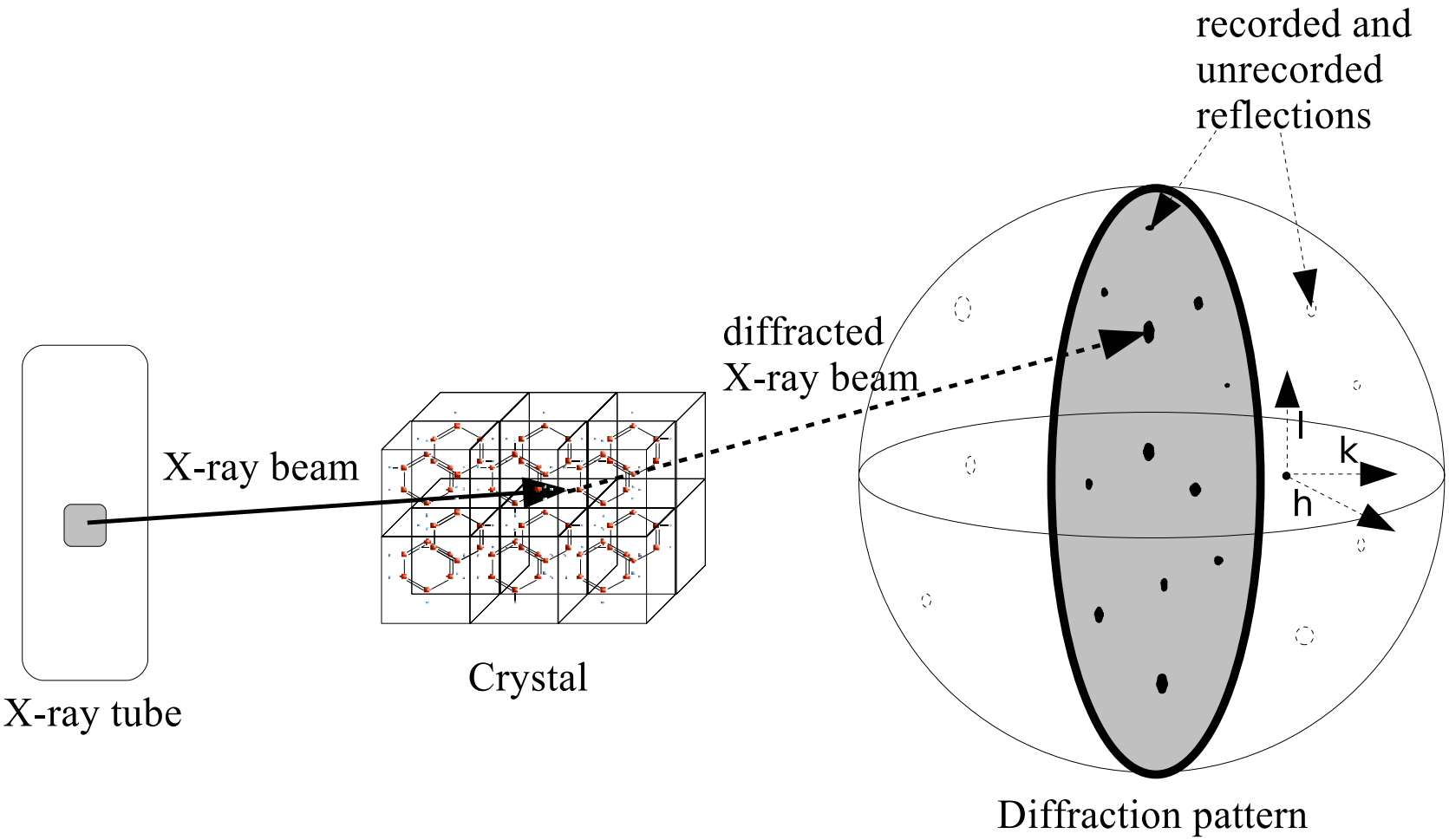
Geometric Constraints for the Phase Problem in X-Ray Crystallography

Corinna Heldt Alexander Bockmayr

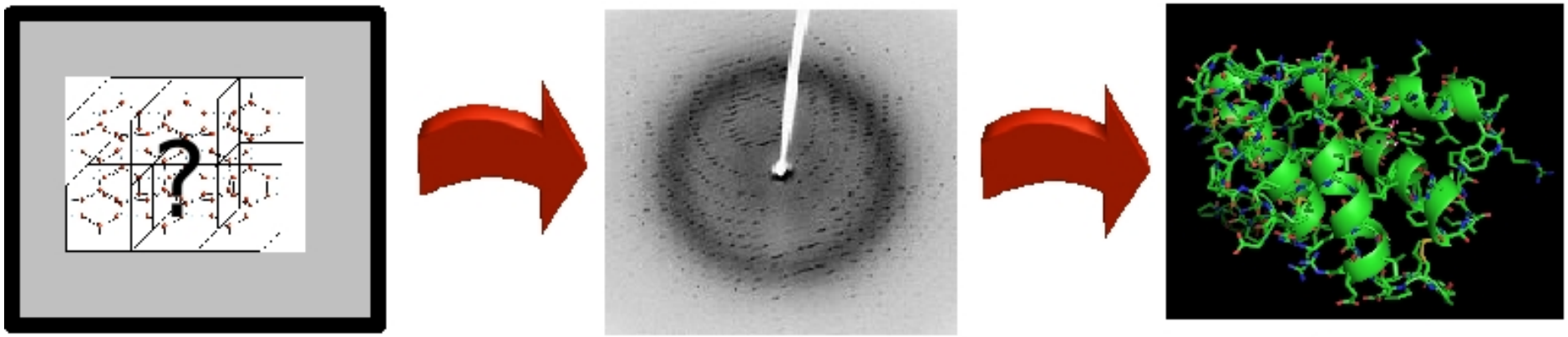
Fachbereich Mathematik und Informatik
Freie Universität Berlin

