems due to ill-behaved tasks.

- **Crystallography context**

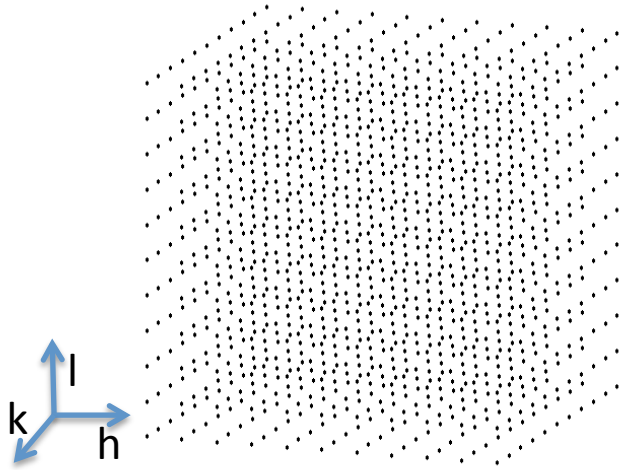
  - We want to study electron density distribution in unit cell

$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$



  - Experimentally we obtain structure factors

$$F(\mathbf{s}) = \int_{V_{cell}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{sr}) dV$$



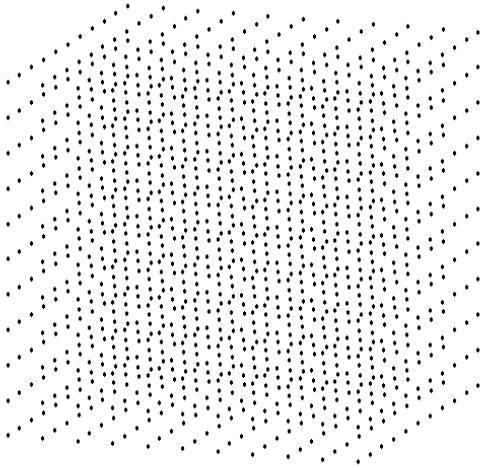
Inverse problems

- **Exact correspondence between *ρ* and *F* is only when all terms in the summation are present**

$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr}) \quad \Longleftrightarrow \quad F(\mathbf{s}) = \int_{V_{cell}} \rho(\mathbf{r}) \exp(2\pi i \mathbf{sr}) dV$$

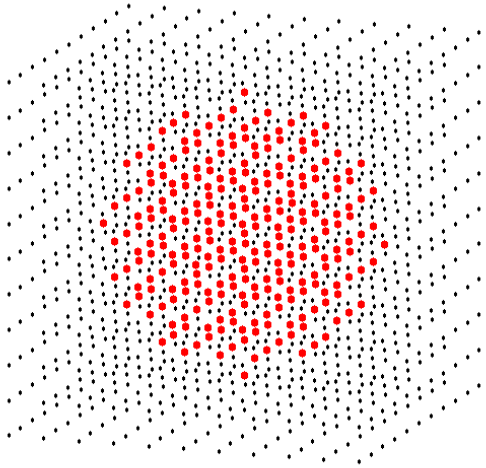
- **In reality only a subset of all *F* is measured**

All reflections



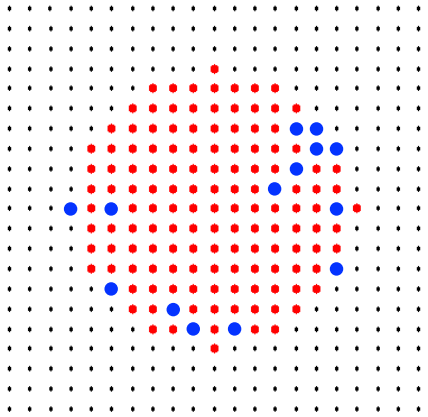
Infinite number

Measured reflections (red)



Reflections in sphere R=1/d_{min}
d_{min} - highest resolution of dataset

Measured reflections: 2D slice



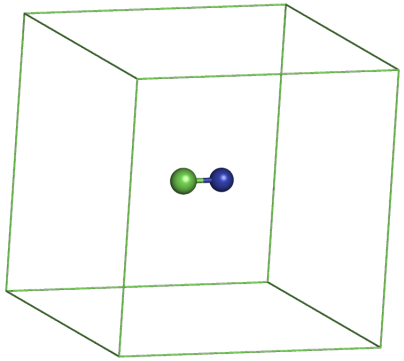
Some reflections in sphere R=1/d_{min} may be missing (blue): incomplete dataset

- **Incomplete hkl set means density is not accurate anymore: Fourier image of finite resolution**

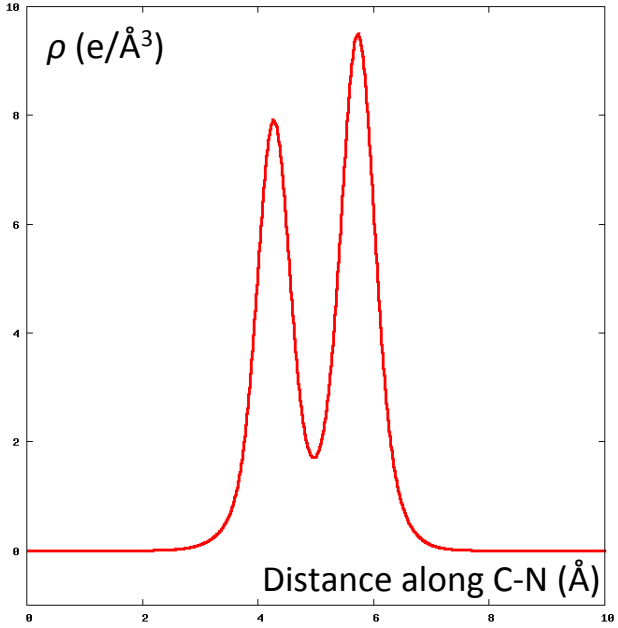
$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr}) \quad \Longrightarrow \quad \rho_{image}(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h_{min}}^{h_{max}} \sum_{k_{min}}^{k_{max}} \sum_{l_{min}}^{l_{max}} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$

