

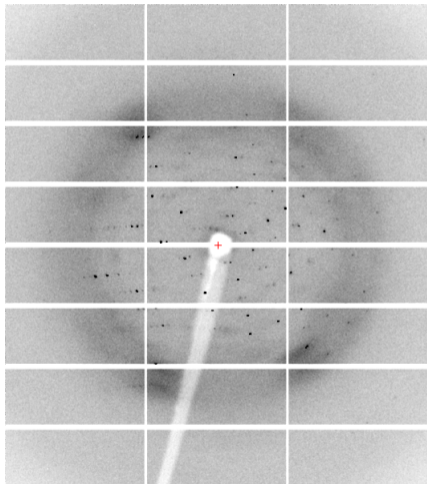
PAIREF: Introduction to paired refinement

Petr Kolenko

5th January 2023

- 1 High-resolution diffraction limit
- 2 Paired refinement with *PAIREF*
- 3 What about the influence on the electron density map?
- 4 FAQ
- 5 PAIREF tutorial

Why do we have to limit the resolution? (make the cut off?)



Aim of the data processing:

- Include as much experimental information as possible.
- Avoid noisy or even bad data.

Cut in resolution is important!

How to make the right decision?

Data quality indicators:

$$R_{\text{merge}} = \frac{\sum_{\mathbf{h}} \sum_i |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \frac{\sum_{\mathbf{h}} \sum_i \sqrt{\frac{n_{\mathbf{h}}}{n_{\mathbf{h}}-1}} |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$CC_{1/2} = \frac{\sum_{m=1}^{N/2} (I_m - \bar{I})(J_m - \bar{J})}{\sqrt{\sum_{m=1}^{N/2} (I_m - \bar{I})^2} \sqrt{\sum_{m=1}^{N/2} (J_m - \bar{J})^2}}$$

Further indicators: $\langle I/\sigma \rangle$, $R_{\text{p.i.m.}}$, completeness, ...

How to make the right decision?

Data quality indicators:

$$R_{\text{merge}} = \frac{\sum_{\mathbf{h}} \sum_i |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \frac{\sum_{\mathbf{h}} \sum_i \sqrt{\frac{n_{\mathbf{h}}}{n_{\mathbf{h}}-1}} |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$\text{CC}_{1/2} = \frac{\sum_{m=1}^{N/2} (I_m - \bar{I})(J_m - \bar{J})}{\sqrt{\sum_{m=1}^{N/2} (I_m - \bar{I})^2} \sqrt{\sum_{m=1}^{N/2} (J_m - \bar{J})^2}}$$

Further indicators: $\langle I/\sigma \rangle$, $R_{\text{p.i.m.}}$, completeness, ...

Criteria used in past:

- $\langle I/\sigma \rangle$ higher than 2
- R_{merge} lower than 0.15
- $R_{\text{merge,HighRes}}$ lower than 0.6
- R_{meas} lower than 0.15
- $\text{CC}_{1/2, \text{HighRes}}$ higher than 0.3
- ...

How to make the right decision?

Data quality indicators:

$$R_{\text{merge}} = \frac{\sum_{\mathbf{h}} \sum_i |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \frac{\sum_{\mathbf{h}} \sum_i \sqrt{\frac{n_{\mathbf{h}}}{n_{\mathbf{h}}-1}} |I_{\mathbf{h}i} - \langle I_{\mathbf{h}} \rangle|}{\sum_{\mathbf{h}} \sum_i \langle I_{\mathbf{h}} \rangle}$$

$$\text{CC}_{1/2} = \frac{\sum_{m=1}^{N/2} (I_m - \bar{I})(J_m - \bar{J})}{\sqrt{\sum_{m=1}^{N/2} (I_m - \bar{I})^2} \sqrt{\sum_{m=1}^{N/2} (J_m - \bar{J})^2}}$$

Further indicators: $\langle I/\sigma \rangle$, $R_{\text{p.i.m.}}$, completeness, ...