Workshop on Constraint Based Methods for Bioinformatics
July 21th, 2010 Edinburgh

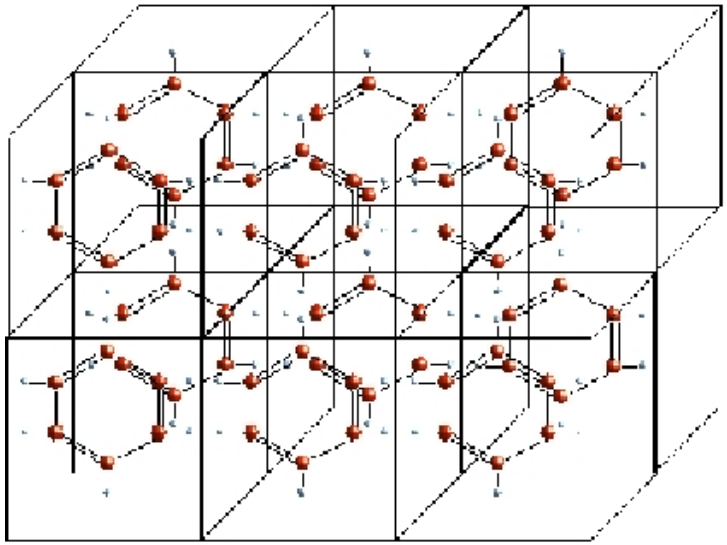
X-ray Diffractometer



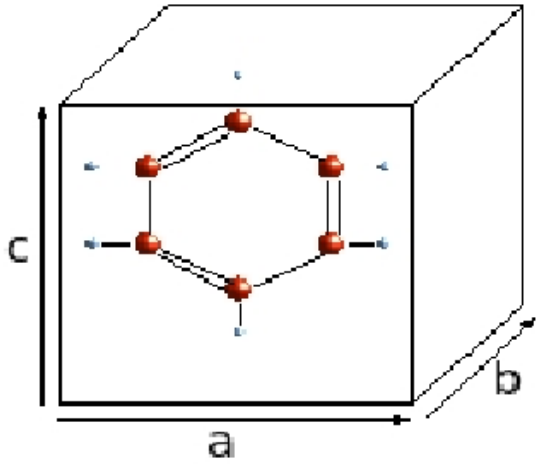
X-ray Crystallography



Electron Density in the Unit Cell



Crystal



Unit cell

Electron density distribution $\rho(\mathbf{x})$ is a periodic function

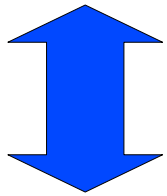
$\Rightarrow$ can be developed into a Fourier series

Electron Density vs. Structure Factors

Fourier coefficients

Electron density: $$\rho(\mathbf{x}) = \frac{1}{V_{cell}} \sum_{\mathbf{h}\in\mathbb{Z}^3} \mathbf{F}(\mathbf{h}) \exp(-2\pi i(\mathbf{h}\cdot\mathbf{x})),\ \mathbf{x}\in V$$

Volume of the unit cell — Unit cell



Fourier transform

Structure factors: $$\mathbf{F}(\mathbf{h}) = \int_V \rho(\mathbf{x}) \exp(2\pi i(\mathbf{h}\cdot\mathbf{x}))d\mathbf{x},\ \mathbf{h}\in\mathbb{Z}^3$$

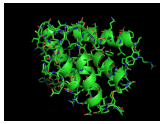
$$= |\mathbf{F}(\mathbf{h})| e^{i\varphi(\mathbf{h})}$$

complex number

magnitude — phase

The Phase Problem

Electron density



$\rho(\mathbf{x})$

$$\rho(\mathbf{x}) = \frac{1}{V} \sum_{\mathbf{h} \in \mathbb{Z}^3} \mathbf{F(h)} e^{-2\pi i(\mathbf{h}\cdot\mathbf{x})}$$

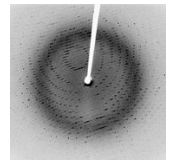
Structure factors

$$\mathbf{F(h)} = |\mathbf{F(h)}| e^{i\varphi(\mathbf{h})}$$

Phase?