Inverse problems

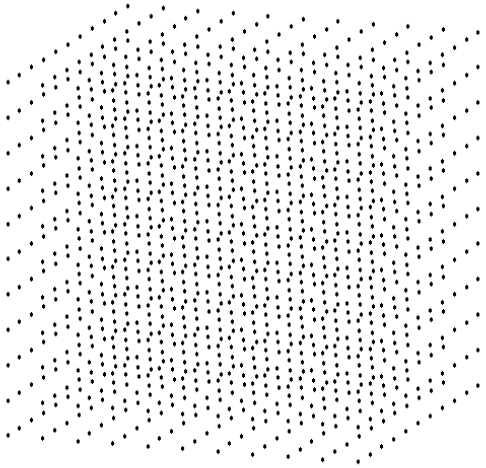
Toy example: C-N in 10*10*10Å P1 box



Electron density distribution along C-N bond vector

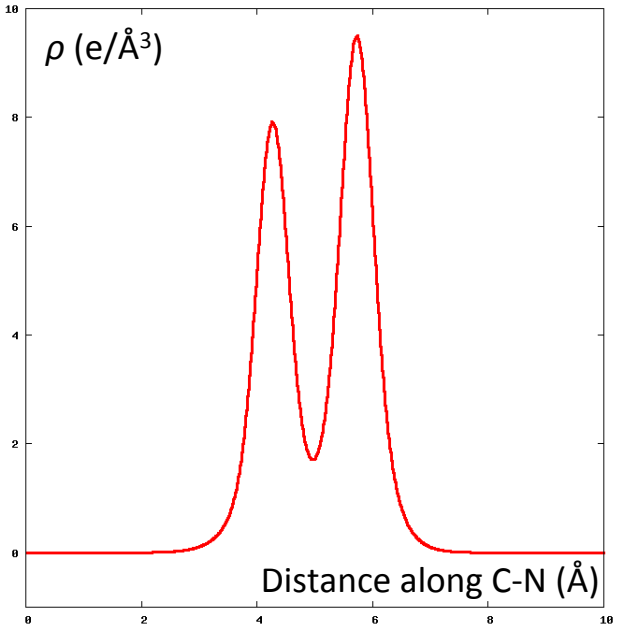


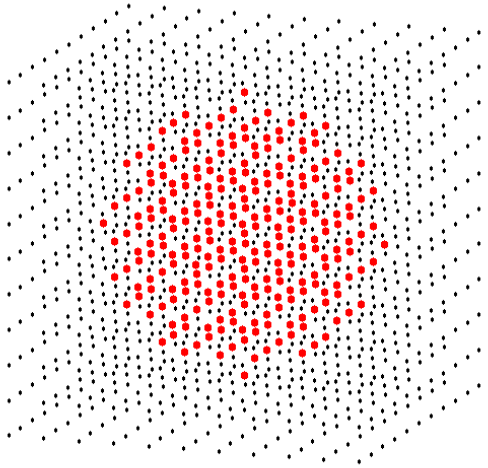
Inverse problems



Exact density

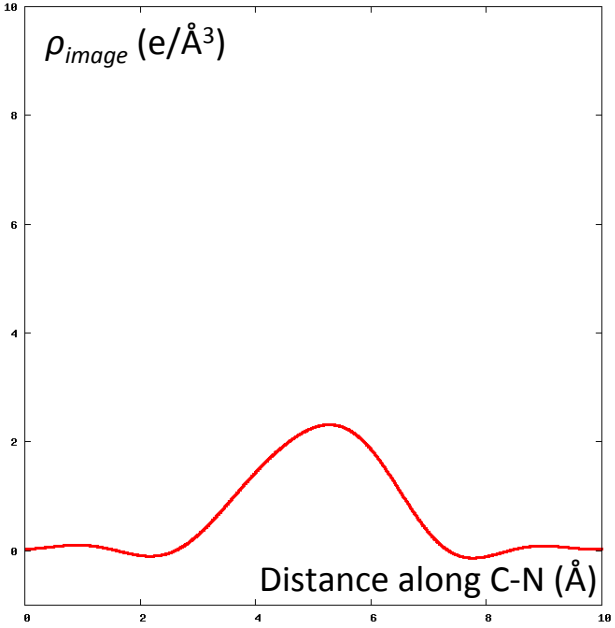
$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$





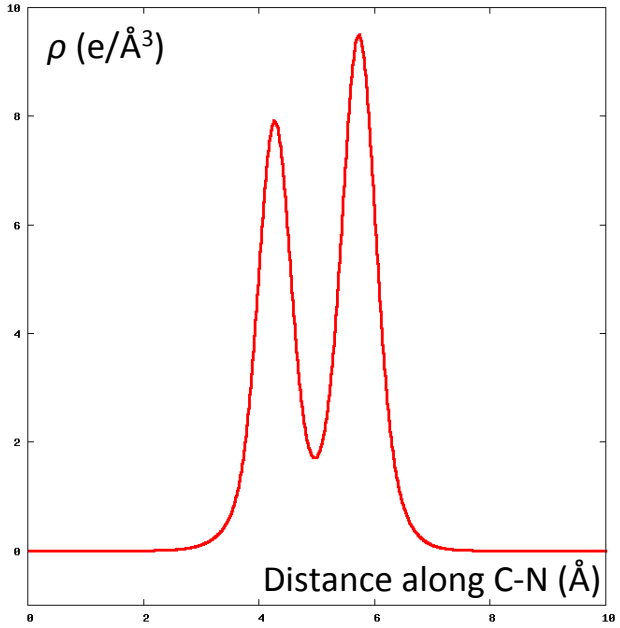
2 Å resolution Fourier image

$$\rho_{image}(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h_{min}}^{h_{max}} \sum_{k_{min}}^{k_{max}} \sum_{l_{min}}^{l_{max}} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$