Criteria used in past:

- $\langle I/\sigma \rangle$ higher than 2
- R_{merge} lower than 0.15
- $R_{\text{merge,HighRes}}$ lower than 0.6
- R_{meas} lower than 0.15
- $\text{CC}_{1/2, \text{HighRes}}$ higher than 0.3
- ...

Is application of these criteria useful?

How to make the right decision?

Data quality indicators:

$$R_{\text{merge}} = \frac{\sum_h \sum_i |I_{hi} - \langle I_h \rangle|}{\sum_h \sum_i \langle I_h \rangle}$$

$$R_{\text{meas}} = R_{\text{r.i.m.}} = \frac{\sum_h \sum_i \sqrt{\frac{n_h}{n_h - 1}} |I_{hi} - \langle I_h \rangle|}{\sum_h \sum_i \langle I_h \rangle}$$

$$CC_{1/2} = \frac{\sum_{m=1}^{N/2} (I_m - \bar{I})(J_m - \bar{J})}{\sqrt{\sum_{m=1}^{N/2} (I_m - \bar{I})^2} \sqrt{\sum_{m=1}^{N/2} (J_m - \bar{J})^2}}$$

Further indicators: $\langle I/\sigma \rangle$, $R_{\text{p.i.m.}}$, completeness, ...

Criteria used in past:

- $\langle I/\sigma \rangle$ higher than 2
- R_{merge} lower than 0.15
- $R_{\text{merge,HighRes}}$ lower than 0.6
- R_{meas} lower than 0.15
- $CC_{1/2,HighRes}$ higher than 0.3
- ...

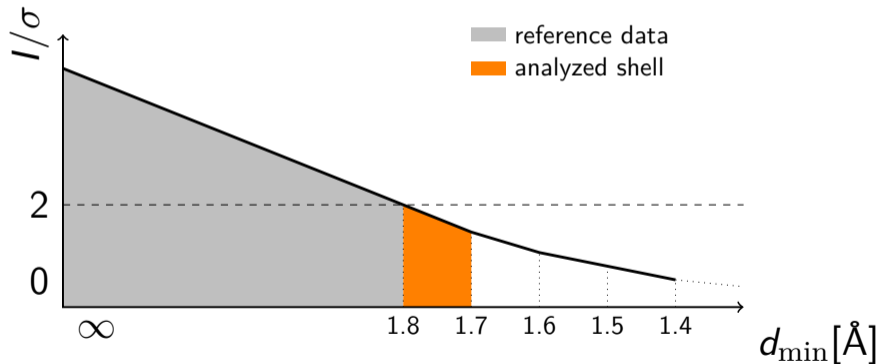
Is application of these criteria useful?

Yes, but only as a starting point.

Paired refinement requires reference data!

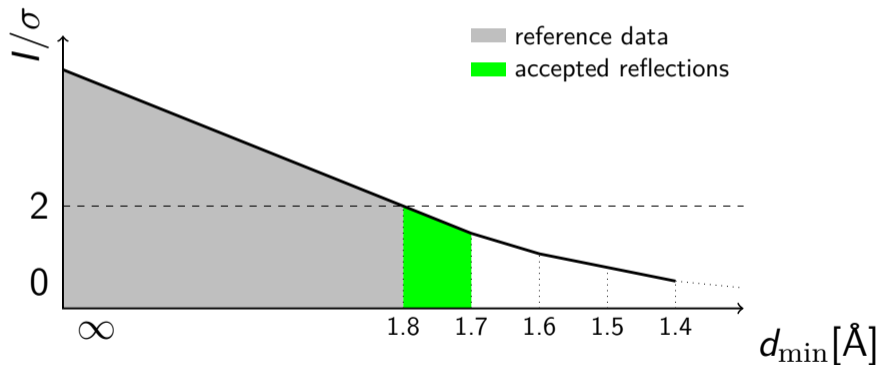
Paired refinement requires reference data!

Does addition of new reflections improve my model?



