Informative priors in scaling & merging

Doeke Hekstra

February 22, 2021

Aim: Using informative priors in scaling & merging

For time-resolved crystallography data, we have two kinds of useful information

- High-quality synchrotron reference data, E_{ref}
- Knowledge that E_{on} and E_{off} tend to be highly correlated

$$
\begin{tikzcd}
P(E_1)P(E_2) \arrow[r, "\text{Correlation w\textbackslash\ reference}"] \arrow[d, "\text{Correlation among on, off}"'] & P\left(E_1|E_{ref}\right)P(E_2|E_{ref}) \arrow[d] \\
P(E_1, E_2) \arrow[r] & P(E_1, E_2|E_{ref})
\end{tikzcd}
$$

Aim: Using informative priors in scaling & merging

For time-resolved crystallography data, we have two kinds of useful information

- High-quality synchrotron reference data, E_{ref}
- Knowledge that E_{on} and E_{off} tend to be highly correlated

For complex structure factors, all of these correlations are easily expressed as extensions of the Wilson model. For example,

$$P(E_1, E_2, E_3) = P\left(E_{1x}, E_{2x}, E_{3x}, E_{1y}, E_{2y}, E_{3y}\right) = N(0, C)$$

$$C = \frac{1}{2}\begin{bmatrix} 1 & r_x & r & 0 & 0 & 0 \\ r_x & 1 & r & 0 & 0 & 0 \\ r & r & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & r_x & r \\ 0 & 0 & 0 & r_x & 1 & r \\ 0 & 0 & 0 & r & r & 1 \end{bmatrix}$$

Aim: Using informative priors in scaling & merging

For time-resolved crystallography data, we have two kinds of useful information

- High-quality synchrotron reference data, E_{ref}
- Knowledge that E_{on} and E_{off} tend to be highly correlated

For complex structure factors, all of these correlations are easily expressed as extensions of the Wilson model. For example,

$$P(E_1, E_2 | E_3) = P\left(E_{1x}, E_{2x}, E_{1y}, E_{2x} | E_{3x}, E_{3y}\right) = N\left(rE_3, C_{1,2|3}\right)$$

$$C_{1,2|3} = \frac{1}{2}\begin{bmatrix} 1-r^2 & r_x - r^2 & 0 & 0 \\ r_x - r^2 & 1-r^2 & 0 & 0 \\ 0 & 0 & 1-r^2 & r_x - r^2 \\ 0 & 0 & r_x - r^2 & 1-r^2 \end{bmatrix}$$

Aim: Using informative priors in scaling & merging

For time-resolved crystallography data, we have two kinds of useful information

- High-quality synchrotron reference data, E_{ref}
- Knowledge that E_{on} and E_{off} tend to be highly correlated

For complex structure factors, all of these correlations are easily expressed as extensions of the Wilson model. But, marginalizing to structure factor amplitudes is a hassle.

Acentric

- $P(E_1) \sim Wilson$
- $P(E_1|E_2) \sim Rice$
- $P\left(E_1\middle|E_{ref}\right) \sim Rice$
- $P(|E_1|, |E_2|), P(|E_1|, |E_2| \mid E_{ref})$?

$$P(|E_1|^2, |E_2|^2 \mid E_{ref}) \sim \text{Bivariate Non-central } \chi^2$$

https://www.researchgate.net/publication/220557408

Aim: Using informative priors in scaling & merging

For now:

1. Normalization for better E_{ref}
2. Suitability of the Rice and Folded-Normal distributions

Pending

1. Use of the $\text{Bivariate Non}-\text{central }\chi^2$ distribution (implemented*, but don't know how to pick parameters).

Acentric

- $P(E_1) \sim Wilson$
- $P(E_1|E_2) \sim Rice$
- $P\left(E_1\middle|E_{ref}\right) \sim Rice$
- $P(|E_1|, |E_2|)$ tbd
- $P(|E_1|, |E_2| \mid E_{ref})$?

$$P(|E_1|^2, |E_2|^2 \mid E_{ref}) \sim \text{Bivariate Non-central } \chi^2$$

* https://www.researchgate.net/publication/220557408

Normalizing structure factors

Three steps

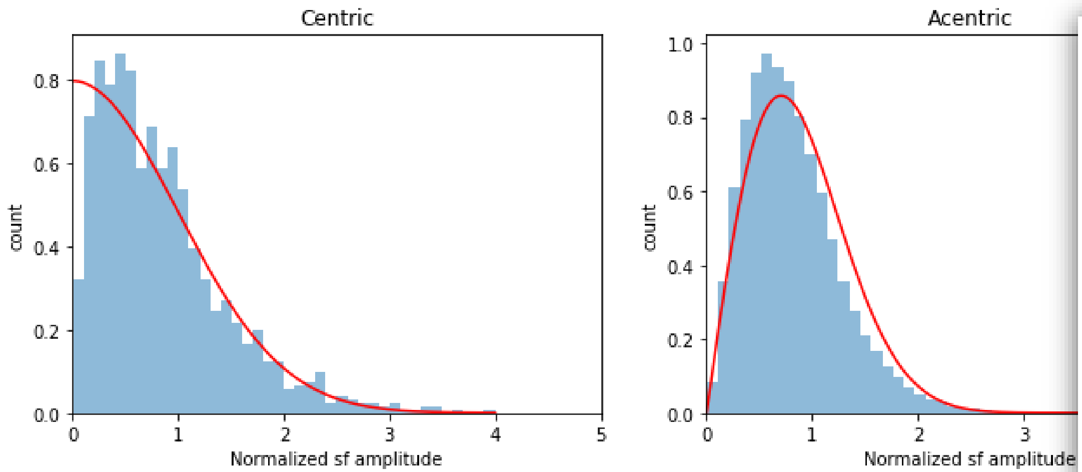
1. Fit $\boxed{E_h = f e^{-s^T B s} \frac{F_h}{\sqrt{\varepsilon}},}$ using the Wilson distributions as the loss function, with $s = 1/d_h$ and h short for (h, k, l).

2. Fit $\boxed{E_h' = E_h\left(\sum_m A_m \cos\left(2\pi\tilde{h}\cdot m\right) + B_m \sin\left(2\pi\tilde{h}\cdot m\right)\right),}$ with $\tilde{h} = \left(\frac{h}{N_h}, \frac{k}{N_k}, \frac{l}{N_l}\right)$ and m short for (m, n, p), and the same loss function for increasing $m_{max} = n_{max} = p_{max}$. Pick best Fourier order by cross-validation.

3. Perform k-nearest neighbor regression on the E'. Obtains Σ as the local estimate of $\langle E'^2\rangle$. Then $E_{knn} = E'/\sqrt{\Sigma}$.