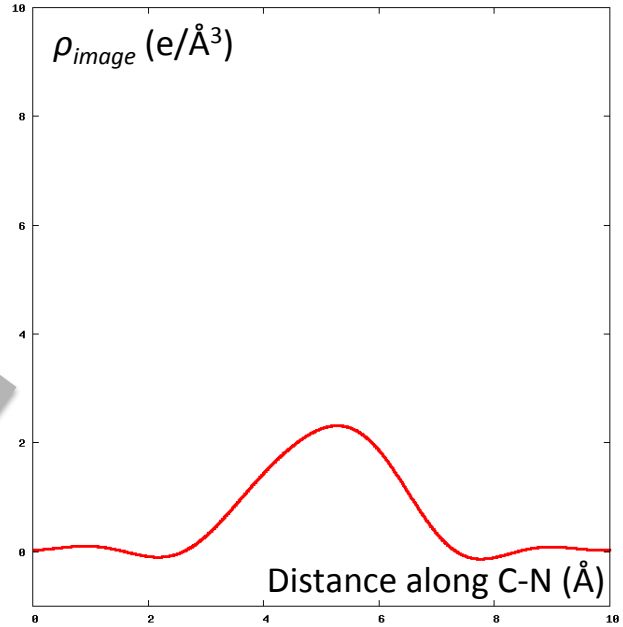


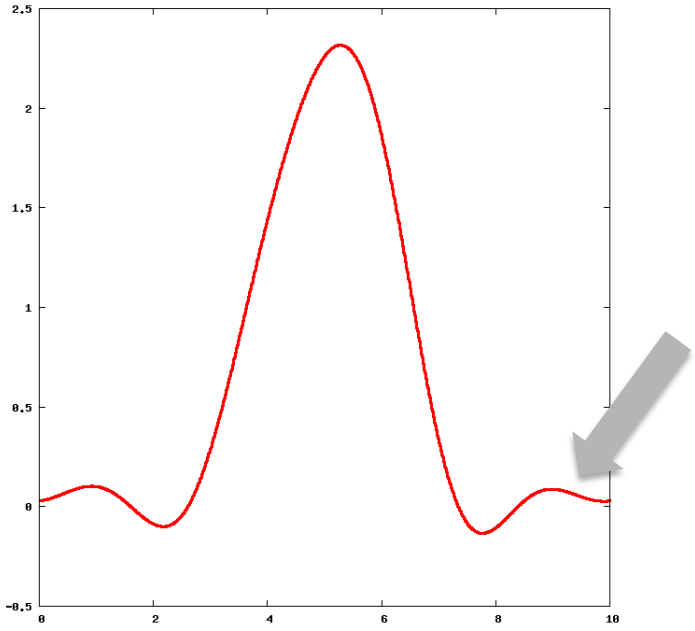
Inverse problems

Exact density



2 Å resolution Fourier image

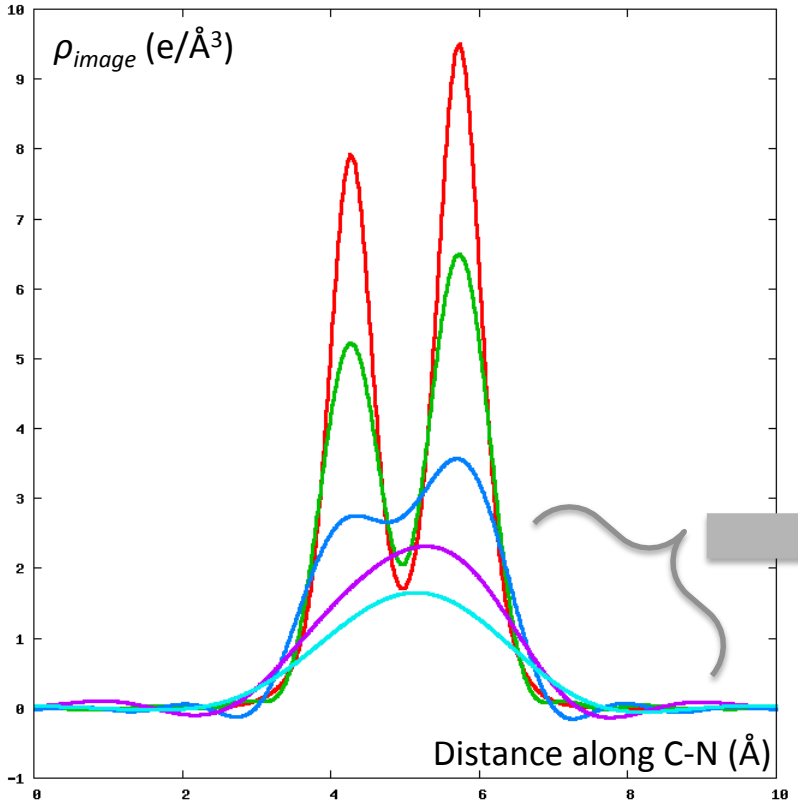


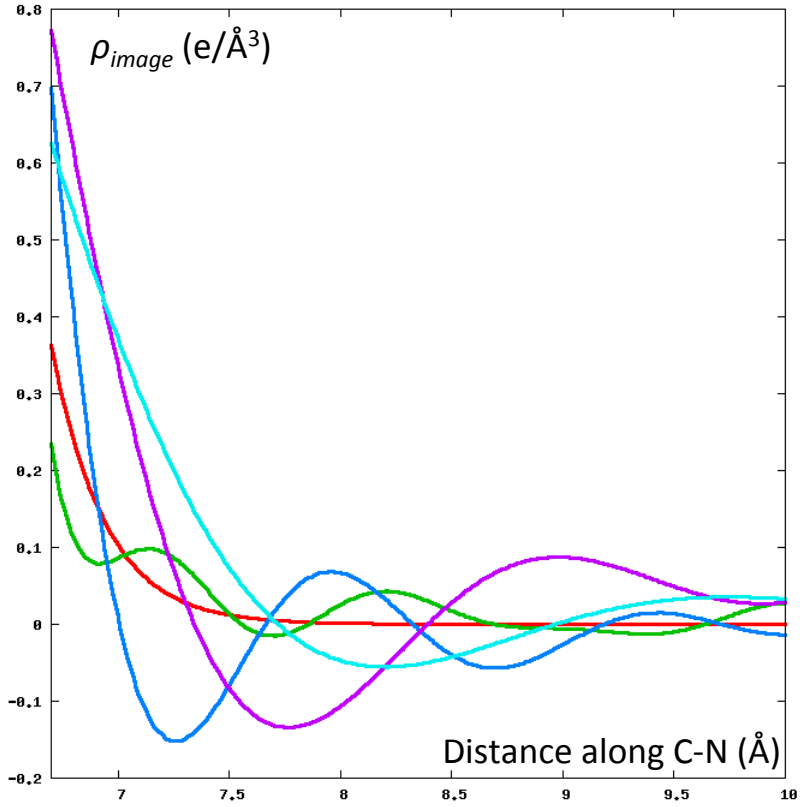


Positive and negative spurious peaks – Fourier truncation ripples (artifacts)

Inverse problems

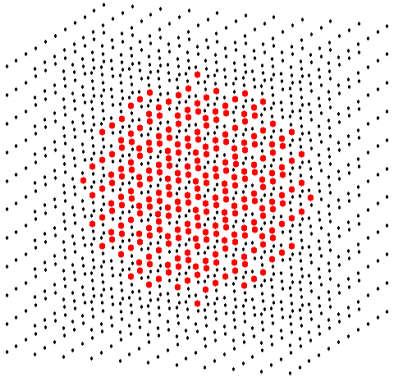
More examples: exact density (red) and 1, 1.5, 2 and 2.5 Å resolution Fourier images





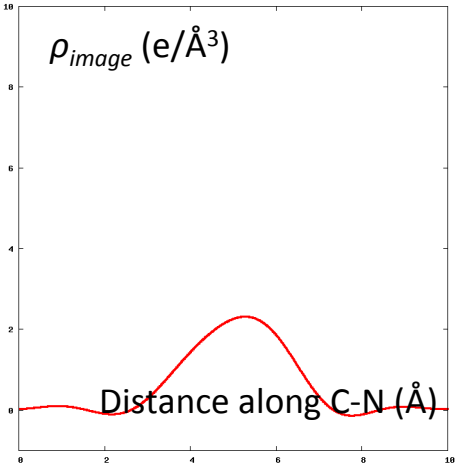
Inverse problems

2 Å resolution set of F(s)



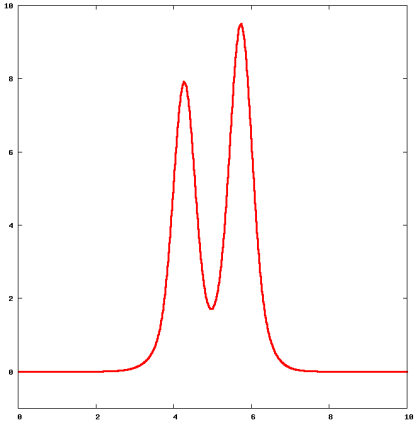
$$\rho_{image}(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h_{min}}^{h_{max}} \sum_{k_{min}}^{k_{max}} \sum_{l_{min}}^{l_{max}} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$

2 Å resolution Fourier image



Exact density

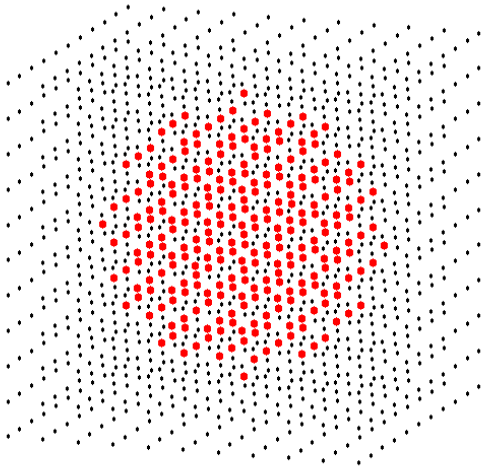
$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{sr})$$