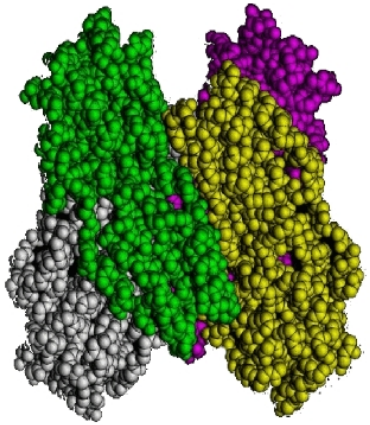
$$I(\mathbf{h}) \propto |\mathbf{F(h)}|$$

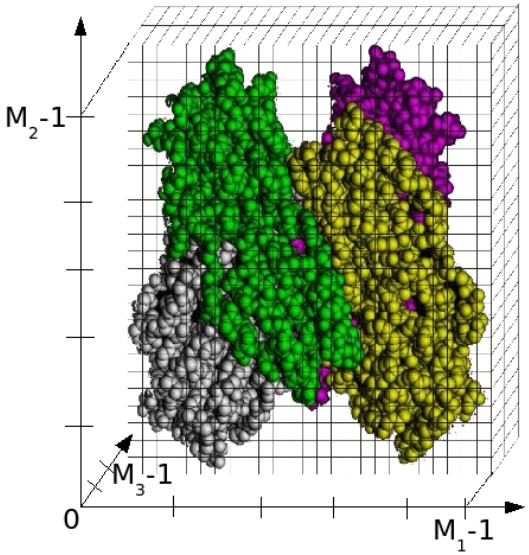
Experimental data



$I(\mathbf{h})$

Grid Electron Density



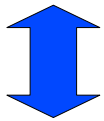


$$\Pi = [0, M_1 - 1] \times [0, M_2 - 1] \times [0, M_3 - 1] \subseteq \mathbb{Z}^3$$

$$\mathbf{M} = \begin{pmatrix} M_1 & 0 & 0 \\ 0 & M_2 & 0 \\ 0 & 0 & M_3 \end{pmatrix}, \quad M = M_1 M_2 M_3$$

$$\rho_g(\mathbf{j}) = \rho(\mathbf{M}^{-1}\mathbf{j})$$

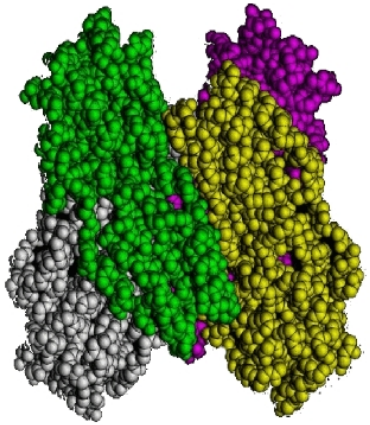
Grid electron densities: $\rho_g(\mathbf{j}) = \sum_{\mathbf{h} \in \Pi} \mathbf{F}_g(\mathbf{h}) \exp(-2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{j} \in \Pi$

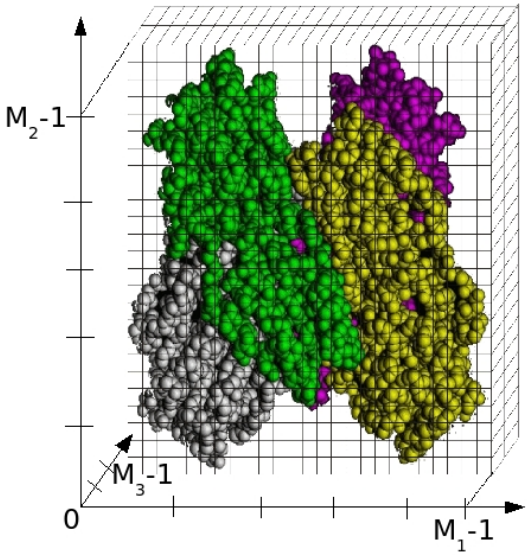


Discrete Fourier transform

Grid structure factor: $\mathbf{F}_g(\mathbf{h}) = \frac{1}{M} \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{h} \in \Pi$

Grid Electron Density





$$\Pi = [0, M_1 - 1] \times [0, M_2 - 1] \times [0, M_3 - 1] \subseteq \mathbb{Z}^3$$

$$\mathbf{M} = \begin{pmatrix} M_1 & 0 & 0 \\ 0 & M_2 & 0 \\ 0 & 0 & M_3 \end{pmatrix}, \quad M = M_1 M_2 M_3$$

$$\rho_g(\mathbf{j}) = \rho(\mathbf{M}^{-1}\mathbf{j})$$

Grid electron densities: $\rho_g(\mathbf{j}) = \sum_{\mathbf{h} \in \Pi} \mathbf{F}_g(\mathbf{h}) \exp(-2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{j} \in \Pi$

Discrete Fourier transform

Grid structure factor: $\mathbf{F}_g(\mathbf{h}) = \frac{1}{M} \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})), \ \forall \mathbf{h} \in \Pi$