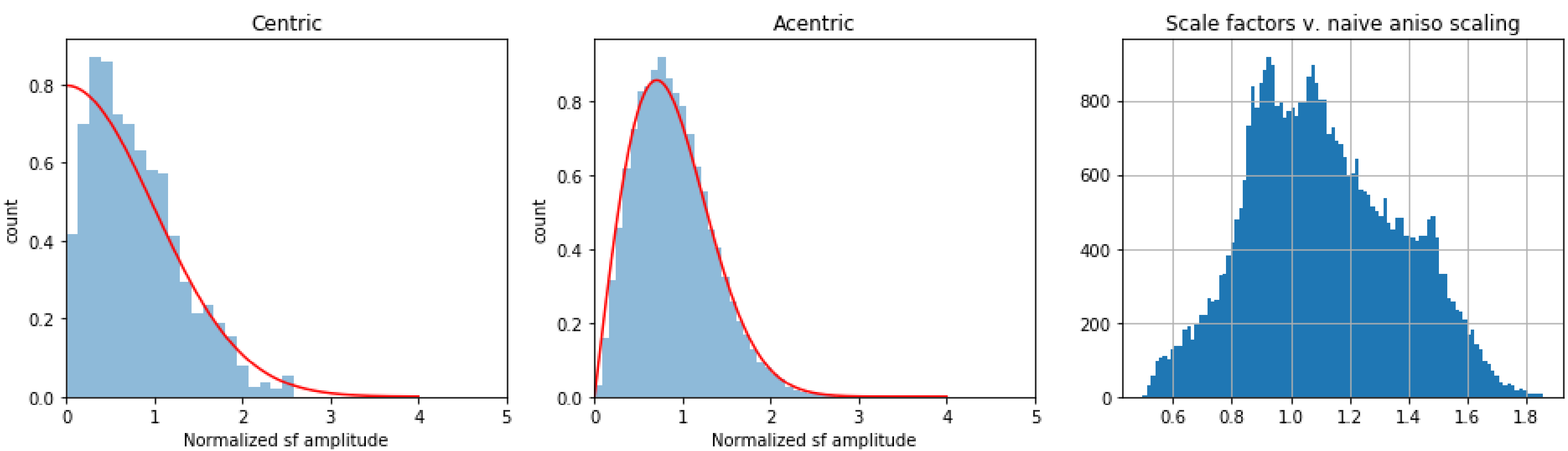
Example: normalizing 1OTB

After simple anisotropic scaling of 1OTB:



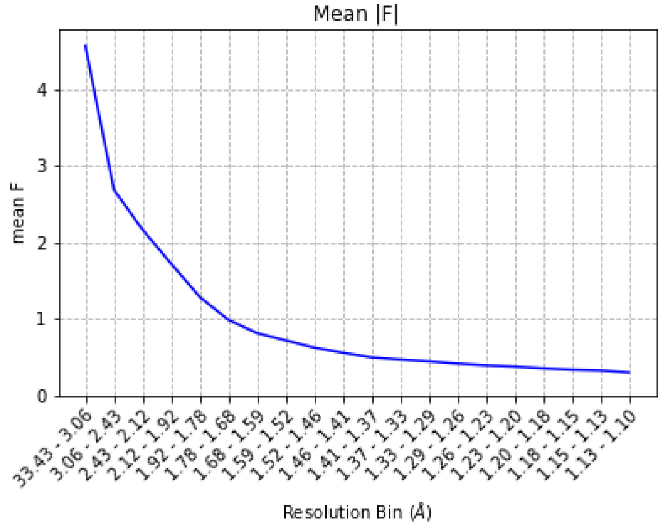

```
For n = 1 the test loss = 6180.78
Elapsed time: 3.284 s
For n = 2 the test loss = 5248.84
Elapsed time: 5.117 s
For n = 3 the test loss = 5125.22
Elapsed time: 37.18 s
For n = 4 the test loss = 5075.84
Elapsed time: 183.1 s
For n = 5 the test loss = 5083.3
Elapsed time: 896.2 s
```

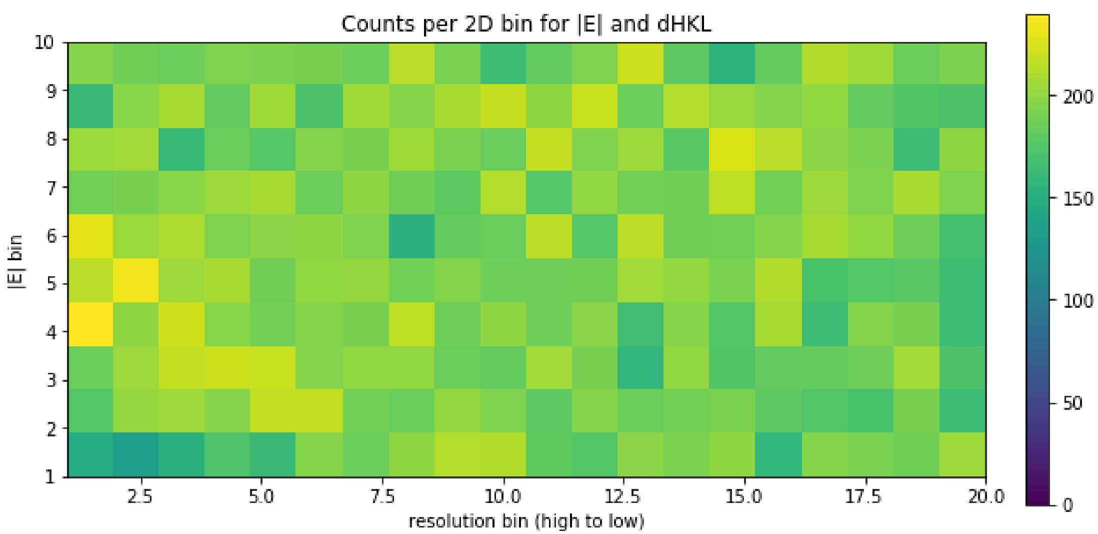
After anisotropic scaling with Fourier corrections of 1OTB:



Example: normalizing 1OTB

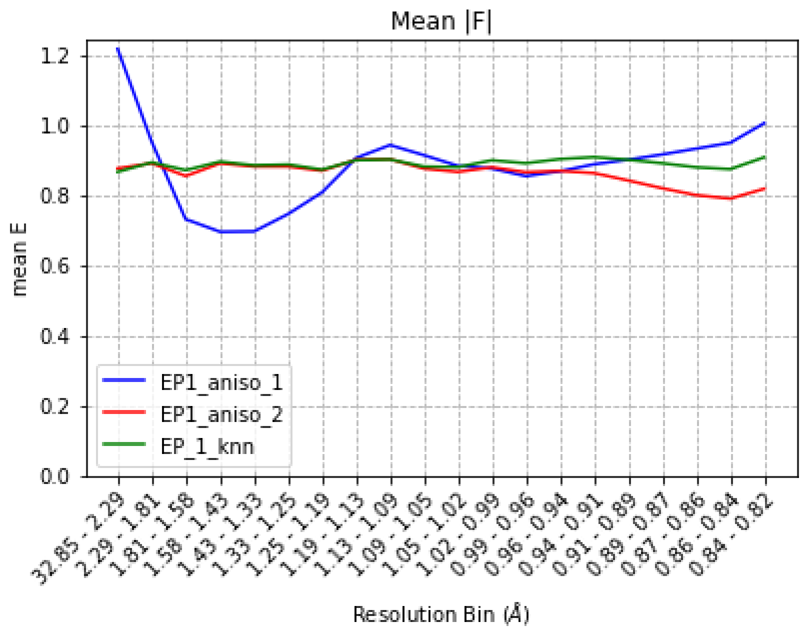
Naïve structure factor amplitudes:

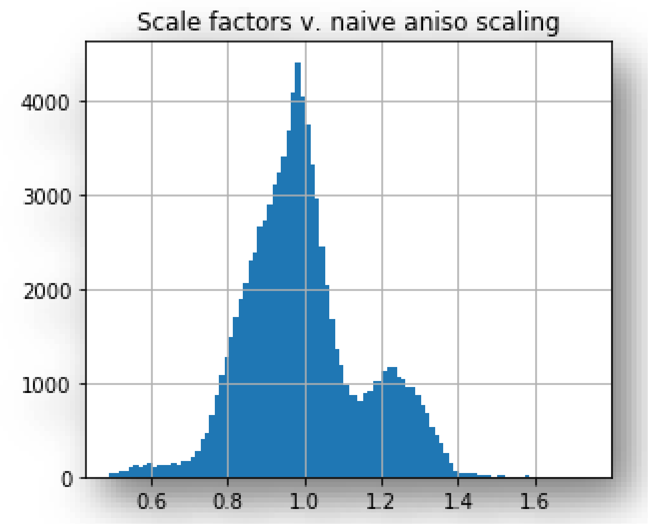


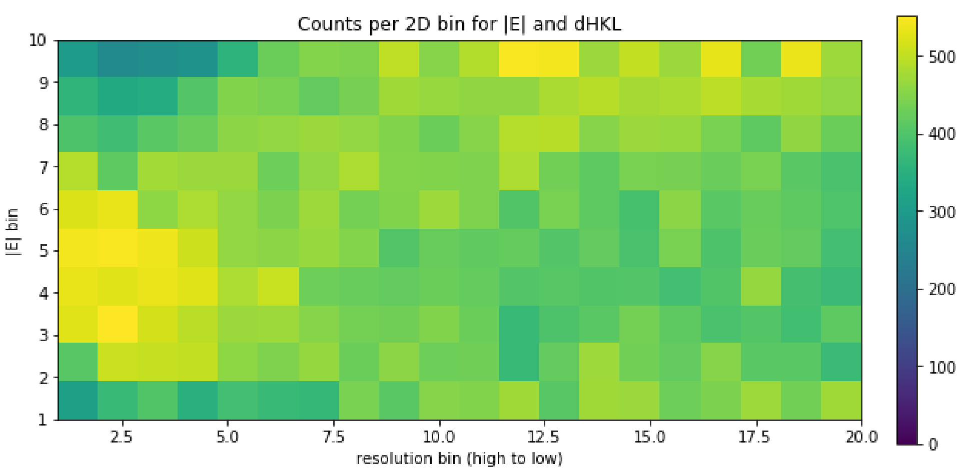


Example: normalizing 1NWZ

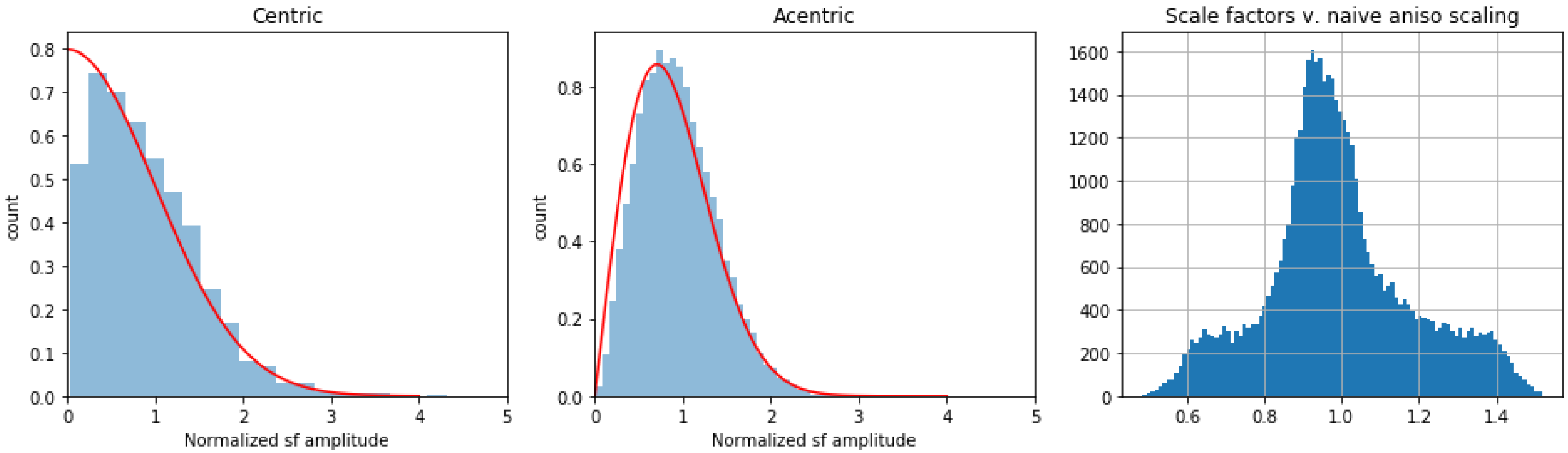
After anisotropic scaling with Fourier corrections of 1NWZ:

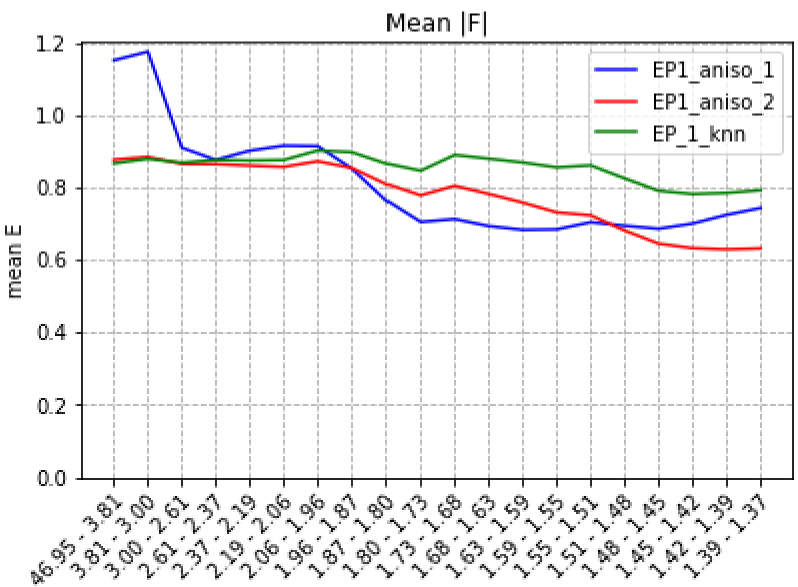




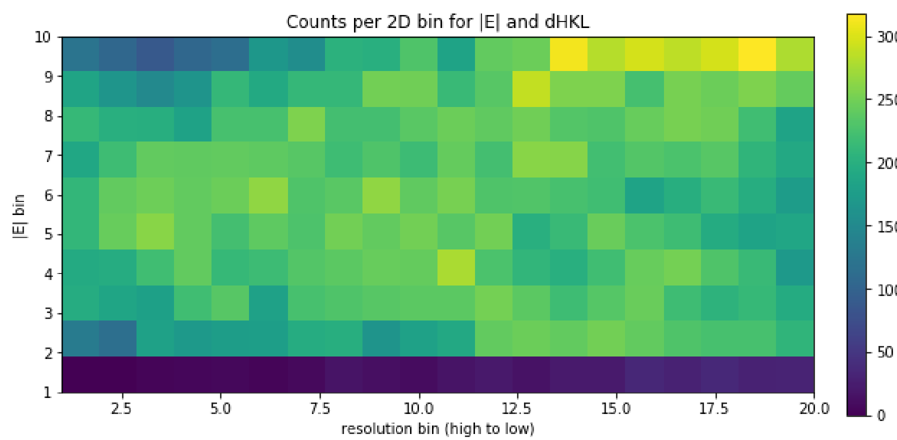


A troublemaker: **GFP@RT**





This dataset (here for *n* = 4) did not scale very well. *k*-NN may be more appropriate. Perhaps reflects truncation approach…



