Advanced Crystal Structure Analysis

Modern Cross Validation in Crystallography

Tim Grüne
Georg-August-Universität
Institut für Strukturchemie

http://shelx.uni-ac.gwdg.de/~tg
tg@shelx.uni-ac.gwdg.de

17th July 2014

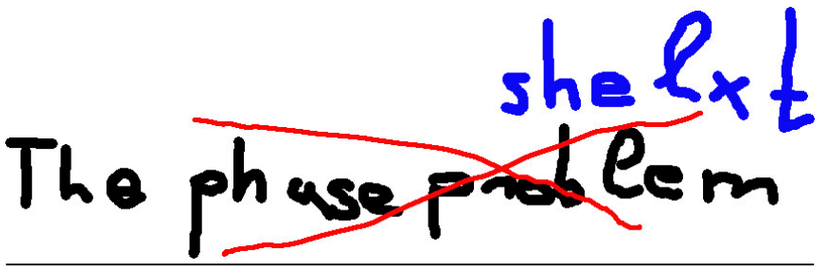
Overview

The Phase Problem in Refinement

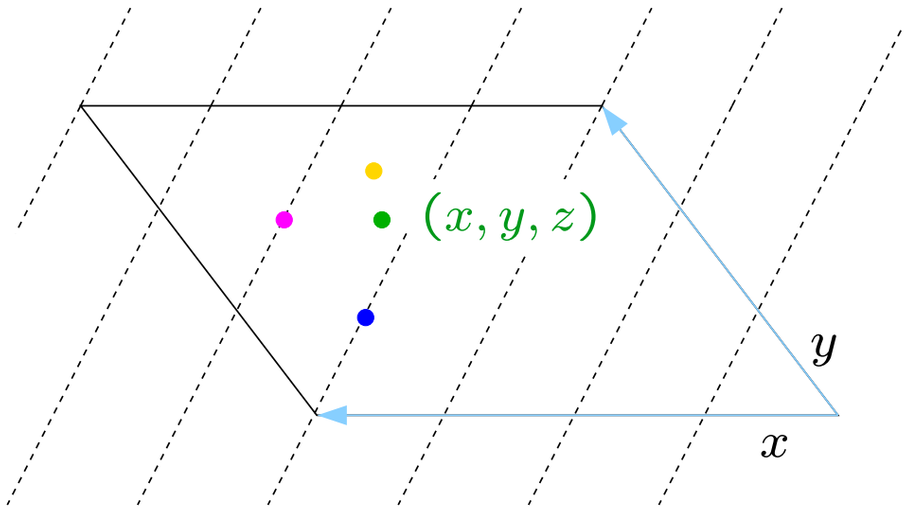