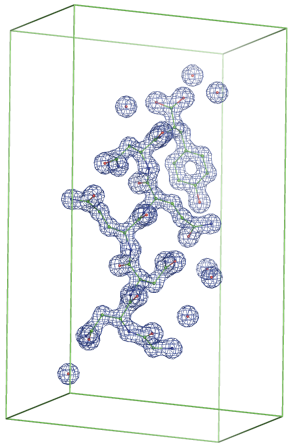


- Ill-posed problem in crystallography: we want to reconstruct image damaged due to finite amount of measured data
- Regularization involves introducing additional information in order to solve an ill-posed problem
- It involves encoding prior knowledge in terms of constraints on the solution space like positivity or smoothness for example.
- Example: density is positive and total charge F000

Inverse problems

$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{s}\mathbf{r})$$

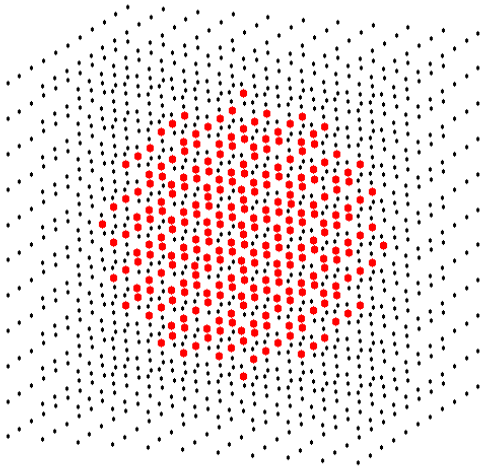




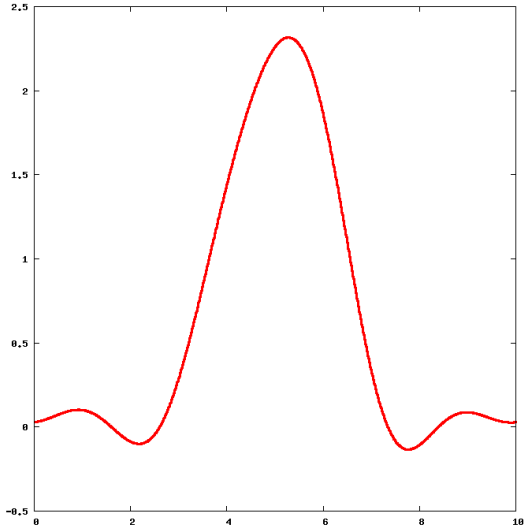
$$F(\mathbf{s}) = \frac{V_{cell}}{N_x N_y N_z} \sum_{j_x=0}^{N_x-1} \sum_{j_y=0}^{N_y-1} \sum_{j_z=0}^{N_z-1} \rho(j_x, j_y, j_z) \exp\left(2\pi i \left[h j_x + k j_y + l j_z\right]\right)$$ (Sayre, 1951)

- One can iterate back and forth any number of times – this will not change *F(**s**)* and *ρ(**s**)*
- Values of F calculated for Miller indices that were not used in calculation of *ρ(**s**)* will always be zero.

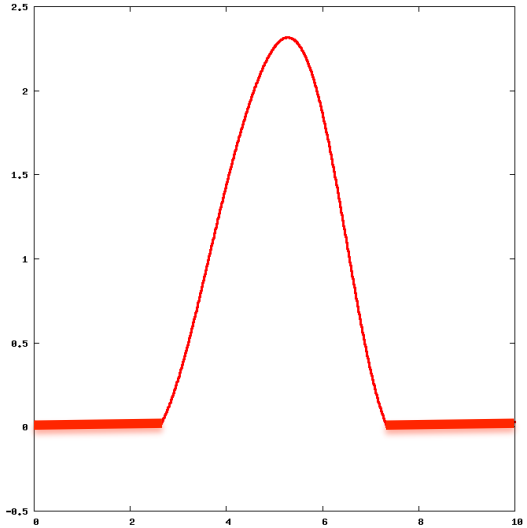
Inverse problems



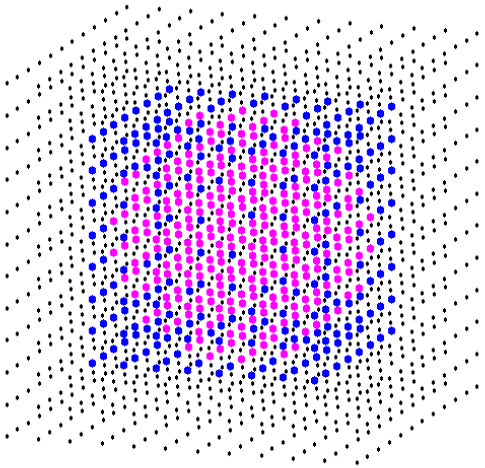
$$\rho(\mathbf{r}) = \frac{1}{V_{cell}} \sum_{h=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} \sum_{l=-\infty}^{+\infty} F(\mathbf{s}) \exp(-2\pi i \mathbf{s}\mathbf{r})$$



Modify density in some way



$$F(\mathbf{s}) = \frac{V_{cell}}{N_x N_y N_z} \sum_{j_x=0}^{N_x-1} \sum_{j_y=0}^{N_y-1} \sum_{j_z=0}^{N_z-1} \rho(j_x, j_y, j_z) \exp\left(2\pi i \left[h j_x + k j_y + l j_z\right]\right.$$



- This is a foundation for collection of regularization methods that in crystallography called Density Modification

Inverse problems

- There are many ways to modify density:

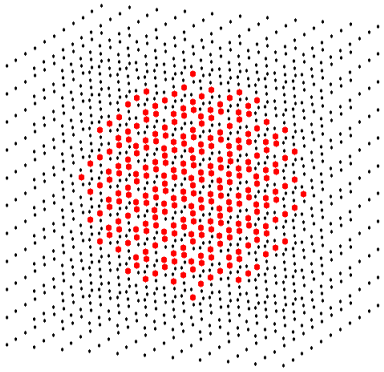
  1. Atomicity (Hoppe & Gassmann, 1964)
  2. Positivity (Barrett & Zwick, 1971)
  3. Noncrystallographic symmetry (Bricogne, 1974)
  4. Solvent flatness (Bricogne, 1974)
  5. Map connectivity (continuity) (Bhat & Blow, 1982)
  6. Histogram matching (Lunin, 1988)
  7. MEM (maximum entropy methods) (Collins, 1982)

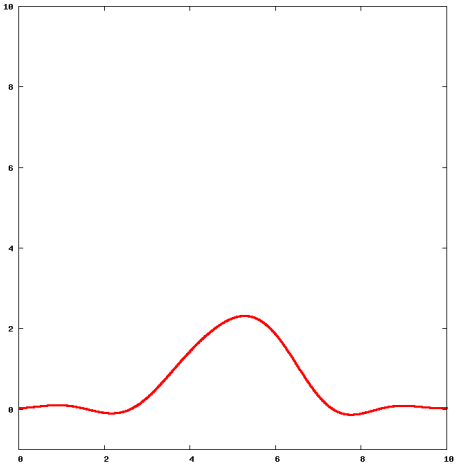
- Good reviews: Podjarny, Rees & Urzhumtsev, 1996; Cowtan, 2012; Trwilliger (for statistical Density Modification)