Normalizing HEWL anomalous

```
# For simplicity...
ds1 = ds1[(ds1["I(+)"]>=0) & (ds1["I(-)"]>=0)]
```

1A_Anom_dataset_prep_and_scaling

Normalizing HEWL anomalous

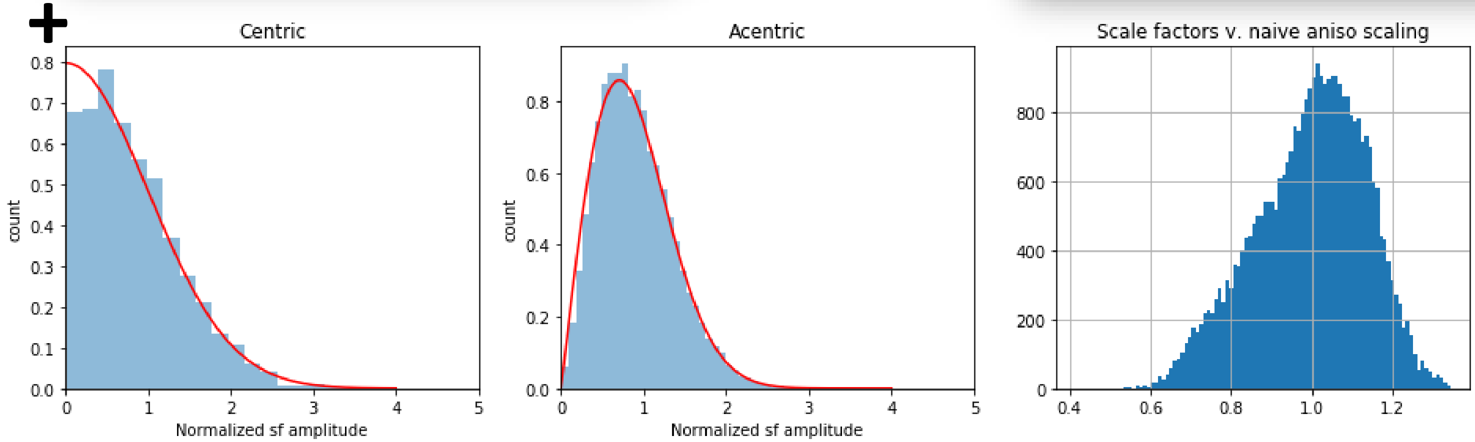
+

```
For n = 1 the test loss = 5849.07
Elapsed time: 0.4382 s
For n = 2 the test loss = 5667.69
Elapsed time: 4.897 s
For n = 3 the test loss = 5628.02
Elapsed time: 37.26 s
For n = 4 the test loss = 5580.48
Elapsed time: 176.7 s
For n = 5 the test loss = 5568.9
Elapsed time: 898.9 s
```

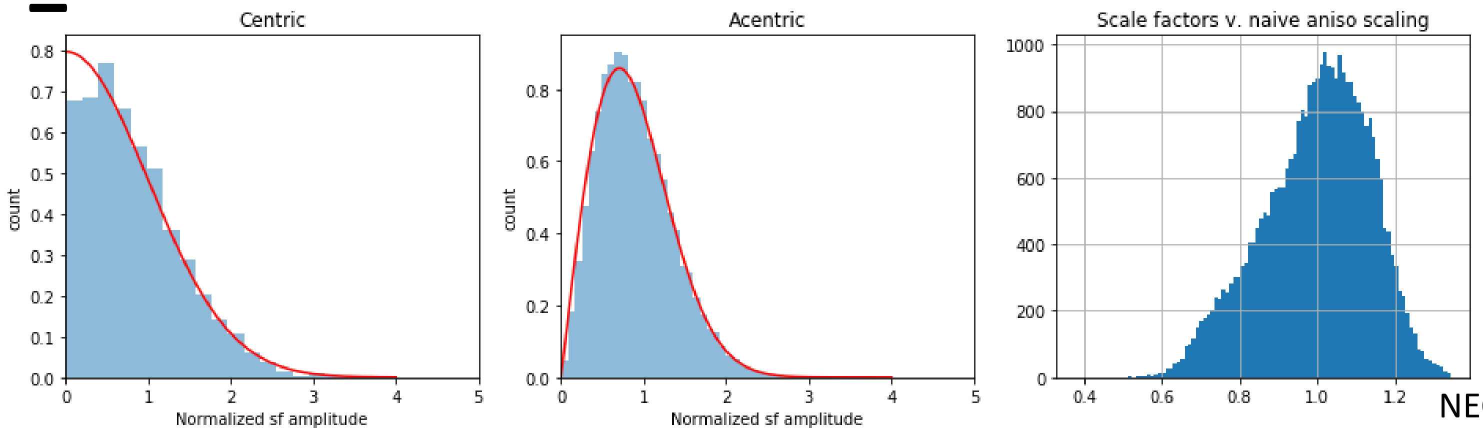
−

```
For n = 1 the test loss = 5840.85
Elapsed time: 0.2482 s
For n = 2 the test loss = 5657.45
Elapsed time: 4.063 s
For n = 3 the test loss = 5619.43
Elapsed time: 28.2 s
For n = 4 the test loss = 5574.26
Elapsed time: 155.6 s
For n = 5 the test loss = 5565.04
Elapsed time: 830.5 s
```