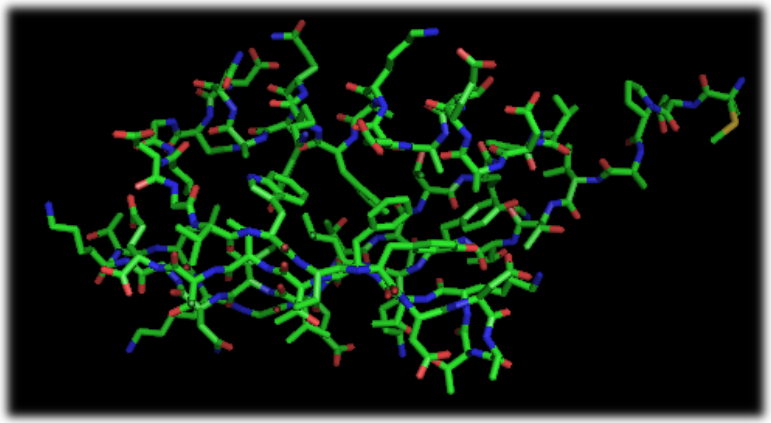
Binary Electron Density

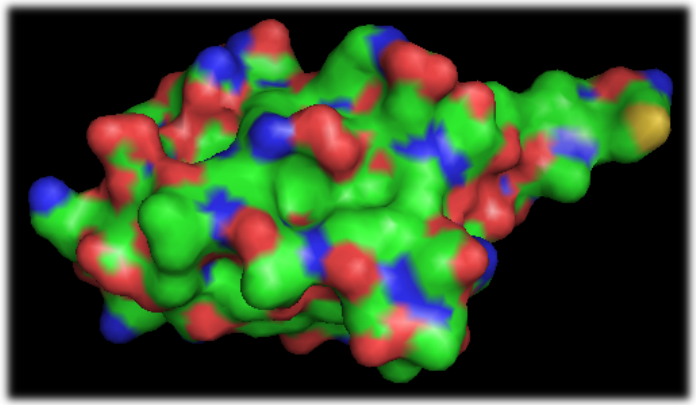
Binary envelope instead of real electron density distribution

$$\rho_g(\mathbf{j}) \in \mathbb{R} \to z_j \in \{0,1\},\ \mathbf{j} \in \Pi,$$

represents areas with electron density above a certain level L:

$$z_{\mathbf{j}} = \begin{cases} 1, & \text{if } \rho_g(\mathbf{j}) \geq L \\ 0, & \text{otherwise.} \end{cases}$$





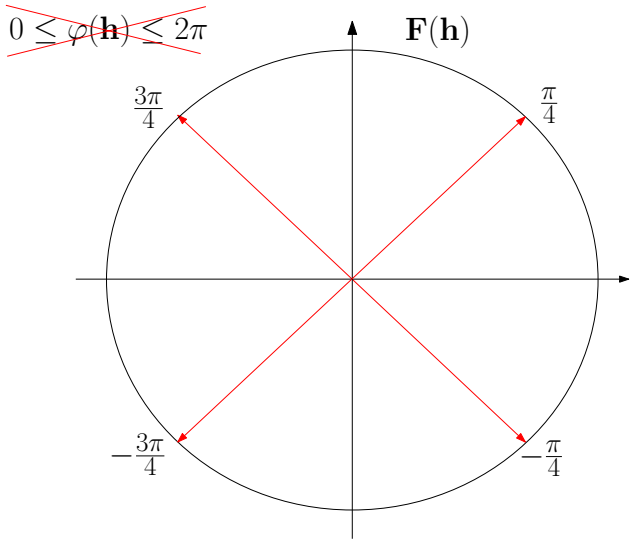
Phase Restrictions

All phase values from 0 to 2π are possible

$\varphi \in [0, 2\pi]$ continous problem

Restriction to four possible phase values:

$$\varphi(\mathbf{h}) \in \left\{ \frac{\pi}{4}, \frac{3\pi}{4}, -\frac{\pi}{4}, -\frac{3\pi}{4} \right\}$$



Introduction of two new binary variables:

$\alpha(\mathbf{h}), \beta(\mathbf{h}) \in \{0, 1\}$

Constraint System

$$\left| \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})) - \frac{M}{V_{cell}} \mathbf{F}(\mathbf{h}) \right| \leq \varepsilon(\mathbf{h}), \ \forall \mathbf{h} \in \Pi$$

Real part

$$\left| \sum_{\mathbf{j} \in \Pi} \cos(2\pi(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})) z_{\mathbf{j}} - \frac{M}{V_{cell}} (2\alpha(\mathbf{h}) - 1) \frac{1}{\sqrt{2}} |F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}),$$

$$\left| \sum_{\mathbf{j} \in \Pi} \sin(2\pi(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})) z_{\mathbf{j}} - \frac{M}{V_{cell}} (2\beta(\mathbf{h}) - 1) \frac{1}{\sqrt{2}} |F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}).$$