- Purpose: improve density by improving phases and extending the resolution

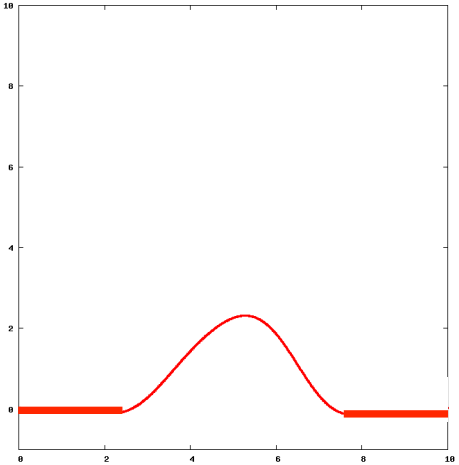
- **Positivity – one of the least model committal constraint (less chances to introduce model bias) and quick and easy to apply.**

Inverse problems

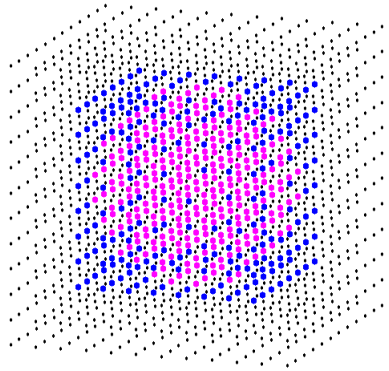




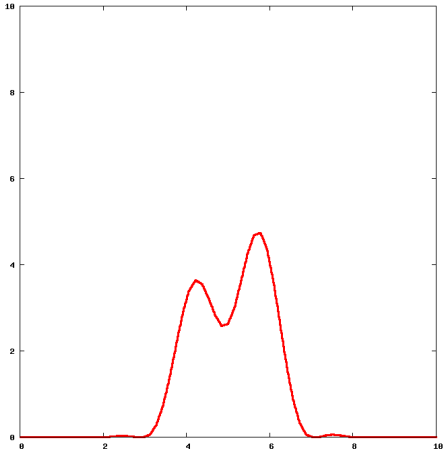
Original map



Modified map

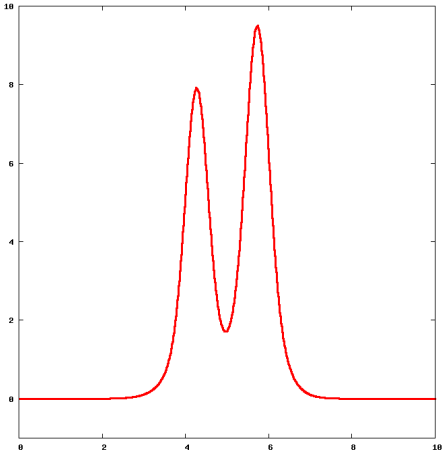


New set of reflections



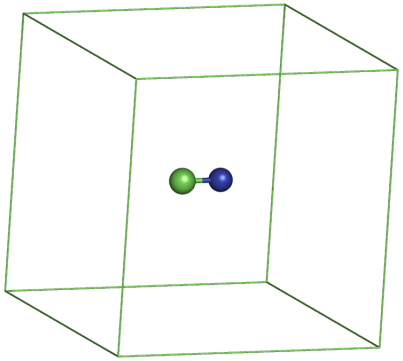
Improved map

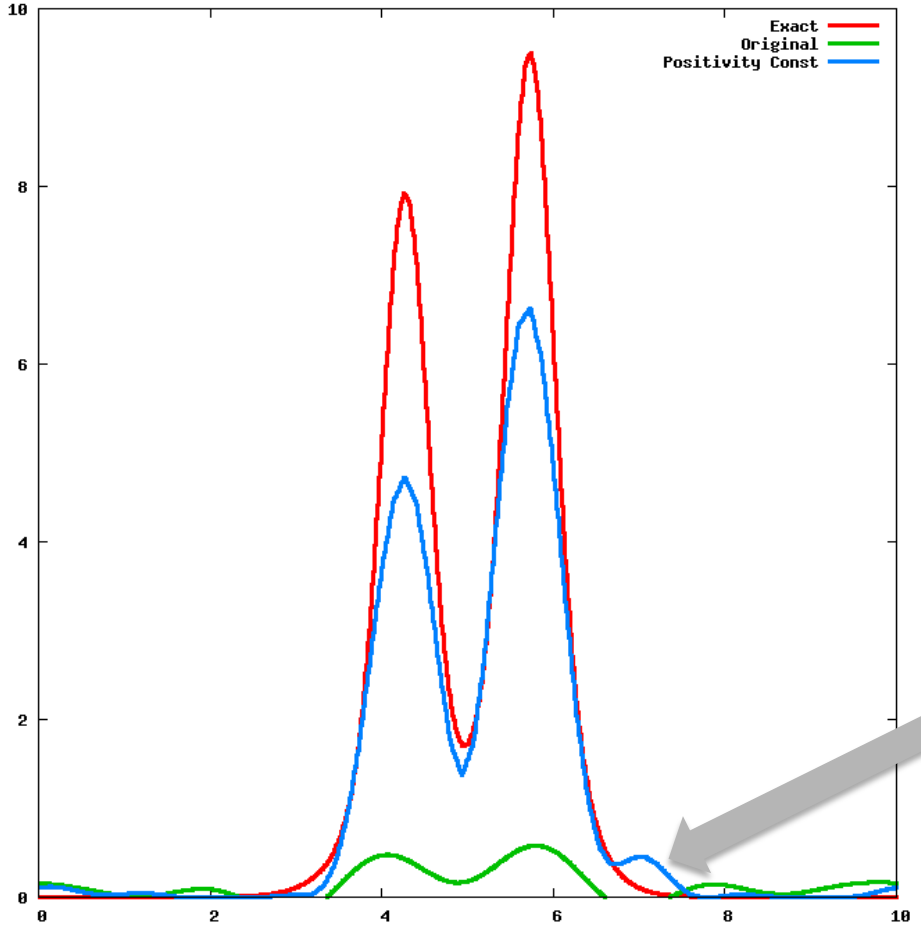
This is may still be far from ideal density



Positivity constraint: example

Example: C-N in 10*10*10Å P1 box





Original data (1.5 Å) is highly incomplete and has 70 degrees phase error

Problem: may grow spurious peaks.

Maximum Entropy map modification

- Originates from information theory (Jaynes, 1957)
- Useful publications:
  Collins, 1982
  Gull & Daniel, 1978
  Wu, 1984

- No paper that provide details enough for painless implementation!

- Starting from a positive electron density distribution (which can be flat) we want to "grow" such a new density distribution:
  a) that is as flat as possible (= has highest, maximum, entropy), and
  b) the structure factors calculated from this new distribution are close enough to original data within prescribed tolerance

- There are many admissible maps: the maps that are close to the original data within the tolerance

- During the process entropy drops, and NOT increase. This is counterintuitive as the method is called maximum entropy

- Method is called maximum entropy because we choose that map (among all admissible ones) that is the flattest one (has highest entropy)