+

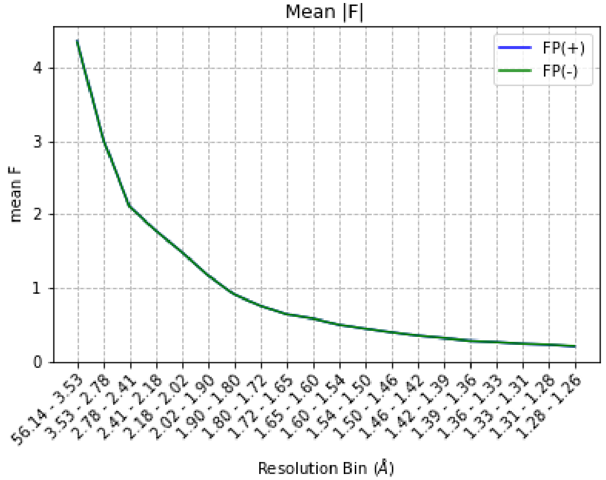


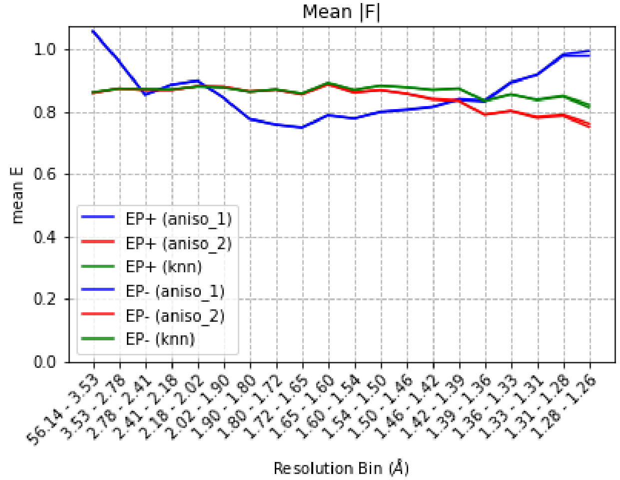
−

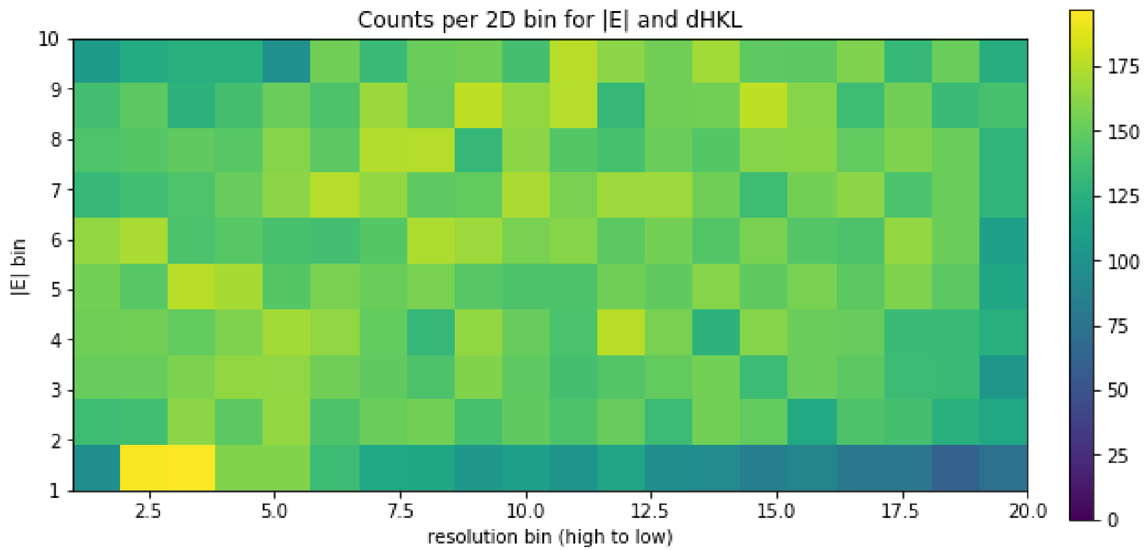


NECAT_HEWL_RT_NaI_82_XDS

Normalizing HEWL anomalous







1A_Anom_dataset_prep_and_scaling

Double-Wilson model

In the DW model, the real and imaginary components of two data sets are both modeled as correlated random walks:

$$\begin{bmatrix} Re(F^A) \\ Im(F^A) \\ Re(F^B) \\ Im(F^B) \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \frac{1}{2}\Sigma \begin{bmatrix} 1 & 0 & r & 0 \\ 0 & 1 & 0 & r \\ r & 0 & 1 & 0 \\ 0 & r & 0 & 1 \end{bmatrix}\right)$$

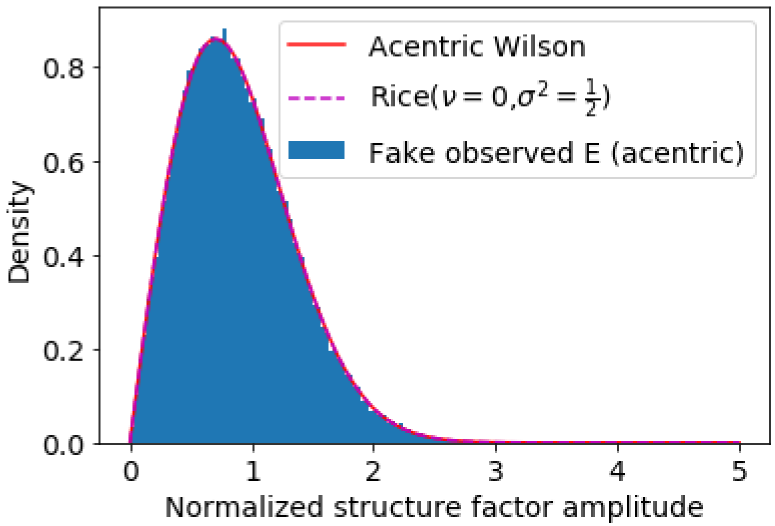
where $\Sigma = \langle |F|^2 \rangle = \mathrm{N}\sigma^2$ for a 2D random walk with N steps each with variance $\frac{1}{2}\sigma^2$ along each dimension. $r = r_{DW}$ governs the correlation between datasets.

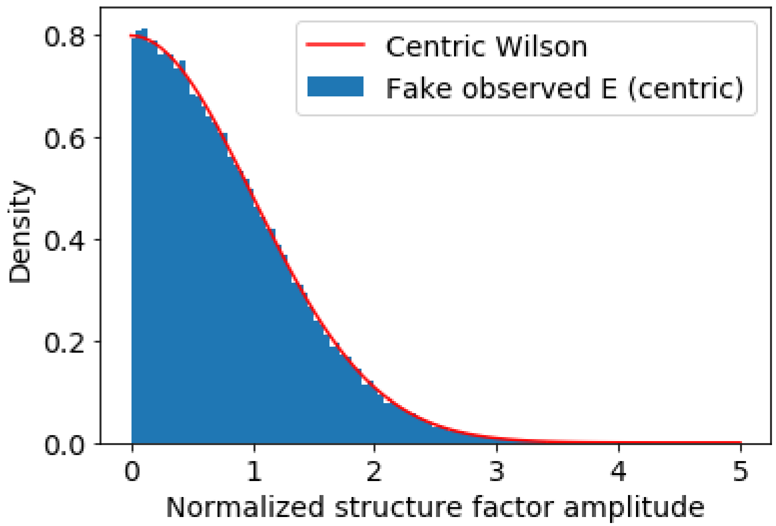
$$\begin{bmatrix} F^A \\ F^B \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}\right)$$

Note that the ½ disappears, because F^A can be thought of as the sum of a random walk in the complex plane added to its own complex conjugate.

Double-Wilson model

The amplitudes of the centric and acentric reflections follow the Wilson distribution:

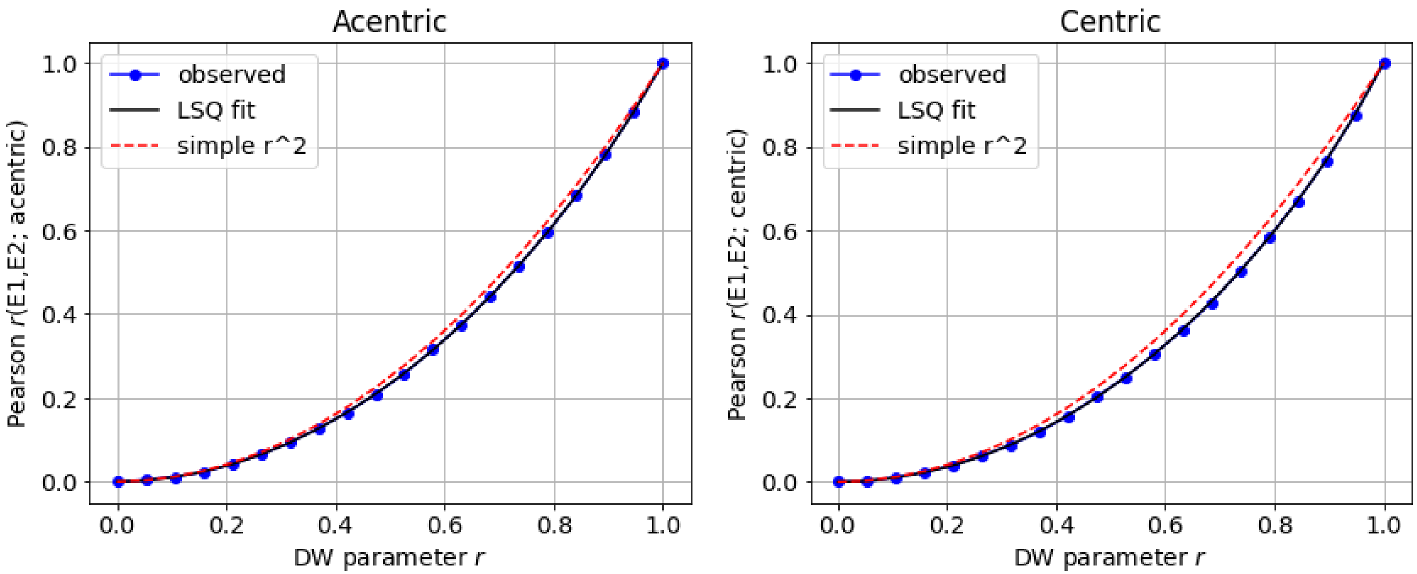




2_Surrogate_data_example

Double-Wilson model

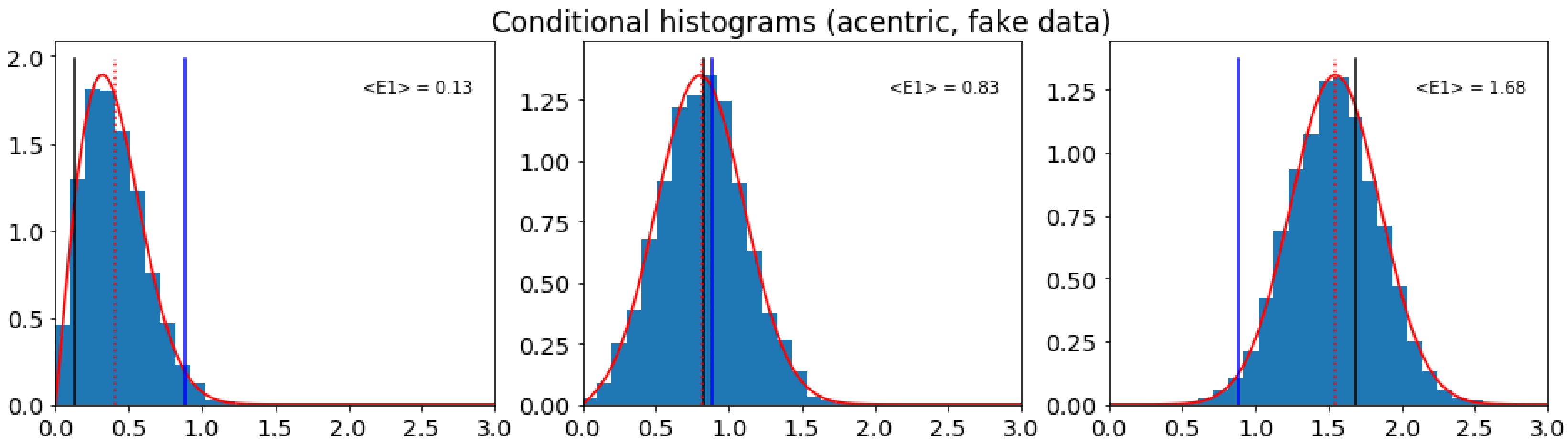
The Pearson correlations between structure factor amplitudes from two correlated data sets almost equal r_{DW}^2.



2_Surrogate_data_example

Double-Wilson model

The conditional distributions of structure factor amplitudes of one data set given the other are described by the Rice distribution (acentric) and Folded Normal (centric).



Conditional mean, $\mathbf{E}(|E_2| \mid |E_1|) = r_{DW}|E_1|$ (centric & acentric)

Conditional variance, $Var(|E_2|||E_1|) = \begin{cases} \frac{1}{2}(1 - r_{DW}^2) & \text{(acentric)} \\ (1 - r_{DW}^2) & \text{(centric)} \end{cases}$

2_Surrogate_data_example

Double-Wilson model

The conditional distributions of structure factor amplitudes of one data set given the other are described by the Rice distribution (acentric) and Folded Normal (centric).

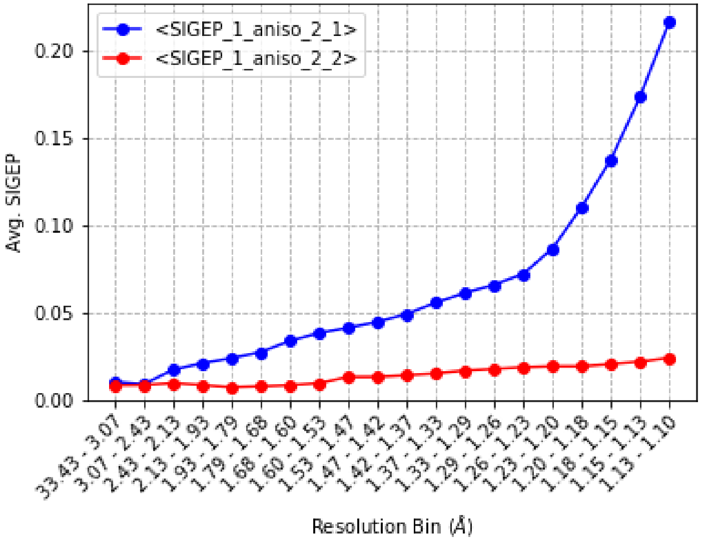
Conditional mean, $\mathbf{E}(|E_2| \mid |E_1|) = r_{DW}|E_1|$ (centric & acentric)

Conditional variance, $Var(|E_2| \mid |E_1|) = \begin{cases} \frac{1}{2}(1 - r_{DW}^2) & \text{(acentric)} \\ (1 - r_{DW}^2) & \text{(centric)} \end{cases}$

```
rice.pdf(    x, cond_mean/np.sqrt(cond_var), scale=np.sqrt(cond_var))
foldnorm.pdf(x, cond_mean/np.sqrt(cond_var), scale=np.sqrt(cond_var))
```

2_Surrogate_data_example

Comparison: 1OTB v 1NWZ



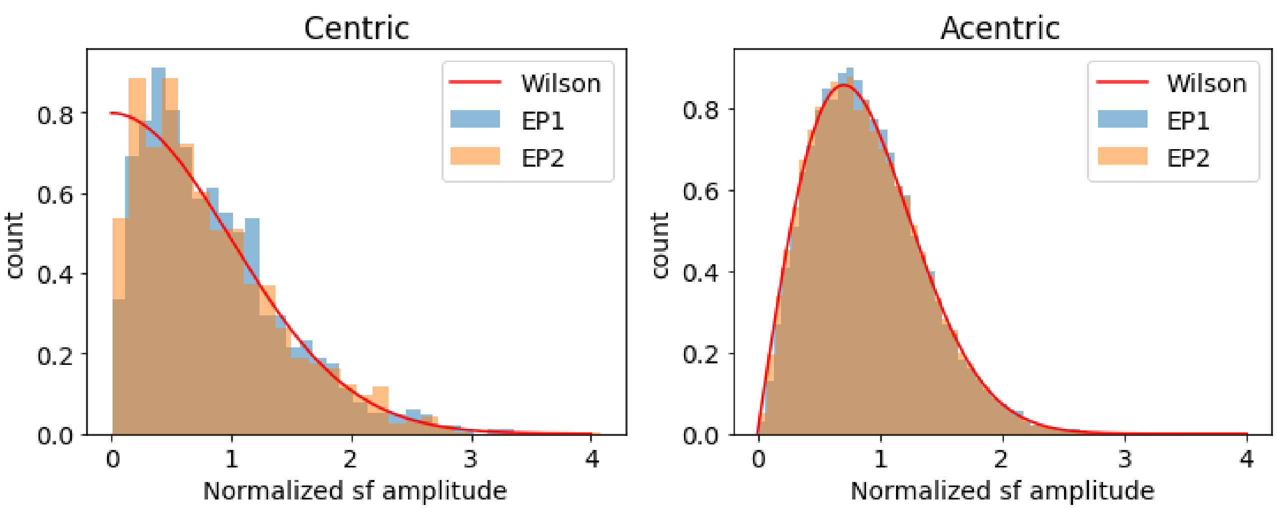
Data quality

Based on Anisotropic + Fourier

For the rest of the analysis,
we’ll cut the datasets to 1.2 Å.