Imaginary part

Constraint System

$$\left| \sum_{\mathbf{j}\in\Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h}\cdot\mathbf{M}^{-1}\mathbf{j})) - \frac{M}{V_{cell}}\mathbf{F}(\mathbf{h}) \right| \leq \varepsilon(\mathbf{h}),\ \forall \mathbf{h}\in\Pi$$

$$\left| \sum_{\mathbf{j}\in\Pi} \cos(2\pi(\mathbf{h}\cdot\mathbf{M}^{-1}\mathbf{j}))z_{\mathbf{j}} - \frac{M}{V_{cell}}(2\alpha(\mathbf{h})-1)\frac{1}{\sqrt{2}}|F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}),$$

$$\left| \sum_{\mathbf{j}\in\Pi} \sin(2\pi(\mathbf{h}\cdot\mathbf{M}^{-1}\mathbf{j}))z_{\mathbf{j}} - \frac{M}{V_{cell}}(2\beta(\mathbf{h})-1)\frac{1}{\sqrt{2}}|F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}).$$

↑

Electron density discretisation

Constraint System

$$\left| \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})) - \frac{M}{V_{cell}} \mathbf{F(h)} \right| \leq \varepsilon(\mathbf{h}), \ \forall \mathbf{h} \in \Pi$$

$$\left| \sum_{\mathbf{j} \in \Pi} \cos(2\pi(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j}))z_{\mathbf{j}} - \frac{M}{V_{cell}}(2\alpha(\mathbf{h}) - 1)\frac{1}{\sqrt{2}}|F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}),$$

$$\left| \sum_{\mathbf{j} \in \Pi} \sin(2\pi(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j}))z_{\mathbf{j}} - \frac{M}{V_{cell}}(2\beta(\mathbf{h}) - 1)\frac{1}{\sqrt{2}}|F(\mathbf{h})| \right| \leq \varepsilon(\mathbf{h}).$$

↑

Phase value restriction

Constraint System

$$\left| \sum_{\mathbf{j} \in \Pi} \rho_g(\mathbf{j}) \exp(2\pi i(\mathbf{h} \cdot \mathbf{M}^{-1}\mathbf{j})) - \frac{M}{V_{cell}} \mathbf{F}(\mathbf{h}) \right| \leq \varepsilon(\mathbf{h}), \ \forall \mathbf{h} \in \Pi$$

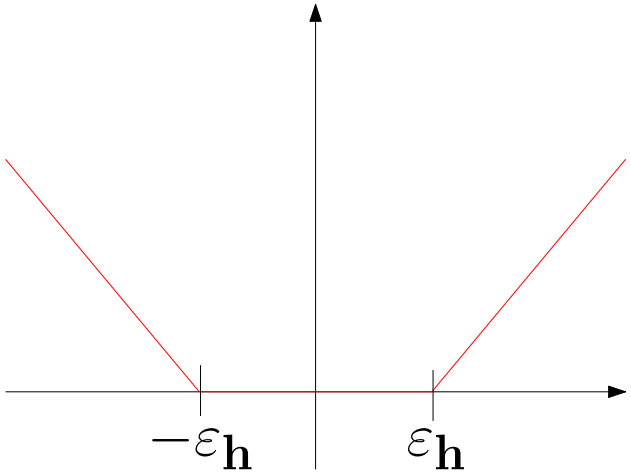
$$\left| A^{Re}(\mathbf{h}, \mathbf{z}, \alpha(\mathbf{h})) \right| \leq \varepsilon(\mathbf{h}),$$

$$\left| A^{Im}(\mathbf{h}, \mathbf{z}, \beta(\mathbf{h})) \right| \leq \varepsilon(\mathbf{h}).$$

Linear Pseudo Boolean Program (0-1 ILP)

$$
\begin{aligned}
& \min && \sum_{\mathbf{h}\in\Pi} (r_{\mathbf{h}}^{Re} + r_{\mathbf{h}}^{Im}) \\
\text{s.t.} \quad & 0 \leq && r_{\mathbf{h}}^{Re},\ 0 \leq r_{\mathbf{h}}^{Im} \\
& -\varepsilon_{\mathbf{h}} - r_{\mathbf{h}}^{Re} \leq && A^{Re}(\mathbf{h},\mathbf{z},\alpha(\mathbf{h})) \leq \varepsilon_{\mathbf{h}} + r_{\mathbf{h}}^{Re} \\
& -\varepsilon_{\mathbf{h}} - r_{\mathbf{h}}^{Im} \leq && A^{Im}(\mathbf{h},\mathbf{z},\beta(\mathbf{h})) \leq \varepsilon_{\mathbf{h}} + r_{\mathbf{h}}^{Im} \\
& && \forall \mathbf{h} \in \Pi,
\end{aligned}
$$