Maximum Entropy map modification: basic definitions

- Find grid function $\{\rho_j\}_{j=0}^{N-1}$ such that maximizes residual $H(\rho) - \frac{\lambda}{2} Q_X(\rho) \to max, \lambda > 0$ and $\rho_j \geq 0 \; \forall j, \sum_{j=0}^{N-1} \rho_j = 1$.
  - $H(\rho) = -\sum_{j=0}^{N-1} \rho_j \ln(\rho_j)$ - entropy
  - $Q_X(\rho) = \sum_{k \in K} w_k \left| \mathbf{F}_k - \mathbf{F}_k^{obs} \right|^2$ - constraint term
  - $\mathbf{F}_k(\rho) = \frac{1}{N} \sum_{j=0}^{N-1} \rho_j \exp\left[2\pi i \frac{jk}{N}\right]$ - complex inverse discrete Fourier transform
  - $\{\mathbf{F}_k^{obs}\}_{k \in K}$ - defined complex numbers, K – given set of reflections
  - $\mathbf{F}_{-k}^{obs} = \overline{\mathbf{F}_k^{obs}}$ (have Hermitian symmetry)
  - Assume: $\rho \ln(\rho) = 0$ if $\rho = 0$, $w_k = \begin{cases} 1 & \text{for } k \in K \\ 0 & \text{otherwise} \end{cases}$

Maximum Entropy map modification: basic definitions

- Find grid function $\{\rho_j\}_{j=0}^{N-1}$ such that maximizes residual $H(\rho) - \frac{\lambda}{2} Q_X(\rho) \to max, \lambda > 0$ and $\rho_j \geq 0 \; \forall j, \sum_{j=0}^{N-1} \rho_j = 1$.
  - $H(\rho) = -\sum_{j=0}^{N-1} \rho_j \ln(\rho_j)$ - entropy
  - $Q_X(\rho) = \sum_{k \in K} w_k \left| \mathbf{F}_k - \mathbf{F}_k^{obs} \right|^2$ - constraint term

Remark #1: for obscure to me reasons, most of papers uses $KL(\rho) = -\sum_{j=0}^{N-1} \rho_j \ln\left(\frac{\rho_j}{\tau_j}\right)$ and call it entropy, while *KL* is Kullback-Leibler divergence (or cross entropy)

Remark #2: I was not able to find the meaning and purpose of τ in crystallographic context, and why sometimes *KL* is used over *H*, and vice versa.

Remark #3: Functions such as –ln(x) or even $-u^{1/2}$ have similar properties as *H(x).* Function $-u^{1/2}$ have no information theory justification! Using any of these functions make no different on final result (Narayan & Nityananda, 1986).

Remark #4: Constraint term Q can be more complex, such as weighted sum of phased and phaseless reflections. I may also include other information such as symmetry, solvent/ macromolecule mask and other *a priori* known information.

Remark #5: This is very similar to crystallographic model refinement where the optimizing target is $T_{data} + w*T_{restraints} \to min$.

Maximum Entropy map modification: basics

- Maximization $H(\rho) - \frac{\lambda}{2} Q_X(\rho) \to max$ is achieved by solving *N* non-linear equations w.r.t. *ρ:*

$$\frac{d}{d\rho}\left(H(\rho) - \frac{\lambda}{2} Q_X(\rho)\right) = 0$$

N – number of grid points

$H(\rho) = -\sum_{j=0}^{N-1} \rho_j \ln \rho_j$ – entropy

$Q_X(\rho) = \sum_{k \in K} w_k \left|\mathbf{F}_k - \mathbf{F}_k^{obs}\right|^2$ – constraint term

- Evaluating derivatives results in *N* equations

$$\rho = A\, exp[\lambda(\rho - \rho_{obs})]$$

that are solved using iterative procedure

Derivatives of $Q_X(\rho)$ w.r.t. ρ

$$\frac{d}{d\rho} Q_X(\rho) = ?$$

$$Q_X = \sum_{h=0}^{N-1} w_h \left| \mathbf{F}_h - \mathbf{F}_h^o \right|^2$$

$$\rho = \left\{ \rho_j \right\}_{j=0}^{N-1} \qquad \mathbf{F}_h = \frac{1}{N} \sum_{j=0}^{N-1} \rho_j \exp\left[2\pi i \frac{jh}{N} \right] \quad , \quad h = 0,\ldots,N-1 \qquad \rho_j = \sum_{h=0}^{N-1} \mathbf{F}_h \exp\left[-2\pi i \frac{jh}{N} \right] \quad , \quad j = 0,\ldots,N-1$$

$$\left| \mathbf{F}_h - \mathbf{F}_h^o \right|^2 = \mathbf{F}_h \overline{\mathbf{F}_h} - \overline{\mathbf{F}_h^o} \mathbf{F}_h - \mathbf{F}_h^o \overline{\mathbf{F}_h} + \mathbf{F}_h^o \overline{\mathbf{F}_h^o}$$

$$\frac{\partial}{\partial \rho_j} \left| \mathbf{F}_h - \mathbf{F}_h^o \right|^2 = \left(\frac{\partial}{\partial \rho_j} \mathbf{F}_h \right) \overline{\mathbf{F}_h} + \mathbf{F}_h \frac{\partial}{\partial \rho_j} \overline{\mathbf{F}_h} - \overline{\mathbf{F}_h^o} \frac{\partial}{\partial \rho_j} \mathbf{F}_h - \mathbf{F}_h^o \frac{\partial}{\partial \rho_j} \overline{\mathbf{F}_h}$$

$$\frac{\partial}{\partial \rho_j} \mathbf{F}_h = \frac{1}{N} \exp\left[2\pi i \frac{jh}{N} \right]$$

$$\frac{\partial}{\partial \rho_j} \overline{\mathbf{F}_h} = \frac{1}{N} \exp\left[-2\pi i \frac{jh}{N} \right]$$

$$\frac{\partial}{\partial \rho_j} Q_X(\rho) = \frac{1}{N} \sum_{h=0}^{N-1} w_h \left(\mathbf{F}_h - \mathbf{F}_h^o \right) \exp\left[-2\pi i \frac{jh}{N} \right] + \overline{\frac{1}{N} \sum_{h=0}^{N-1} w_h \left(\mathbf{F}_h - \mathbf{F}_h^o \right) \exp\left[-2\pi i \frac{jh}{N} \right]}$$

Since Fourier coefficients have Hermitian symmetry result of summation are real numbers that invariant under complex conjugation operation:

$$\frac{\partial}{\partial \rho_j} Q_X(\rho) = \frac{2}{N} \sum_{h=0}^{N-1} w_h \left(\mathbf{F}_h - \mathbf{F}_h^o \right) \exp\left[-2\pi i \frac{jh}{N} \right] \qquad \rho_j^o = \sum_{h=0}^{N-1} w_h \mathbf{F}_h^o \exp\left[-2\pi i \frac{jh}{N} \right] \qquad \sum_{h=0}^{N-1} w_h \mathbf{F}_h \exp\left[-2\pi i \frac{jh}{N} \right] = \left(T * \rho \right)_j$$

$$\frac{\partial}{\partial \rho} Q_X(\rho) = \frac{2}{N} \left(T * \rho - \rho^o \right)$$

Here: T is interferention function and $\rho^o = \rho_{obs}$

Calculation algorithm (inputs)

Inputs

- Structure factors $\left\{\tilde{\mathbf{F}}_k^{obs}\right\}_{k\in K}$
- Total charge $\tilde{F}_0$
- Gridding N (N1, N2, N3)
- Scale C_{obs} that brings $\left\{\tilde{\mathbf{F}}_k^{obs}\right\}_{k\in K}$ onto absolute scale (C_{obs}=1 if $\left\{\tilde{\mathbf{F}}_k^{obs}\right\}_{k\in K}$ on absolute scale)

Remark #1: Input structure factors can be phased or phaseless F_{obs}, $2mF_{obs}$-DF_{model} map coefficients and other intrinsically positive maps.