dataset 1: 1OTB

dataset 2: 1NWZ

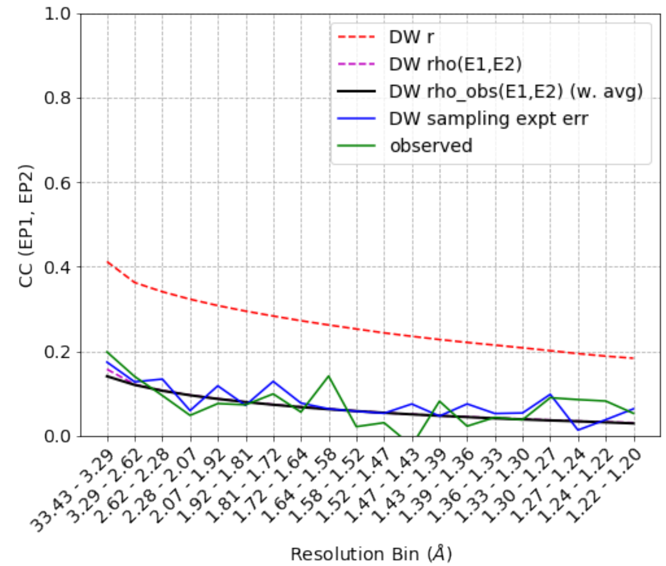


3_Fitting_DW_to_paired_data

Comparison: 1OTB v 1NWZ

```
(a,b) = fitting_dw.fit_ab(ds1_2,labels=[EP1_label,EP2_label],\
                          dHKL_label=dHKL_label, dHKL_bin_label=dHI
print(f"a: {a:.3}")
print(f"b: {b:.3}")
```

```
`ftol` termination condition is satisfied.
Function evaluations 9, initial cost 2.6541e-02, final cost 9.907
8e-03, first-order optimality 2.06e-08.
a: 0.388
b: 0.887
```



$$r_{DW} = a \cdot e^{-bs^2}$$

This is the inferred true r_{DW} after correcting for measurement error.

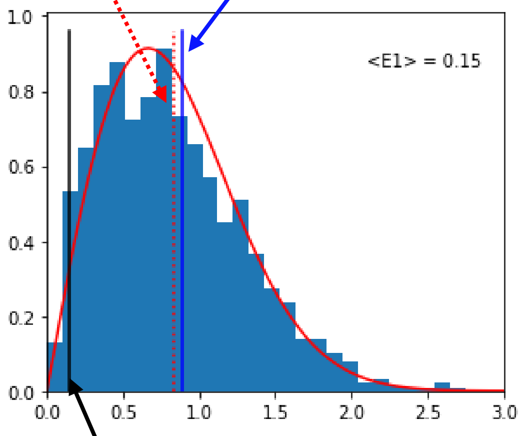
Note that the effective r_{DW} produced by **`fitting_dw.eff_r_dw_per_hkl`** is lower as it includes measurement error

2_Surrogate_data_example
3_Fitting_DW_to_paired_data

Comparison: 1OTB v 1NWZ

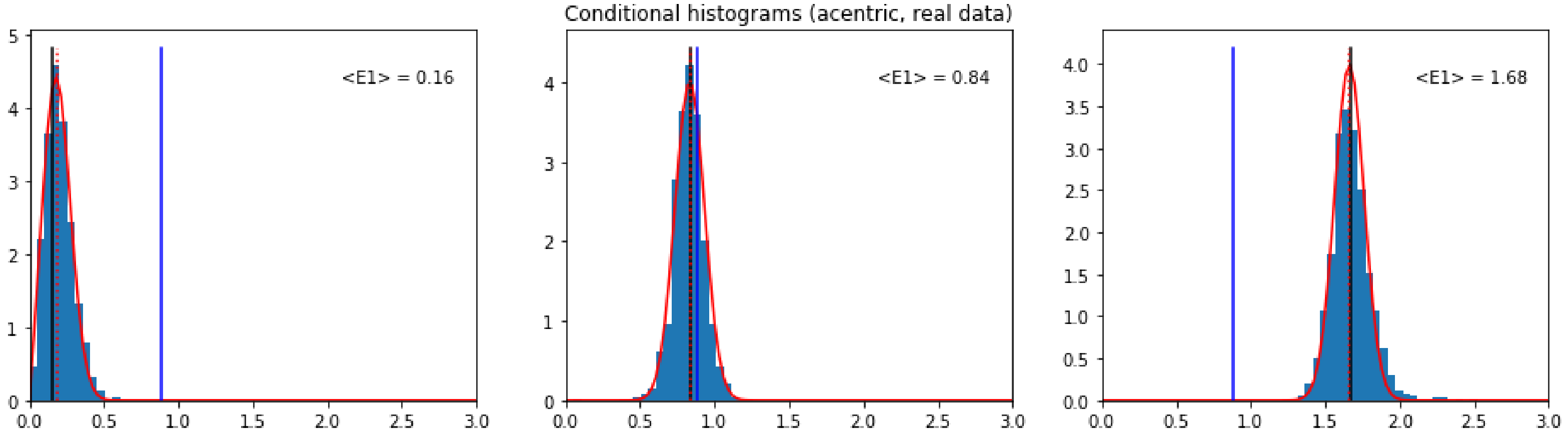
At this low r_{DW}, the changes in prior are already quite small!