Cross-Validation: Meaning and Examples

Means of Cross-Validation

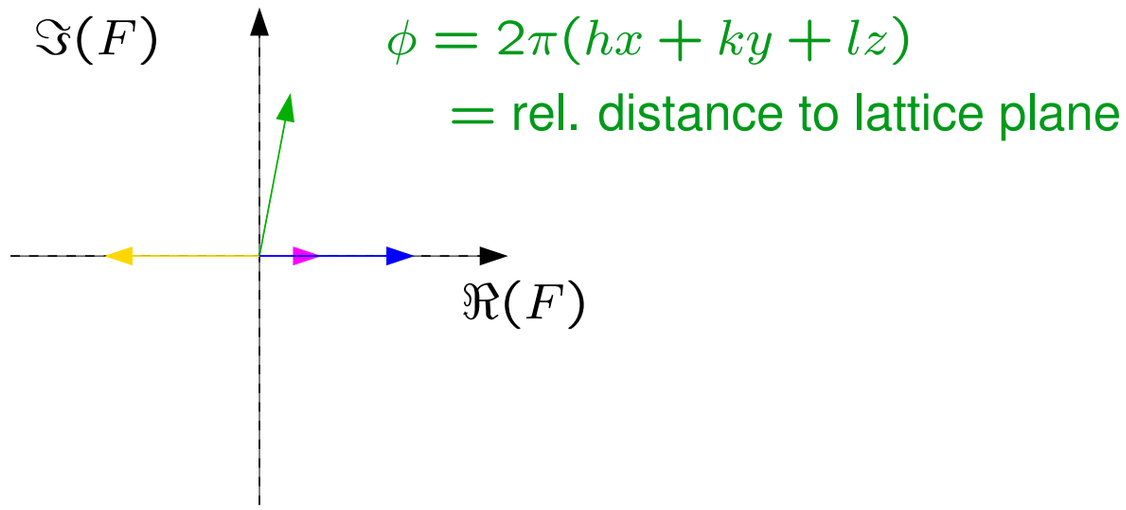
Complete R_{free} with Shelxl


The phase problem
shelxt

$$\underline{f_i e^{-2\pi i(hx_i+ky_i+lz_i)}}$$



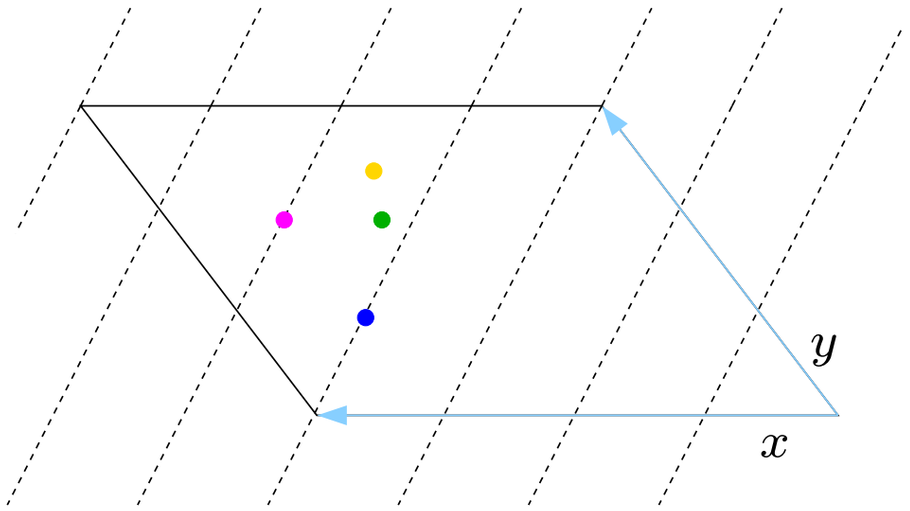
$(hkl) = (430)$



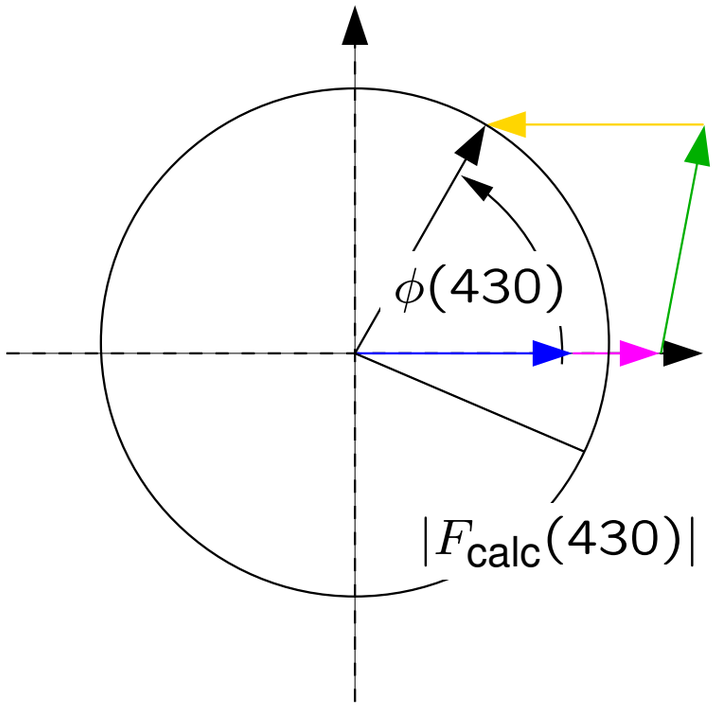
$\phi = 2\pi(hx + ky + lz)$
$=$ rel. distance to lattice plane