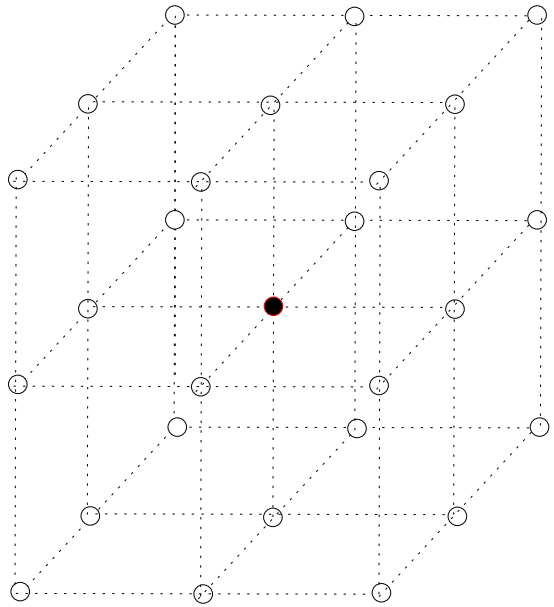
$$
z_{\mathbf{j}},\ \alpha(\mathbf{h}),\ \beta(\mathbf{h}) \in \{0,1\}, \quad \forall \mathbf{h},\mathbf{j} \in \Pi.
$$



Additional Constraints

$\mathbf{j}_1, \mathbf{j}_2 \in \Pi$ *neighbours,* $\mathbf{j}_1 \mathfrak{n} \mathbf{j}_2$, *if:*

$$\| \mathbf{j}_1 - \mathbf{j}_2 \|_2 = 1$$



$z_\mathbf{j} \in \Pi$ *isolated, if:*

$$z_\mathbf{j} = 0 \Rightarrow z_\mathbf{i} = 1, \ \forall \ \mathbf{i} \mathfrak{n} \mathbf{j}$$

or

$$z_\mathbf{j} = 1 \Rightarrow z_\mathbf{i} = 0, \ \forall \ \mathbf{i} \mathfrak{n} \mathbf{j}.$$

Exclusion of isolated interior grid points:

$$-5 \leq z_\mathbf{j} - \sum_{\mathbf{i} \mathfrak{n} \mathbf{j}} z_\mathbf{i} \leq 0, \text{ for all } \mathbf{j} \in \Pi$$

Connectivity Constraint

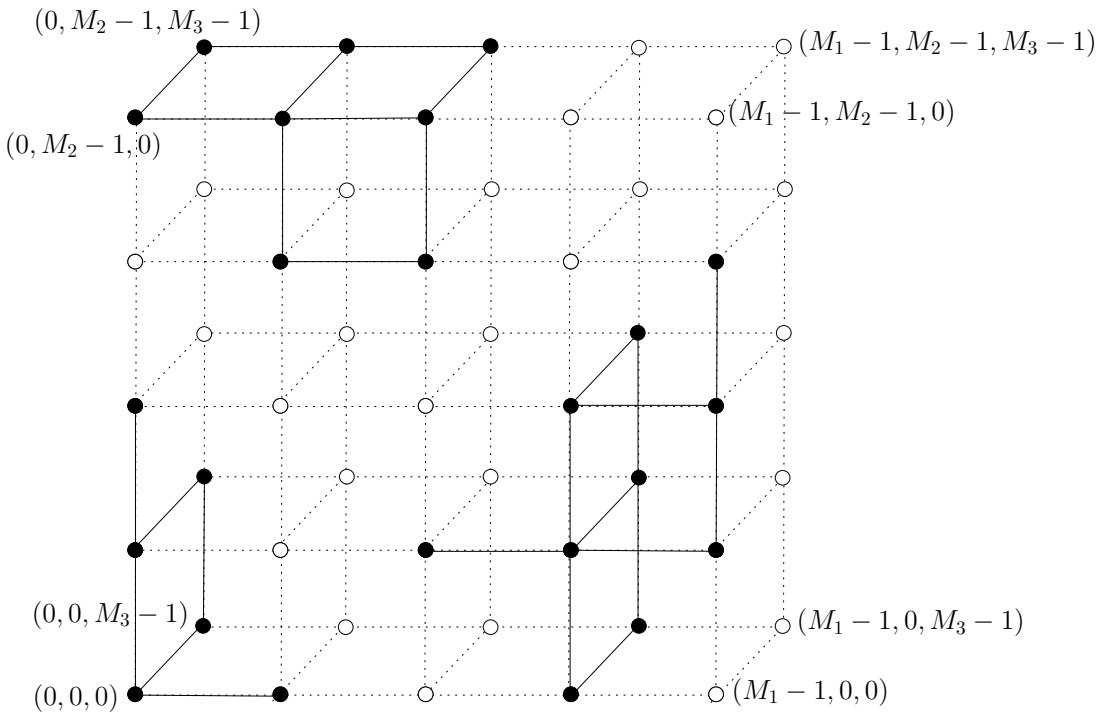


Figure: The graph $G^*_\Pi = (V^*_\Pi, E^*_\Pi)$

$\Omega_\kappa \stackrel{def}{=} \{\mathbf{j} : \rho(\mathbf{j}) \geq \kappa\}$
consists of a small number of connected components