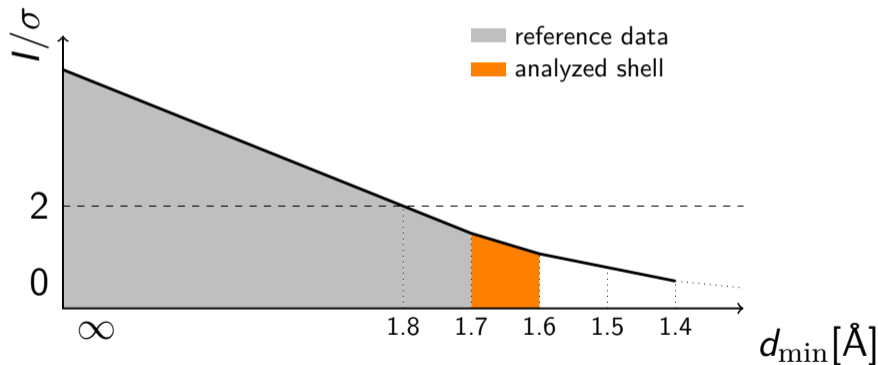
Paired refinement - the (current) optimal way

Does addition of new reflections improve my model? **Yes!**



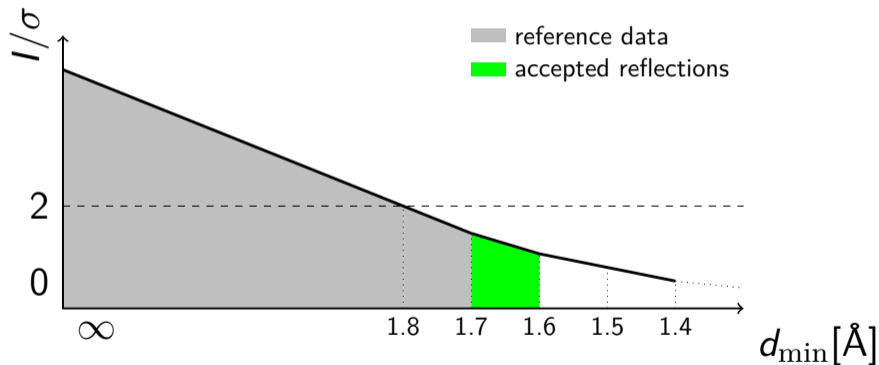
Paired refinement - the (current) optimal way

Does addition of new reflections improve my model?



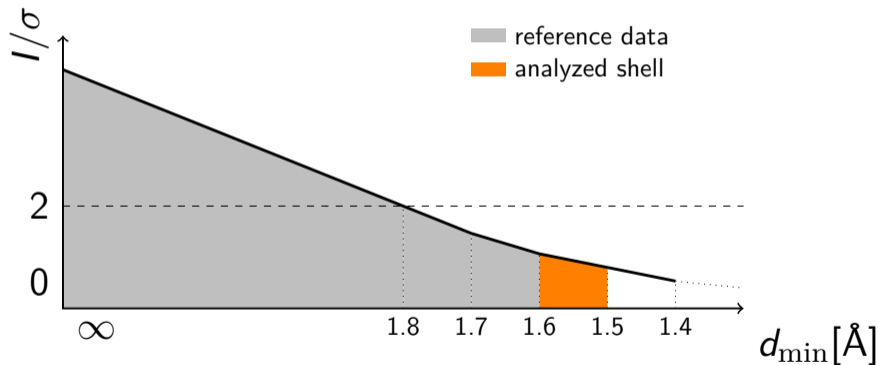
Paired refinement - the (current) optimal way

Does addition of new reflections improve my model? **Yes!**



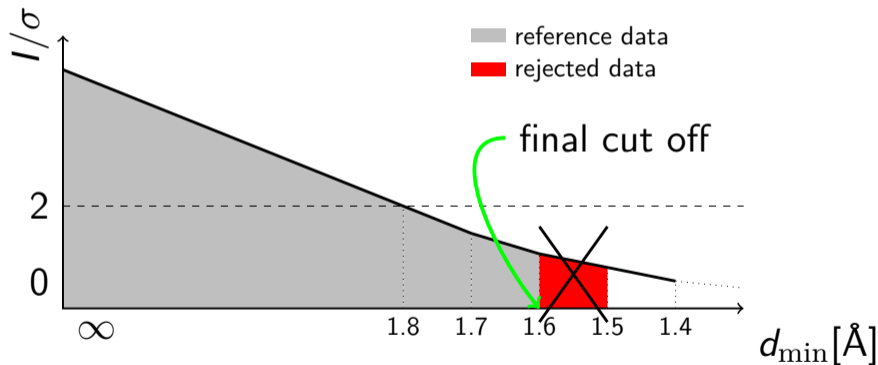
Paired refinement - the (current) optimal way

Does addition of new reflections improve my model?