(atomic constituents for $F(430)$)

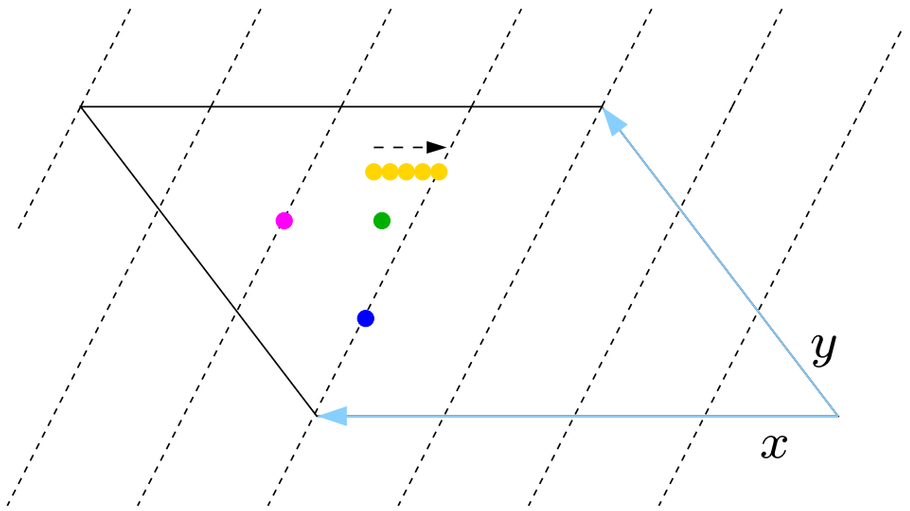
$$F(hkl) = \sum_i f_i e^{-2\pi i(hx_i + ky_i + lz_i)}$$