$E_\Pi = \{e = (v_\mathbf{j}, v_\mathbf{i}) \mid \mathbf{j} \mathfrak{n} \mathbf{i}\}$
$V^*_\Pi = \{v_\mathbf{j} \mid z_\mathbf{j} = 1,\ \mathbf{j} \in \Pi\}$
$E^*_\Pi \subseteq E_\Pi$ set of edges induced by V^*_Π

Connectivity Constraint

Any binary grid electron density distribution $z^* \in \{0,1\}^{M_1 \times M_2 \times M_3}$, satisfying the following constraints contains at most K components:

$$-1 \leq \quad 2e_{\mathbf{j}_1\mathbf{j}_2} - z_{\mathbf{j}_1} - z_{\mathbf{j}_2} \leq 0 \quad \leftarrow \text{ Edge constraints}$$

$$\frac{1}{|T_i|}\sum_{\mathbf{j}\in T_i} z_{\mathbf{j}} \leq \qquad u_{T_i} \leq \sum_{\mathbf{j}\in T_i} z_{\mathbf{j}}, \qquad \leftarrow u_T \stackrel{def}{=} \begin{cases} 1, & \text{if } \sum_{\mathbf{j}\in T} z_{\mathbf{j}} \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\sum_{i=1}^{K+1} u_{T_i} - K \leq \sum_{(\mathbf{j}_1,\mathbf{j}_2)\in\delta(T_1,\ldots,T_{K+1})} e_{\mathbf{j}_1\mathbf{j}_2} \quad \leftarrow \text{ K components}$$

$$u_{T_i},\ z_{\mathbf{j}}, e_{\mathbf{j}_1\mathbf{j}_2} \in \{0,1\},$$

$$\forall \emptyset \neq T_1,\ldots,T_{K+1} \subsetneq \Pi,\ \bigcup_{i=1}^{K+1} T_i = \Pi,\ T_i \cap T_j = \emptyset,$$

$$\forall i \neq j,\ i,j \in \{1,\ldots K+1\}, \forall \mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi \text{ with } \mathbf{j_1} \ \mathfrak{n}\ \mathbf{j_2}$$

Connectivity Constraint

Any binary grid electron density distribution $z^* \in \{0,1\}^{M_1 \times M_2 \times M_3}$, satisfying the following constraints contains at most K components:

$$-1 \leq \quad 2e_{\mathbf{j_1 j_2}} - z_{\mathbf{j_1}} - z_{\mathbf{j_2}} \leq 0 \quad \leftarrow \text{ Edge constraints}$$

$$\frac{1}{|T_i|} \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}} \leq \qquad u_{T_i} \leq \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}}, \qquad \leftarrow \; u_T \stackrel{def}{=} \begin{cases} 1, & \text{if } \sum_{\mathbf{j} \in T} z_{\mathbf{j}} \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\sum_{i=1}^{K+1} u_{T_i} - K \leq \sum_{(\mathbf{j_1}, \mathbf{j_2}) \in \delta(T_1, \ldots, T_{K+1})} e_{\mathbf{j_1 j_2}} \quad \leftarrow \text{ K components}$$

$$u_{T_i},\ z_{\mathbf{j}}, e_{\mathbf{j_1 j_2}} \in \{0,1\},$$

$$\forall \emptyset \neq T_1, \ldots, T_{K+1} \subsetneq \Pi,\ \bigcup_{i=1}^{K+1} T_i = \Pi,\ T_i \cap T_j = \emptyset,$$

$$\forall i \neq j,\ i,j \in \{1, \ldots K+1\}, \forall \mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi \text{ with } \mathbf{j_1} \ \text{n} \ \mathbf{j_2}$$

Connectivity Constraint

Any binary grid electron density distribution $z^* \in \{0,1\}^{M_1 \times M_2 \times M_3}$, satisfying the following constraints contains at most K components:

$$-1 \leq \quad 2e_{\mathbf{j_1 j_2}} - z_{\mathbf{j_1}} - z_{\mathbf{j_2}} \leq 0 \quad \leftarrow \text{Edge constraints}$$

$$\frac{1}{|T_i|} \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}} \leq \qquad u_{T_i} \leq \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}}, \qquad \leftarrow u_T \stackrel{def}{=} \begin{cases} 1, & \text{if } \sum_{\mathbf{j} \in T} z_{\mathbf{j}} \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\sum_{i=1}^{K+1} u_{T_i} - K \leq \sum_{(\mathbf{j_1}, \mathbf{j_2}) \in \delta(T_1, \ldots, T_{K+1})} e_{\mathbf{j_1 j_2}} \quad \leftarrow \text{K components}$$

$$u_{T_i},\ z_{\mathbf{j}}, e_{\mathbf{j_1 j_2}} \in \{0,1\},$$

$$\forall \emptyset \neq T_1, \ldots, T_{K+1} \subsetneq \Pi,\ \bigcup_{i=1}^{K+1} T_i = \Pi,\ T_i \cap T_j = \emptyset,$$

$$\forall i \neq j,\ i,j \in \{1, \ldots K+1\}, \forall \mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi \text{ with } \mathbf{j_1} \ \text{n} \ \mathbf{j_2}$$