Rice mean Wilson mean



Mean $|E_1|$ for a bin with $\sim$1,200 smallest $|E_1|$

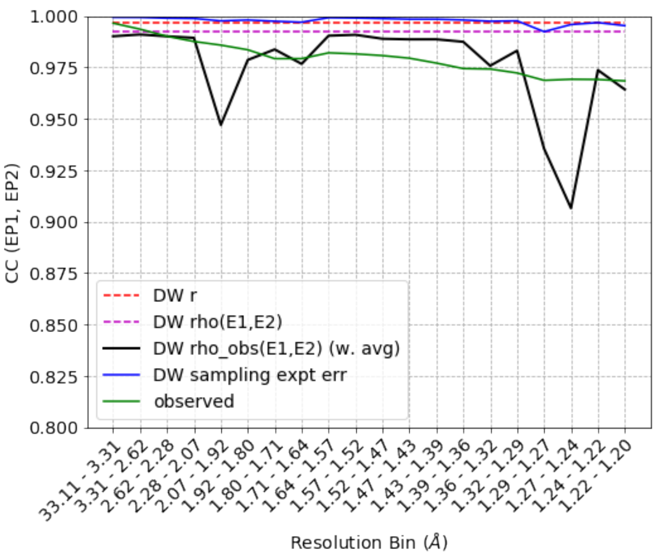
Comparison: **3PYP** v 1NWZ



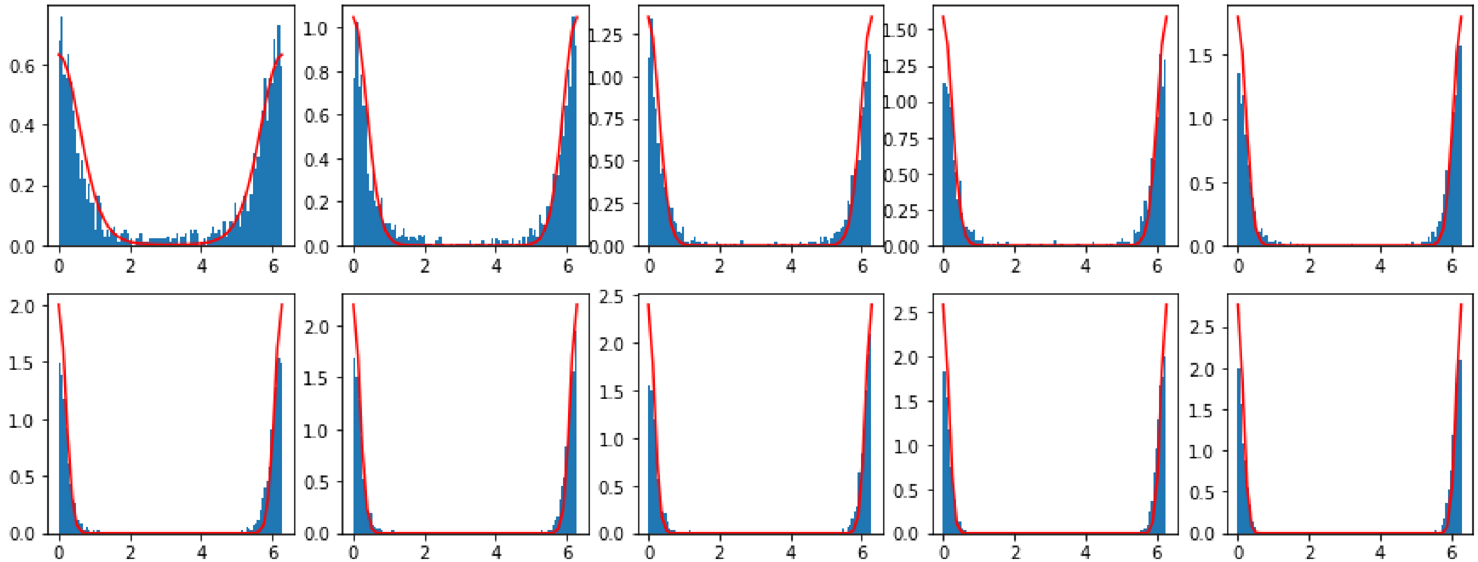
In this case, the priors are highly informative!
(shown for $r_{DW} = 0.99$)

In this case, the experimental errors dominate the correlation coefficient.

Variability in the black line suggests that expt error estimates are conservative, but not perfect.



Comparison: **3PYP** v 1NWZ



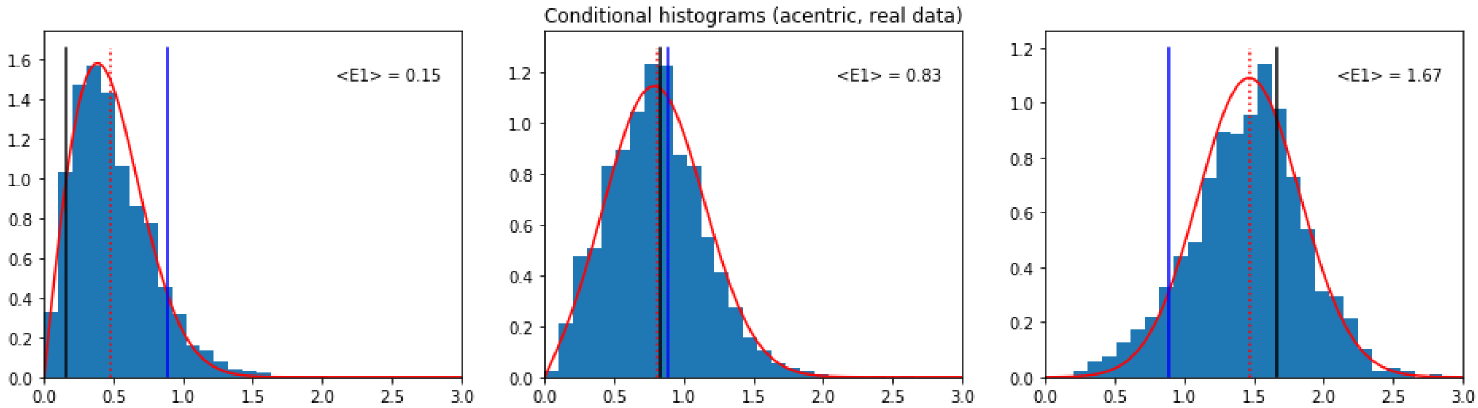
Phase differences are well-described by the corresponding Von Mises distribution.
(shown here for the 10 smallest bins of $|E_1| \cdot |E_2|$)
Red fit for $r_{DW} = 0.99$.

```
vonmises.pdf(x, E1*E2/cond_var)
```

3_Fitting_DW_to_paired_data

Comparison: DHFR (RT, cryo)

	4KJK	4KJJ	4PST	4PSS	
4KJK	1,0	0.90, 0.43	0.95, 0.01	0.88, 0.51	2013 RT
4KJJ		1,0	0.88, 0.26	0.93, 0.16	2013 cryo
4PST			1,0	0.89, 0.22	2005 RT
4PSS				1,0	2015 cryo



fit using $r_{DW} = 0.85$

Keedy, *Structure*, 2013
https://www.sciencedirect.com/science/article/pii/S0969212614001403

4KJK, 4KJJ cut to 1.5Å
all else cut to 1.2Å

Summary of (*a*, *b*) estimates

n	Dataset pair	*a*	*b*	Res. range	Details
0	(5KVX, 5KW3)	0.94	0.79	Cut to 1.7Å	Thaumatin (100K, 278K)
1	(2VWR, 5E1Y)	0.93	0.15	To 1.35Å	LNX2/PDZ2 (100K, 277K)
2	(3PYP, 1NWZ)	1.00	0.00	Cut to 1.1Å	PYP (cryotrapped lit, dark, both 100K)
3	(1NWZ, 1OTB)	0.39	0.89	Cut to 1.2Å	PYP (100K, 295K)*
4	(4EUL, GFP_1.37A) (4EUL, GFP_{PHENIX})	0.66 0.67	0.00 0.4	Cut to 1.8Å Cut to 1.6Å	GFP (100K, 277K**)
5	DHFR	~0.9	0-0.5	Cut to 1.2Å	(see previous slide)
-	HEWL/NaI anom.	1.00	0.00	To 1.26Å	NECAT_HEWL_RT_NaI_82_XDS

*31.9 v 35.3% solvent (1NWZ/1OTB)
**RT data set looks rather crappy; second row using "Filtered" FPs from PHENIX refinement MTZ

Specifying priors

- Ultimately, we do not know *a priori* the correlation between a reference data set and a target data set which is to be scaled and merged.
- To parametrize priors, we need to know:
  - The normalized structure factor amplitudes of the reference
  - Initial (a, b) or r_{DW} calculated per s.f. using **`eff_r_dw_per_hkl`** (in **`fitting_dw.py`**)

4_Making_Priors
4A_Making_Priors_Anom

Specifying priors

- **5_Parsing_DW_parameters** summarizes how to formulate priors based on the provided r_{DW} and $|E|$.