Remark #2: MEM can be applied to residual (example: mF_{obs}-DF_{model}) maps. For this two sets of coefficients are calculated first mF_{obs} and DF_{model} , then corresponding maps are subject to MEM procedure. The difference between two MEM modified maps is the desired residual map.

Remark #3: If amplitudes of input structure factors are measured data (F_{obs}) and experimental uncertainties (σ) are available, then residual Q_X is simply χ^2, and its statistical properties can be used to choose appropriate weight λ and determine convergence.

Calculation algorithm (parameters)

Parameters

- Weight λ, which defines how much we allow deviate new map coefficients from $\left\{\tilde{\mathbf{F}}_k^{obs}\right\}_{k\in K}$
- Memory coefficients β (0.1-1.0), which insures convergence by preventing oscillations
- Number of iterations

Calculation algorithm (optimization procedure - I)

Optimization procedure

1. Prepare inputs
   - Scale input data $\mathbf{F}_k^{obs} = \frac{C_{obs}}{NF_0}\tilde{\mathbf{F}}_k^{obs}$
   - Compute synthesis $\rho_j^{obs} = \sum_{k\in K} w_k \mathbf{F}_k^{obs} \exp\left[-2\pi i \frac{jk}{N}\right]$
   - Define $A_{GD}=1/N$, N – number of grid points (used to enforce Z (total charge) = 1)
2. Obtain initial approximation. Any of three works and does not change the outcome:
   - $\rho_j^{(0)} = \frac{\rho_n^{obs} - \rho_{\min}^{obs}}{\sum_n \left(\rho_n^{obs} - \rho_{\min}^{obs}\right)}$
   - Flat map: set all grid points to a constant value
   - LDE (replace negative density with some small values)

Calculation algorithm (optimization procedure – II)

Optimization procedure

3. Iterations, starting from $\rho^{(0)}$, at (n+1) step given $\rho^{(n)}$ from previous step:

   - Compute

$$\mathbf{F}_k^{(n)} = \frac{1}{N}\sum_{j=0}^{N-1}\rho_j^{(n)}\exp\left[2\pi i\frac{jk}{N}\right] \qquad \rho_j^{\mathrm{mod}} = \sum_{k\in K} w_k \mathbf{F}_k^{(n)}\exp\left[-2\pi i\frac{jk}{N}\right] \qquad \Delta_j = \rho_j^{\mathrm{mod}} - \rho_j^{obs}$$

If $\Delta_j \geq 0$:
$$\tilde{\rho}_j = \left(1+\frac{\lambda}{N}\rho_j^{(n)}\right)\frac{A_{GD}\exp\left[-\lambda\Delta_j/N\right]}{1+\frac{\lambda}{N}A_{GD}\exp\left[-\lambda\Delta_j/N\right]}$$

If $\Delta_j < 0$:
$$\tilde{\rho}_j = \left(1+\frac{\lambda}{N}\rho_j^{(n)}\right)\frac{A_{GD}}{\exp\left[\lambda\Delta_j/N\right]+\frac{\lambda}{N}A_{GD}}$$

$$Z^{(n+1)} = \sum_{j=0}^{N-1}\tilde{\rho}_j$$

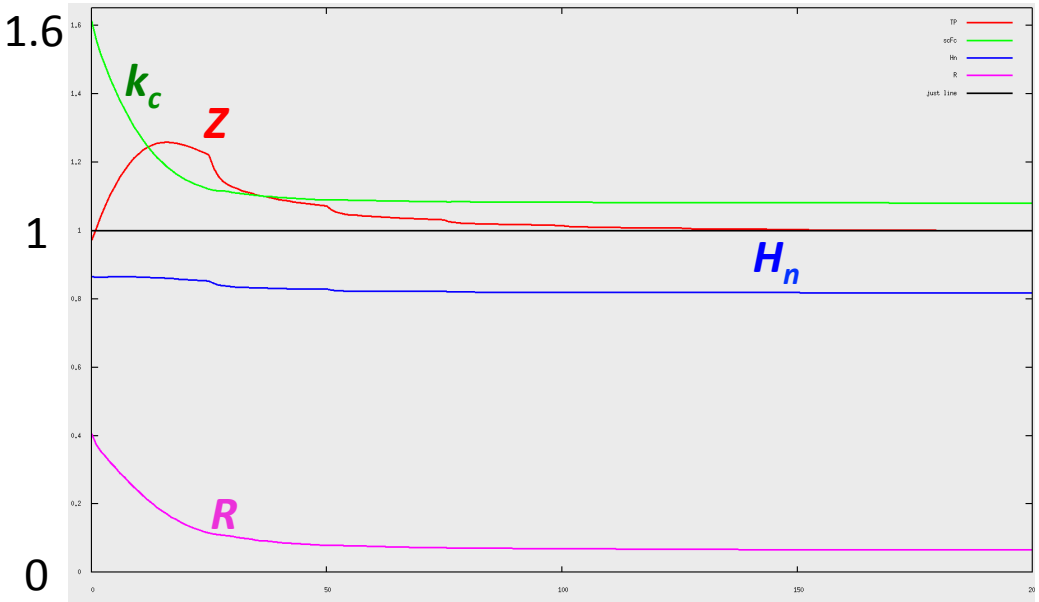
Next iteration map: $\rho_j^{(n+1)} = (1-\beta)\rho_j^n + \beta\frac{1}{Z^{(n+1)}}\tilde{\rho}_j$

Update every 5-25 iterations: $A_{GD}^{new} = A_{GD}^{old} / Z^{(n+1)}$

Calculation algorithm (control parameters)

Control parameters

1. Optimization targets *H(ρ)* and *Q(ρ)*: both should decrease
2. Total unit cell charge $Z = \sum_{j=0}^{N-1} \rho_j$: it should converge to 1
3. Normalized entropy $H_n(\rho) = -\frac{1}{\ln N} \sum_{j=0}^{N-1} \frac{\rho_j}{Z} \ln \frac{\rho_j}{Z}$: it reaches its max value if $\rho_j = 1/N \;\forall\; j$
4. R-factor: $R(\rho) = \sum_{k \in K} \left| \mathbf{F}_k - \mathbf{F}_k^{obs} \right| \Big/ \sum_{k \in K} \left| \mathbf{F}_k^{obs} \right|$
5. Scale factor k_c that minimizes function $\sum_{k \in K} w_k \left| \kappa_c \mathbf{F}_k - \mathbf{F}_k^{obs} \right|^2$: k_c should converge to 1.