Connectivity Constraint

Any binary grid electron density distribution $z^* \in \{0,1\}^{M_1 \times M_2 \times M_3}$, satisfying the following constraints contains at most K components:

$$-1 \leq \quad 2e_{\mathbf{j_1 j_2}} - z_{\mathbf{j_1}} - z_{\mathbf{j_2}} \leq 0 \quad \leftarrow \text{Edge constraints}$$

$$\frac{1}{|T_i|} \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}} \leq \qquad u_{T_i} \leq \sum_{\mathbf{j} \in T_i} z_{\mathbf{j}}, \qquad \leftarrow u_T \stackrel{def}{=} \begin{cases} 1, & \text{if } \sum_{\mathbf{j} \in T} z_{\mathbf{j}} \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

$$\sum_{i=1}^{K+1} u_{T_i} - K \leq \sum_{(\mathbf{j_1},\mathbf{j_2}) \in \delta(T_1,\ldots,T_{K+1})} e_{\mathbf{j_1 j_2}} \quad \leftarrow \text{K components}$$

$$u_{T_i},\ z_{\mathbf{j}}, e_{\mathbf{j_1 j_2}} \in \{0,1\},$$

$$\forall \emptyset \neq T_1, \ldots, T_{K+1} \subsetneq \Pi,\ \bigcup_{i=1}^{K+1} T_i = \Pi,\ T_i \cap T_j = \emptyset,$$

$$\forall i \neq j,\ i,j \in \{1, \ldots K+1\}, \forall \mathbf{j}, \mathbf{j_1}, \mathbf{j_2} \in \Pi \text{ with } \mathbf{j_1} \ \mathfrak{n} \ \mathbf{j_2}$$

Separation

Number of constraints grows exponentially in number of nodes

$$\Rightarrow \text{Separation algorithm}$$

- Calculate solution with only some K-component constraints
- Solution optimal or find violated constraint
- Add cutting plane

$$\frac{1}{|\tilde{T}_i|} \sum_{\mathbf{j} \in \tilde{T}_i} z_{\mathbf{j}} \leq \qquad u_{\tilde{T}_i} \leq \sum_{\mathbf{j} \in \tilde{T}_i} z_{\mathbf{j}}$$

$$\sum_{i=1}^{K+1} u_{\tilde{T}_i} - K \leq \sum_{(\mathbf{j_1}, \mathbf{j_2}) \in \delta(\tilde{T}_1, \ldots, \tilde{T}_{K+1})} e_{\mathbf{j_1} \mathbf{j_2}}$$

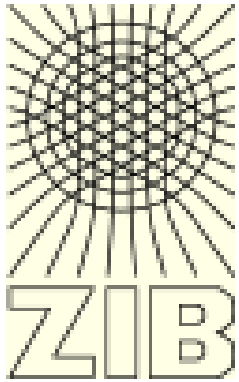
$\emptyset \neq \tilde{T}_1, \ldots, \tilde{T}_{K+1} \subsetneq \Pi$ partition of Π, violating constraints

Implementation

Implementation in SCIP
(Solving Constraint Integer Programs)

IP-solver: CPLEX

SCIP can handle MIP as well as CIP problems



Evaluate Solution Quality

Distance between exact and calculated electron density:

$$D(z_{exact}, z^i_{calc}) \stackrel{def}{=} \sum_{\mathbf{j}\in\Pi} \left| z_{exact}(\mathbf{j}) - z^i_{calc}(\mathbf{j}) \right| \tag{1}$$

Average solution:

$$z_{av}(\mathbf{j}) \stackrel{def}{=} \frac{1}{N} \sum_{i=1}^{N} z^i_{calc}(\mathbf{j}),\ \forall \mathbf{j} \in \Pi, \quad D_{av} \stackrel{def}{=} D(z_{exact}, z_{av}) \tag{2}$$

Solution with minimum distance from all other solutions:

$$D_{sum}(i) \stackrel{def}{=} \sum_{j=1}^{N} \sum_{\mathbf{j}\in\Pi} \left| z^i_{calc}(\mathbf{j}) - z^j_{calc}(\mathbf{j}) \right|,\ \forall i \in \{1, \ldots, N\}, \tag{3}$$

$$z_{ref} \stackrel{def}{=} z^i_{calc},\ \text{with}\ D_{sum}(i) \stackrel{def}{=} \min_{j=1}^{N} \{D_{sum}(j)\}, \quad D_{ref} \stackrel{def}{=} D(z_{exact}, z_{ref}) \tag{4}$$

Computational Results on a $6 \times 6 \times 6$ – *grid*

Constraints	# sol	p_{min}	p_{av}	p_{ref}
none	70	72%	56%	54%
iso	67	72%	62%	54%
connected (2)	49	72%	66%	63%
connected (1)	28	72%	74%	65%
iso, connected (2)	49	72%	69%	68%
iso, connected (1)	28	72%	74%	70%

Consistency with exact binary solution

Ongoing and further work

- find and add more constraints
- decrease running time of the solving algorithm
  - consider bigger grids
  - consider more phase values than just four ones

The End

Thank you for your attention!