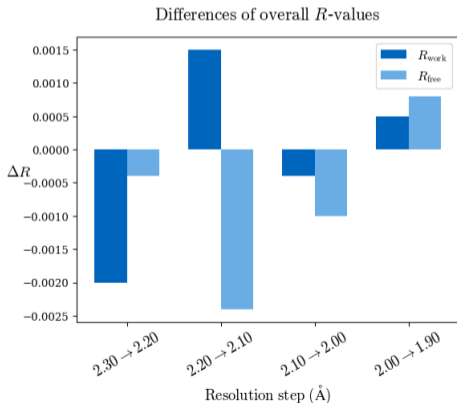
Paired refinement - the (current) optimal way

Does addition of new reflections improve my model? **No!**



How do I know that the model was improved?

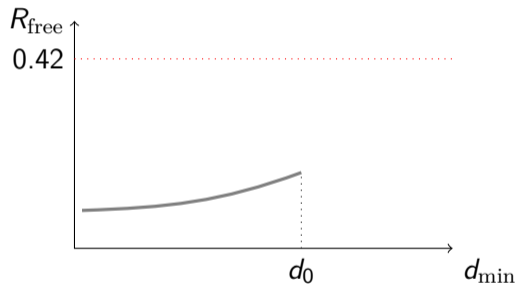
The R -values must be calculated always using the reflections from the reference data!



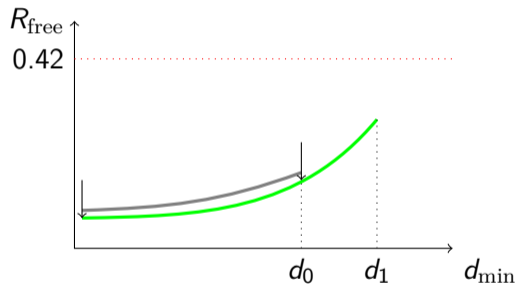
Each run is supplemented with data statistics:

d_max	d_min	#obs	#uniq	mult.	%comp	<I/sI>	r_mrg	r_meas	r_pim	cc1/2	cc*
47.05	5.13	22367	1948	11.48	99.44	38.8	0.042	0.044	0.013	0.999	0.9997
5.13	3.63	42232	3343	12.63	99.79	31.7	0.064	0.067	0.019	0.999	0.9997
3.63	2.97	57138	4246	13.46	99.95	14.3	0.151	0.157	0.042	0.997	0.9992
2.97	2.57	65007	5014	12.97	99.70	4.0	0.622	0.648	0.178	0.951	0.9874
2.57	2.30	73887	5623	13.14	99.89	1.3	1.839	1.914	0.523	0.730	0.9187
2.30	2.20	37405	2818	13.27	99.93	0.6	4.462	4.641	1.263	0.400	0.7559
2.20	2.10	45724	3363	13.60	99.91	0.3	7.623	7.920	2.129	0.196	0.5725
2.10	2.00	49772	4021	12.38	98.51	0.1	16.989	17.721	4.963	0.027	0.2293
2.00	1.90	35920	4046	8.88	81.18	0.0	41.435	43.993	14.417	-0.132	N/A

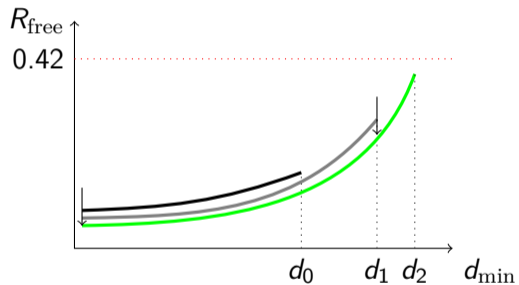
R -values in resolution bins (monitor FREE reflections!)