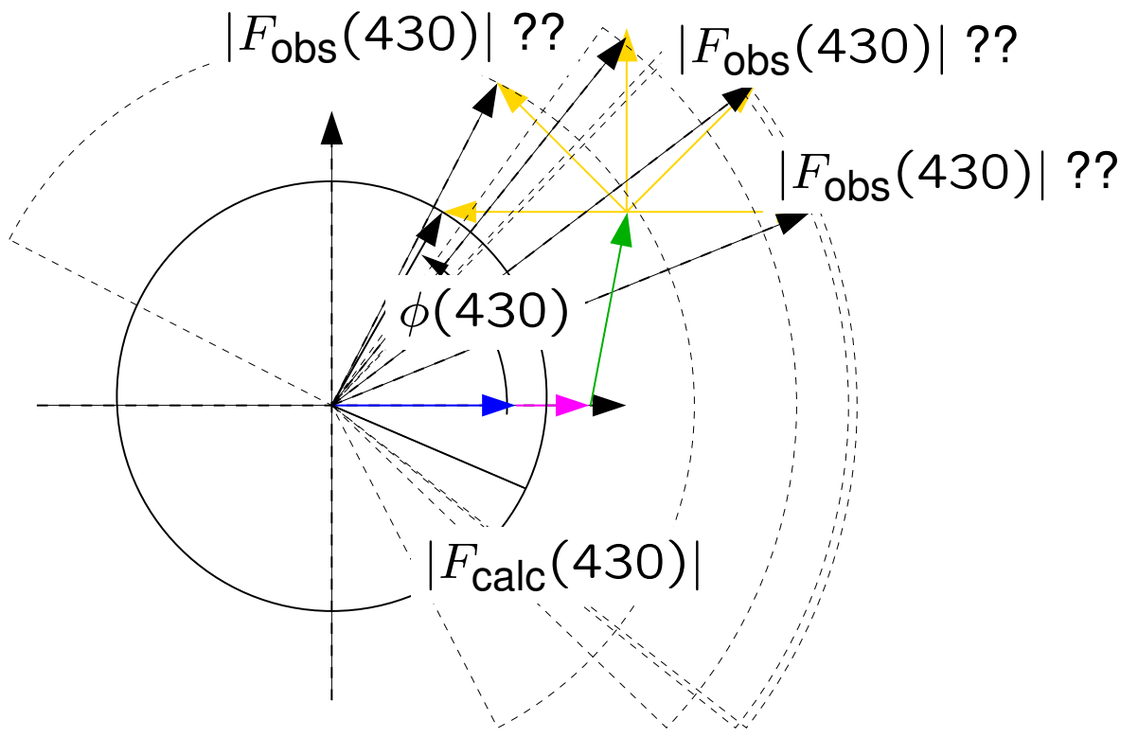
$(hkl) = (430)$



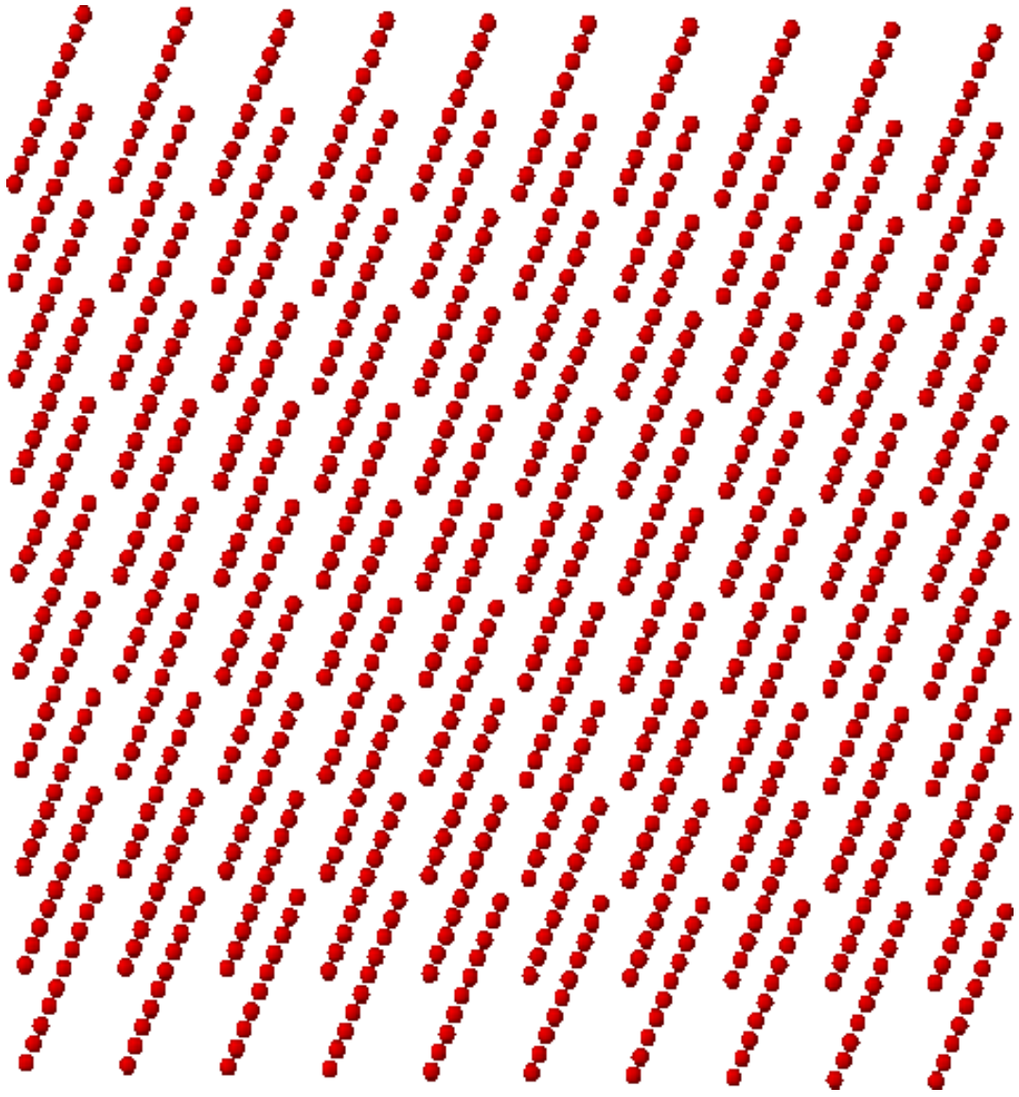
Refinement: match $|F_{\text{calc}}(hkl)|$ and $|F_{\text{obs}}(hkl)$