Implementation in Phenix

Source code

```
cctbx/maptbx/mem.py                          : 288 lines
cctbx/regression/tst_mem.py                  : 116 lines
cctbx/maptbx/statistics.h                    : 75 lines
cctbx/maptbx/boost_python/statistics.cpp     : 24 lines
```

Usage (command line):

```
phenix.max_entropy_map model.pdb map.mtz
```

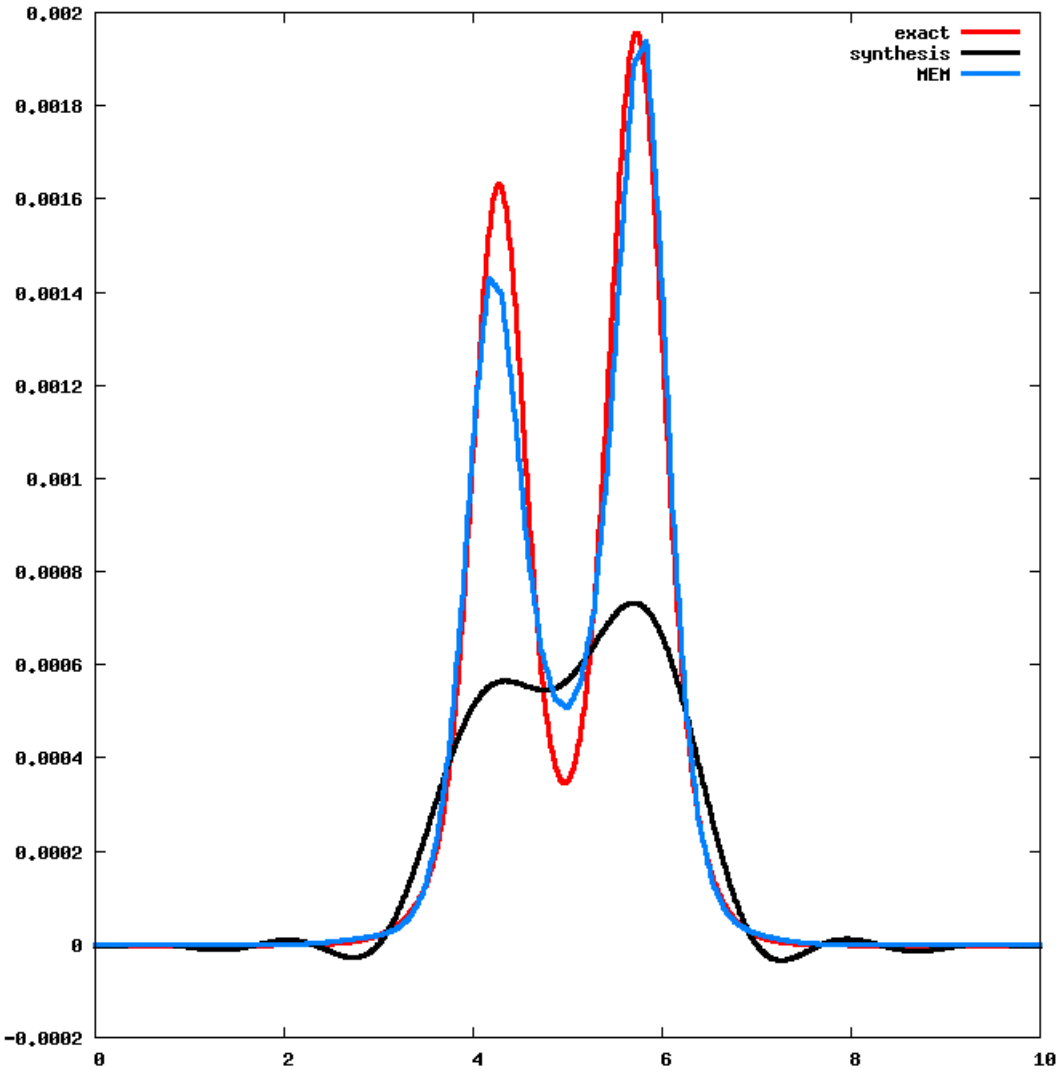
GUI is available (THANKS NAT!)

Program outputs one MTZ file containing two maps: original and MEM maps

Original and MEM maps are scaled such that they have identical cumulative histogram

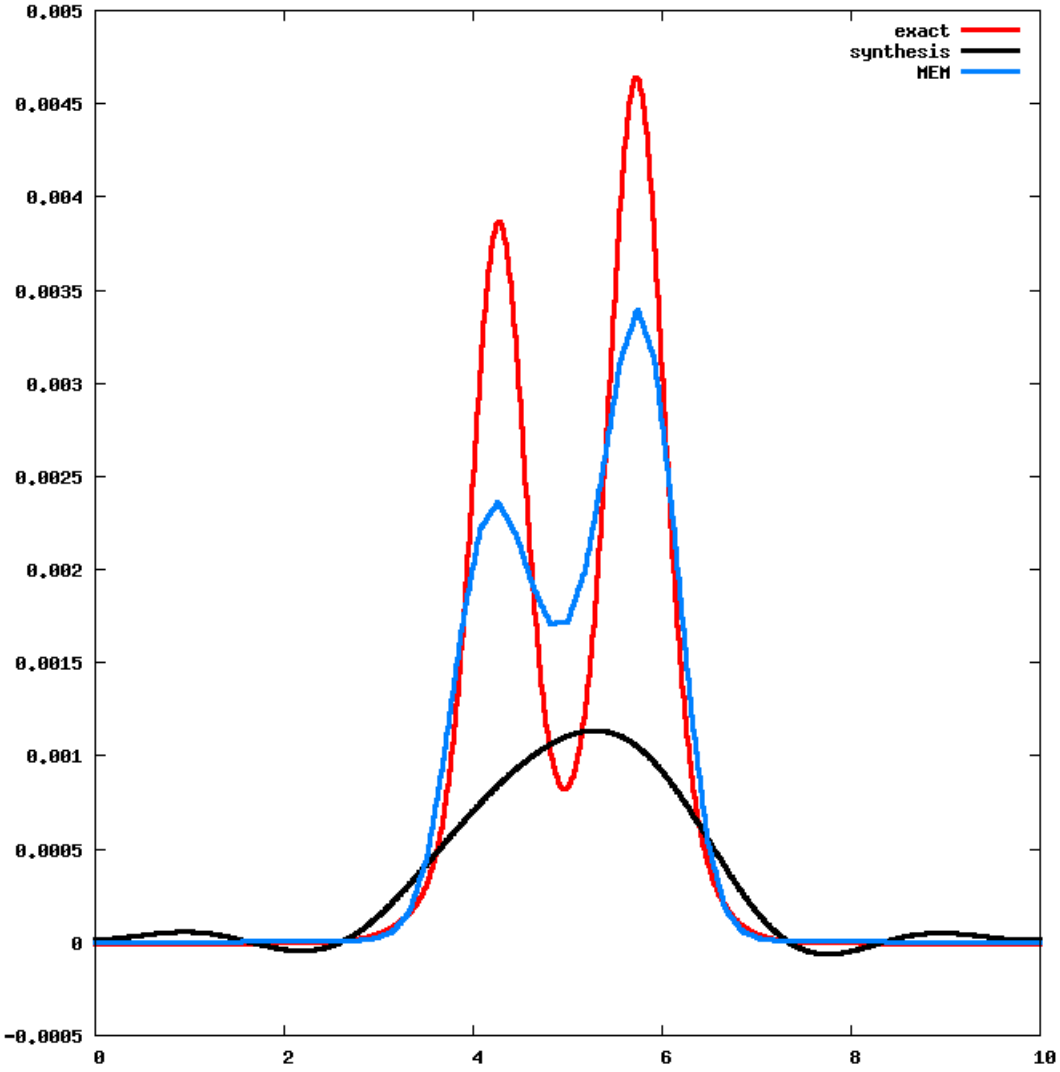
Examples

Restoration of 1.5 Å resolution image



Examples

Restoration of 2 Å resolution image



Bottlenecks

- Runtime: takes from few seconds to few minutes
- Sensitive to F000 estimation (needs to be accurate)
- Starting data is fixed: updating phases should improve the impact:
  - This is why effect with real data is not strong
- Generates map coefficients up to very high resolution: ~0.3..0.5 Å
  - Coot is too slow