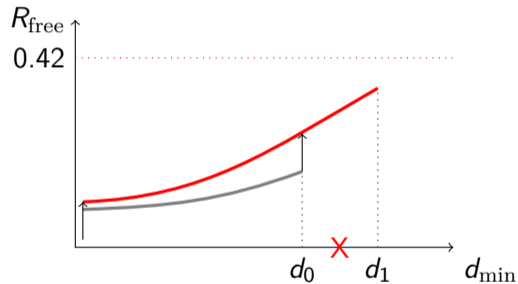
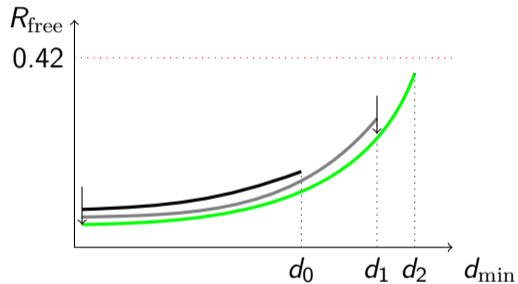
R -values in resolution bins (monitor FREE reflections!)



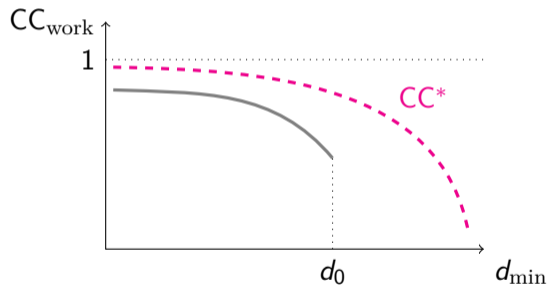
R -values in resolution bins (monitor FREE reflections!)



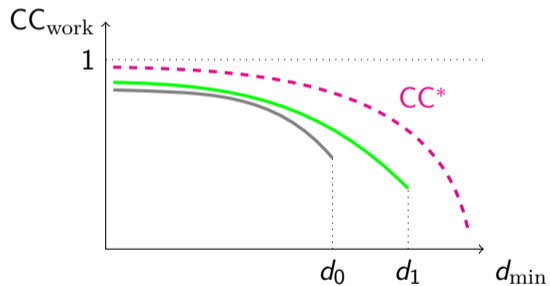
R-values in resolution bins (monitor FREE reflections!)



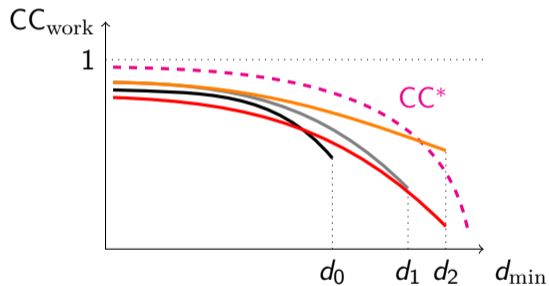
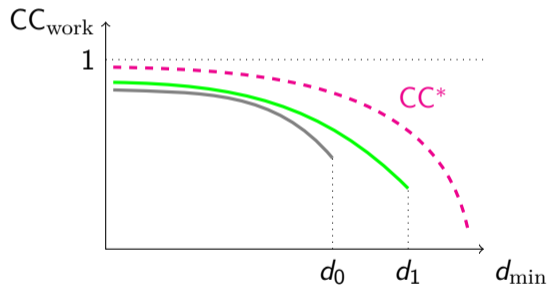
CC_{work} in resolution bins (monitor WORK reflections!)



CC_{work} in resolution bins (monitor WORK reflections!)



CC_{work} in resolution bins (monitor WORK reflections!)