$(hkl) = (430)$

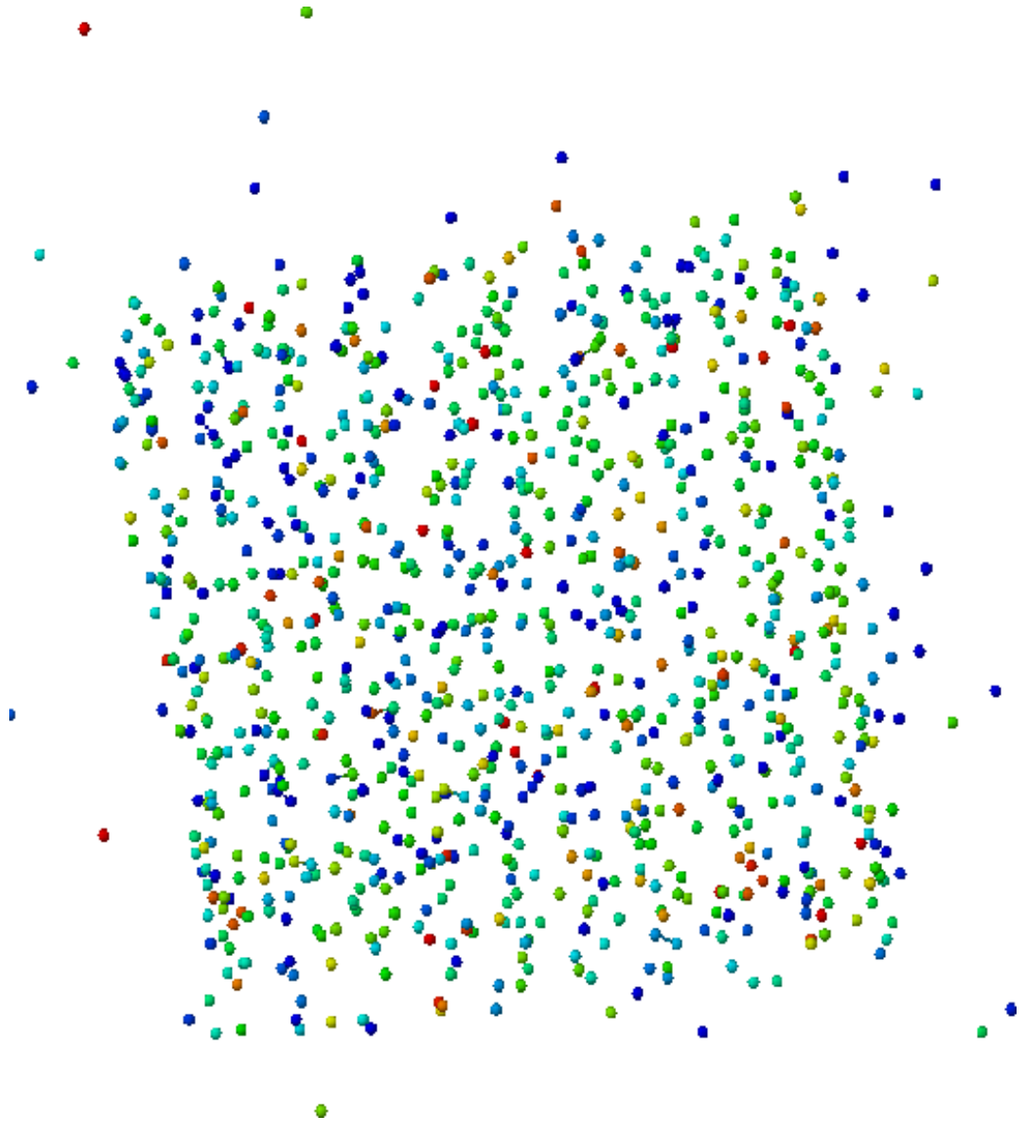


Overfitting $R1$



- Take a 1.2 Å data set
- Fill the unit cell with a grid of atoms
- Free refinement with `shelxl`

Overfitting $R1$



Resolution	$\mathrm{R_{work}}$	$\mathrm{R_{free}}$
3.0 Å	7.2%	55.3%
2.5 Å	17.8%	52.6%
2.0 Å	27.7%	56.6%
1.5 Å	33.7%	54.5%
1.2 Å	38.6%	54.4%

Resolution *vs.* Data to Parameter Ratio

- R_{free} introduced for proteins: increased risk of overfitting (Brünger, [1], 1990)

- Classic work: Kleywegt/Jones [2], 1995 — reverse peptide chain (C- to N-terminus):

$$d = 1.8\text{Å}, R1 = 21.4\ \% \text{ with good geometry } \textit{etc.}$$

- Only a matter of resolution?

Overfitting at $d = 0.44$Å

- 0.44 Å resolution
- 10,000 cycles CGLS unrestrained
- $R1 = 2.86$ %; $wR2 = 6.7$ %

Overfitting at $d = 0.44$Å

- 0.44 Å resolution
- 10,000 cycles CGLS unrestrained
- $R1 = 2.86$ %; $wR2 = 6.7$ %
- 5000 (random) reflections and 6581 parameters
- all 42998 reflections and 6581 parameters: $R1 = 41.3$ %; $wR2 = 77.9$ %

Principle of Cross Validation

Input: 1. N Data points

2. Model

3. Method to calculate Model Quality ($R1wR2, \ldots$)

Method :

1. exclude *free* set with n data points

2. prepare model against $N - n$ data points

3. Calculate Model quality from *free* set

4. **Ideally: Repeat**, so that **each** data point is free exactly once

Crystallographic Problem(s)

- Crystallographic Model Preparation: Iterative process
- Would require N/n entire processes from model solution to model building and refinement
- $\rightarrow$ Impractical to infeasible

"Classical" R_{free}

- **Before** structure solution: mark, say, $k = 500$ reflections as *free*.
- Do not use these 500 reflections **throughout** structure solution and refinement
- Calculate R_{free} from these 500 reflections at each refinement step and compare with $R1$
- $\Rightarrow$ "unbiased" R-value

Problem: Choice of set size k: $\sigma(R_{\text{free}}) \approx R_{\text{free}}/\sqrt{k}$