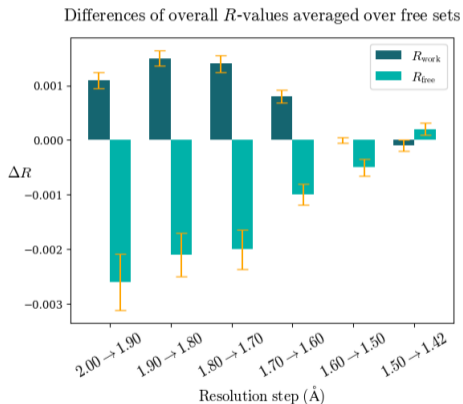


Suggestions from PAIREF procedures

Benevolent vs. strict verdict ...

Shell	Accepted?	Reason
2.30-2.10 Å	Yes	Overall R _{free} decreased while using data in the shell 2.30-2.10 Å
2.10-2.00 Å	Warning	Overall R _{free} decreased while using data in the shell 2.10-2.00 Å R _{free} in high resolution is higher than 0.40 while using data in the shell 2.10-2.00 Å R _{work} in high resolution is higher than 0.40 while using data in the shell 2.10-2.00 Å
2.00-1.90 Å	No	Overall R _{free} decreased while using data in the shell 2.00-1.90 Å R _{free} in high resolution is higher than 0.40 while using data in the shell 2.00-1.90 Å R _{work} in high resolution is higher than 0.40 while using data in the shell 2.00-1.90 Å CC1/2 in high resolution is negative or undefined while using data in the shell 2.00-1.90 Å But statistics deteriorate in a previous resolution shell.

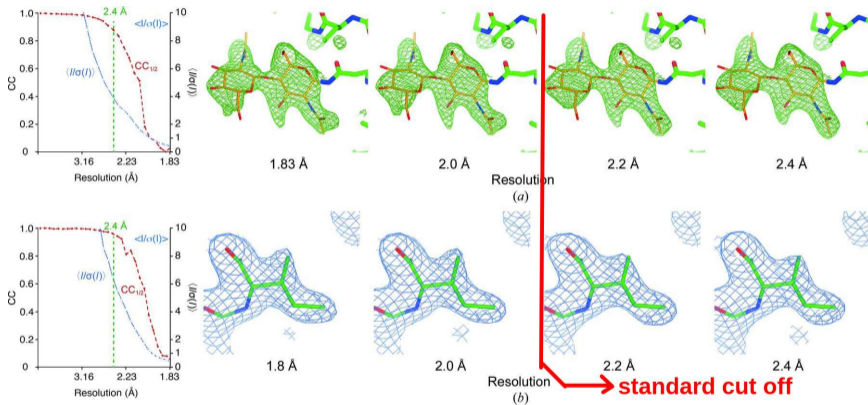
Complete cross-validation procedure in PAIREF



Complete cross-validation procedure:

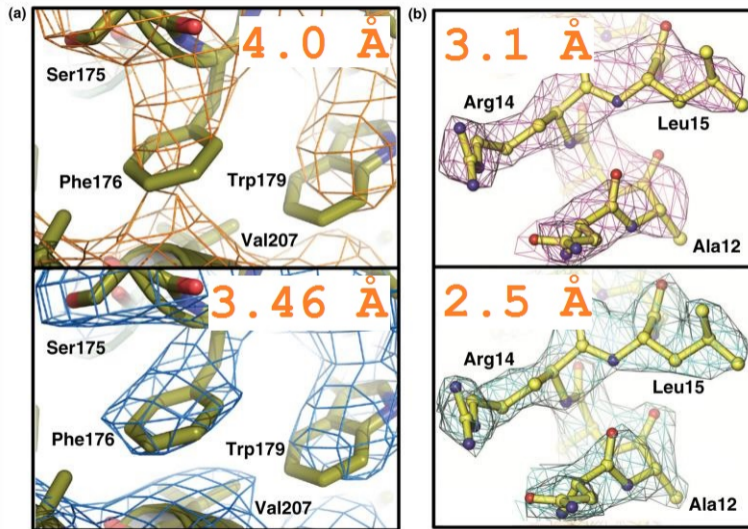
- parallel paired refinement protocol for all free reflections selection
- averaging of the results
- removal of free reflections selection bias
 - shaking of coordinates
 - modification of ADPs

Are there any changes in the electron density?