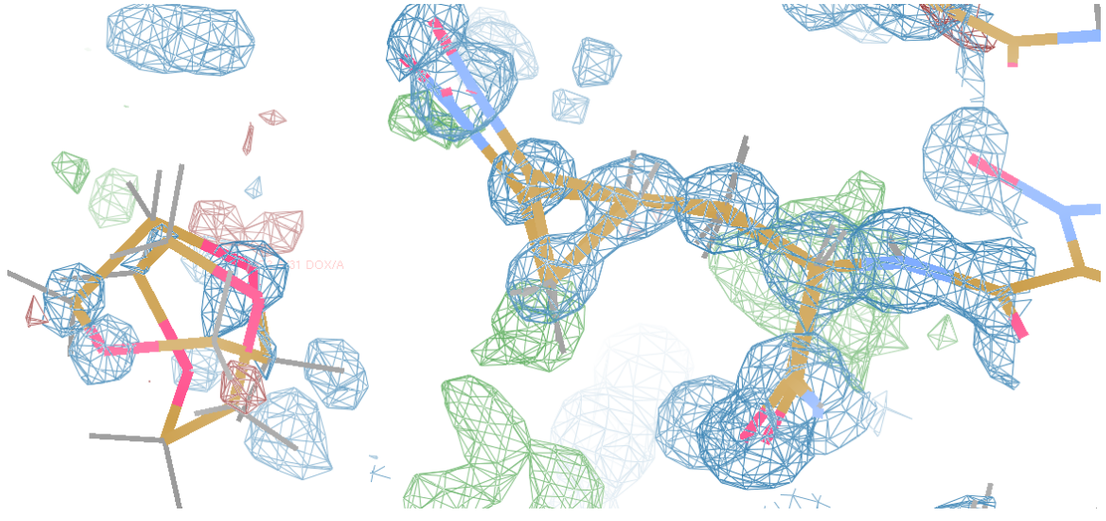
$$k = 500 \Rightarrow 1/\sqrt{k} = 0.04 \Rightarrow R_{\text{free}} = 10.0\% \pm 0.4$$

$k = 500$ and Small Molecule Data

Example: Antibiotic Hormaomycin @ 1.02Å resolution

- 7898 unique reflections, 1921 parameters

$$\frac{7898}{1921} = 4.11 \qquad \frac{7898 - 500}{1921} = 3.85$$



Don’t want to waste any reflection

Practical Complete Cross Validation

(suggested by Brünger, [1])

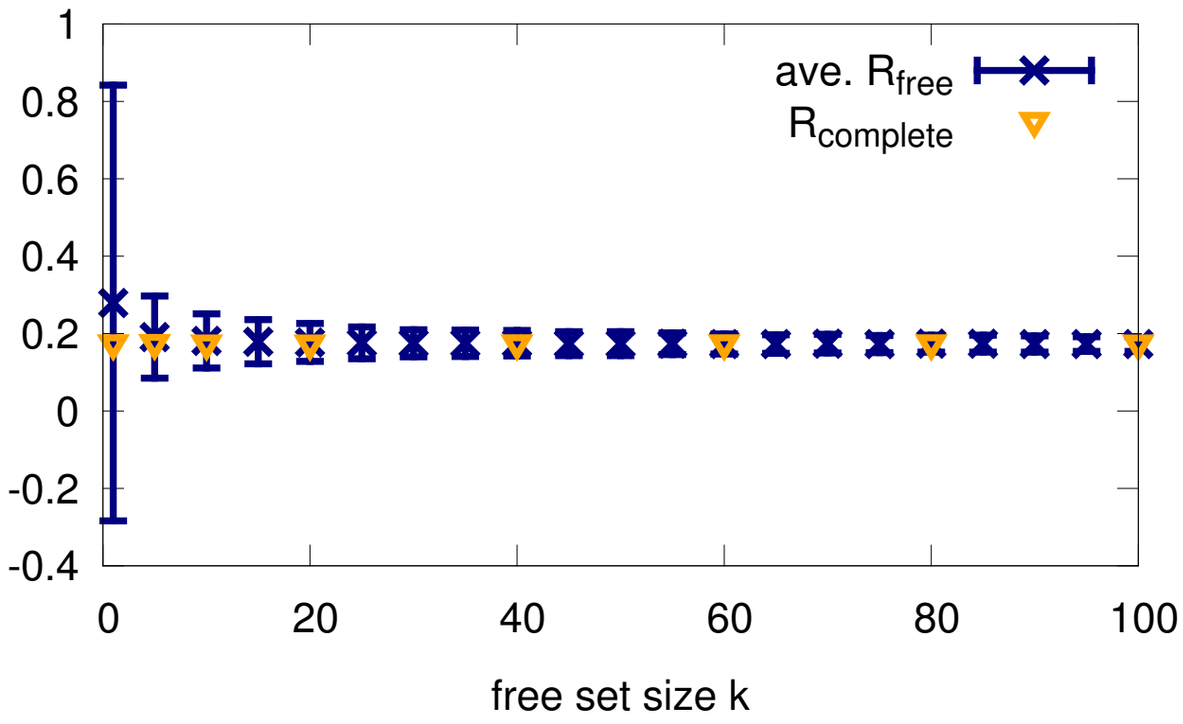
- Create one data set for every *e.g.* 50 reflections (*crossflaghkl -k50*)
- Make model *independent* from reflections
- Refine against $N - 50$ reflections
- $N/50$ refinement run, no model building required
- Calculate R from all left-out reflections

$$R_{\text{complete}} = \frac{\sum_h |F_{obs}(h)| - k|F_{calc}(h)|}{\sum_h |F_{obs}(h)|}$$

Average R_{free} and Complete R_{complete}

Background: PDB_REDO calculates average R_{free} instead of R_{complete}

$\sigma(R_{\text{free}}) \approx R_{\text{free}}/\sqrt{k}$



R_{free} becomes both unstable and unreliable the smaller k.

R_{complete} is independent from choice of k

Average R_{free} and Complete R_{complete}

- Both R_{complete} and average R_{free} are calculated using **all** reflections
- tempting: average R_{free} provides mean and standard deviation
- average R_{free} becomes unstable with small free sets (=low k)
- R_{complete} is stable independent of k: make your own choice!
- R_{complete} has no average: it is **the unbiased** $R1$

Independence of *free* reflections: `WIGL`

- SHELXL command WIGL: random noise to
  1. coordinates
  2. U_{ij} values
  3. `WIGL -0.2 0.2`: repetitively random shifts
  4. **Caveat:** `WIGL 0.2 0.2` (default): each run identical

Overfitting d=0.44Å with 5000 reflections:

	`WIGL -0.2 0.3`	no `WIGL`
$R1$	5.6 % $\pm$ 0.6 %	2.83 % $\pm$ 0.01 %
R_{complete}	**127 %**	5.59 %
	obvious non-sense	obvious bias

WIGL makes coordinate independent from reflections

HOWTO: Complete Cross Validation with SHELXL

SHELXL lst-file contains

```
Nfree(I>2sig)      40    Sum|Fo-Fc| 0.163348E+02    Sum(Fo) 0.403483E+03
Nfree(all)         50    Sum|Fo-Fc| 0.224635E+02    Sum(Fo) 0.416327E+03
                         ^^^^^^^^^^^^^^^^^^^^^^^    ^^^^^^^^^^^^^^^^^^^^
                               nominator                  denominator
```

```
#> grep "Nfree(all)" kcross*.lst > Nfree_all.data
#> awk '{ sumDF += $5; sumFo += $7; } END {print "Rcomplete = ", sumDF/sumFo; }'
Nfree_all.data
```

Conclusions

1. Cross validation important even with data:parameter = 5–10

2. R_{complete} **independent** of test set size k
   - $k = 1$ nearly all reflections available for refinement; calculating R_{complete} can take a day to a week
   - $k = 100$ fast calculation of R_{complete}; less reflections for refinement

3. `WIGL`: **real independence** for free sets
   - $\Rightarrow R_{\text{complete}}$ can be calculated at **any stage**

References

1. A. T. Brünger, *Free R Value: Cross-Validation in Crystallography*; Meth. Enzymol. 1990, Vol. 343, 366–396

2. G. J. Kleywegt & T. A. Jones, *Where freedom is given, liberties are taken*, Structure 1995, 3, 535–540.