P.R. Evans, G.N. Murshudov. (2013). How good are my data and what is the resolution?
*Acta Cryst D***69**, 1204-1214.

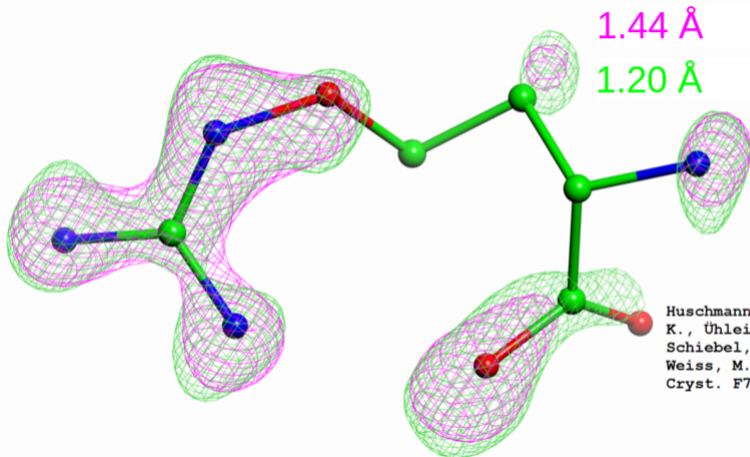
Are there any changes in the electron density?



Current Opinion in Structural Biology

Karplus, P. A. & Diederichs, K. (2015). *Curr. Opin. Struct. Biol.* 34, 60–68.

Is there anything practical?



Huschmann, F. U., Linnik, J., Sparta, K., Ühlein, M., Wang, X., Metz, A., Schiebel, J., Heine, A., Klebe, G., Weiss, M. S. & Mueller, U. (2016). *Acta Cryst.* F72, 346-355.

Frequently asked questions:

- *So what is the right value for I/σ (or $CC1/2$) cutoff?*
There is none! In our data, from 0.3 to 1.7 (or from 0.027 to 0.652) or even higher.
- *In which stage of structure refinement should I run PAIREF?*
It can be run earlier, but we recommend to run PAIREF during the final stages.
- *May I use PAIREF with other refinement programs?*
Yes, also with *phenix.refine*. Please, use our own GUI for that.
- *When will it be available for CCP4 users?*
Already distributed as a binary file together with our GUI. i2 interface will be soon (008 update?).
- *What would I do if I had some recommendation?*
Just send an email to [petr.kolenko \[at\] fjfi.cvut.cz](mailto:petr.kolenko@fjfi.cvut.cz) or [martin.maly.mm \[at\] email.cz](mailto:martin.maly.mm@email.cz)

References:

- M. Malý, K. Diederichs, J. Dohnálek, P. Kolenko. (2020). Paired refinement under the control of PAIREF. *IUCrJ*, **7**, 681-692.
- M. Malý, K. Diederichs, J. Dohnálek, P. Kolenko. (2021). *PAIREF*: paired refinement also for Phenix users. *Acta Cryst*, **F77**, 226-229.
- P.A. Karplus, K. Diederichs. (2012). Linking Crystallographic Model and Data Quality. *Science*, **336**, 1030-1033.
- P.A. Karplus, K. Diederichs. (2015). Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.*, **34**, 60-68.
- P.R. Evans, G.N. Murshudov. (2013). How good are my data and what is the resolution? *Acta Cryst D***69**, 1204-1214.
- K. Diederichs, P.A. Karplus. (2013). Better models by discarding data? *Acta Cryst D***69**, 1215-1222.

- Martin Malý
- Kay Diederichs
- Jan Dohnálek

Some people, who helped us without actually knowing that:

- Andrew Karplus
- James Holton
- Manfred Weiss
- Jan Wollenhaupt
- ...



`pairef.fjfi.cvut.cz`

`https://pairef.fjfi.cvut.cz/dokuwiki/doku.php?id=lunchtime_byte_2023`