JOURNAL OF APPLIED CRYSTALLOGRAPHY

Volume 52 (2019)

Supporting information for article:

BraggNet: Integrating Bragg Peaks using Neural Networks

Brendan Sullivan, Rick Archibald, Jahaun Azadmanesh, Venu Gopal Vandavasi, Patricia S. Langan, Leighton Coates, Vickie Lynch and Paul Langan

S1. Notes on Intensity Statistics for Ideal Crystals

Intensity statistics for the ratios of intensity moments $<I^2>/<I>^2$, $<F>^2/<F^2>$, and $<|E^2-1|>$ arise naturally from the idealized probability distributions of intensities *p(I)*, which have been known since 1949 (Wilson, 1949). The ideal probability distribution function (PDF) for acentric reflections is:

$$p_A(I)dI = \exp(\frac{-I}{< I >})d\left(\frac{I}{< I >}\right) = \gamma_1(\frac{I}{< I >})d\left(\frac{I}{< I >}\right)$$

While the PDF for centric distributions is:

$$p_C(I)dI = \sqrt{\frac{2 < I >}{\pi}}\exp(\frac{-I}{2 < I >})d\left(\frac{I}{2 < I >}\right) = \gamma_{1/2}(\frac{I}{2 < I >})d\left(\frac{I}{2 < I >}\right)$$

Consider the case for acentric peaks. It is common to consider resolution-normalized data. We define resolution-normalized intensities, Z, and resolution-normalized structure factors, E, as follows:

$$Z = \frac{I}{< I >} \qquad E = \sqrt{Z} = \sqrt{\frac{I}{< I >}}$$

Which allows the PDF to be expressed naturally in terms of Z:

$$p_A(Z)dZ = \exp(-Z)d(Z) = \gamma_1(Z)d(Z)$$

From which the cumulative distribution function (CDF), N(z), can be expressed:

$$N_A(z) = \int_0^z p_A(Z)dZ = \int_0^z e^{-Z}dZ = 1 - e^{-z}$$

And the ratio of moments is determined as usual:

$$< I^2 > = \int_0^\infty I^2 p_A(I)dI = 2 < I >^2$$

And so it follows:

$$\frac{< I^2 >}{< I >^2} = \frac{2 < I >^2}{< I >^2} = 2$$

Similarly,

$$< F > = < \sqrt{I} \geq \int_0^\infty \sqrt{I}\, p_A(I)dI = \frac{\sqrt{\pi}}{2} < I >$$

$$< F^2 > = < \left(\sqrt{I}\right)^2 > = < I >$$

So,

$$\frac{< F >^2}{< F^2 >} = \frac{\frac{\pi}{4} < I >}{< I >} = \frac{\pi}{4} \approx 0.785$$

Finally, the expectation value of <|E^2-1|>:

$$< |E^2 - 1| > = \int_0^\infty |E^2 - 1| p_A(I) dI = \int_0^\infty |Z - 1| e^{-Z} dZ = \frac{2}{e} \approx 0.736$$

Following a similar analysis for centric peaks, one finds the CDF for acentric peaks is:

$$N_C(z) = \int_0^z p_C(Z) dZ = \mathrm{erf}(\sqrt{\frac{Z}{2}})$$

Where erf is the error function. The ideal ratios of moments are given in Table 2.

The *L* test was proposed in 2003 (Padilla & Yeates, 2003) as a method to assess data quality using local intensity differences, particularly as a robust test for twinning. The authors define the unitless quantity *L* by comparing two peaks near each other in reciprocal space:

$$L = \frac{I_1 - I_2}{I_1 + I_2} \rightarrow I_2 = I_1 \frac{1 - L}{1 + L}$$

Following the authors' original derivation, the CDF is found by integrating:

$$N(L) = \int_0^\infty \int_{I_1 \frac{(1-L)}{(1+L)}}^\infty P(I_1, I_2) dI_2 dI_1$$

$$= \int_0^\infty \int_{I_1 \frac{(1-L)}{(1+L)}}^\infty \frac{1}{< I >^2} e^{-\frac{I_1 + I_2}{<I>}} dI_2 dI_1$$

$$= \frac{(L + 1)}{2}$$

Which can be differentiated to give the probability density function P(L):

$$P(L) = \frac{d(N(L))}{dL} = \frac{1}{2}$$

Which is again integrated to give the CDF of |L|, N(|L|):

$$N(|L|) = |L|$$

As is shown in Figure 4. The expectation values of |L| and |L^2| are straightforward to arrive at from here:

$$< |L| > = \int_{-1}^0 -L P(L) dL + \int_0^1 L P(L) dL = \frac{1}{2}$$

$$< |L^2| > = \int_{-1}^1 L^2 P(L) dL = \frac{1}{3}$$

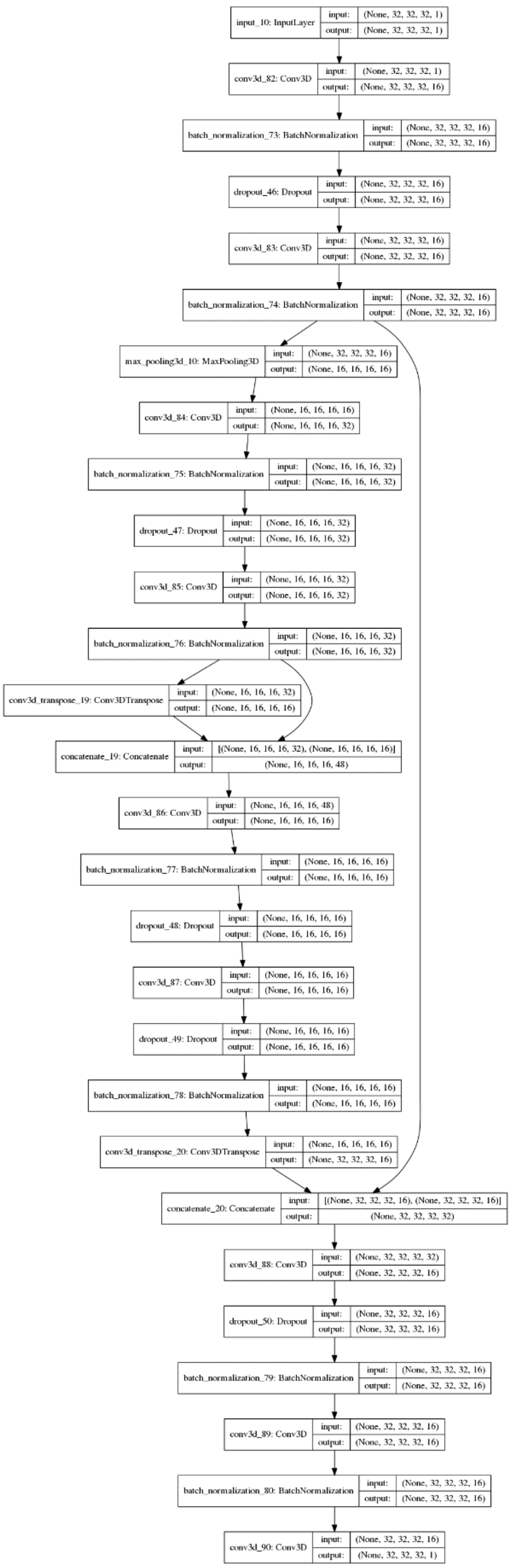

Figure S1 Full schematic of the neural network used for neural network integration.

Table S1 Merging statistics for peaks with I/σ > 1 for a given integration method to a resolution of 1.65 Å. While it is difficult to compare merging statistics from different peak sets, it is clear that neural networks have the possibility to extent completeness at high-resolution shells without compromising data quality.

Neutron Unit Cell Parameters	a = b = 73.3 Å, c = 99.0 Å, α = β = 90°, γ = 120°			
Space Group	$P3_221$			
Number of Orientations	5			
Resolution Range (Å)	13.97-1.65 (171-1.65)			
	Neural Network	**Profile Fitting**	***k*-NN**	**Spherical**
Number of Unique Reflections	36,253 (3,252)	35,503 (2,984)	35,184 (3,028)	36,446 (3,465)
Completeness	95.93% (87.61%)	93.95% (80.39%)	93.10% (81.57%)	96.44% (93.35%)
Multiplicity	3.75 (2.20)	3.57 (1.93)	3.51 (1.98)	3.47 (2.52)
Mean I/σ	9.8 (2.7)	10.9 (2.1)	7.9 (2.1)	8.0 (4.4)
R_{merge}	11.8% (36.5%)	12.4% (24.3%)	20.4% (41.2%)	17.2% (26.6%)
R_{pim}	6.4% (26.1%)	6.8% (18.4%)	11.3% (30.7%)	9.7% (18.7%)
$CC_{1/2}$	0.991 (0.353)	0.987 (0.389)	0.963 (0.073)	0.977 (-0.021)

Table S2 Summary statistics for peaks with I/σ > 1 for the given integration method to a resolution of 1.65 Å (left, shaded) and for peaks with I/σ > 1 for all three integration methods to a resolution of 1.8 Å. These data show that intensity statistics depend more strongly on the integration method than peak selection.

Model	$\langle I^2\rangle/\langle I\rangle^2$	$\langle F\rangle^2/\langle F^2\rangle$	$\langle L\rangle$	$\langle L^2\rangle$	$\langle I^2\rangle/\langle I\rangle^2$	$\langle F\rangle^2/\langle F^2\rangle$	$\langle L\rangle$	$\langle L^2\rangle$
Theory	**2.0**	**0.785**	**0.518**	**0.33**	**2.0**	**0.785**	**0.518**	**0.333**
NN	1.869	0.831	0.429	0.255	1.859	0.830	0.431	0.254
***k*-NN**	1.714	0.859	0.393	0.218	1.772	0.850	0.402	0.226
PF	1.710	0.863	0.378	0.207	1.773	0.850	0.400	0.225

Table S3 Crystallographic data and refinement statistics for X-ray data. Values for the outer resolution shell are given in parentheses.

Diffraction source	Rigaku FRE SuperBright Cu Kα rotating-anode generator
Wavelength (Å)	1.5418
Temperature (K)	296
Detector	R-Axis IV^{++}
Crystal-detector distance (mm)	135
Rotation range per image (°)	0.5
Exposure time per image (s)	60
Space group	$P3_221$
$a = b$ (Å)	73.40
c (Å)	99.43
$\alpha = \beta$ (°)	90
γ (°)	120
Mosaicity (°)	0.31
Resolution range (Å)	50.0–1.57 (1.60–1.57)
Total No. of reflections	459516
No. of unique reflections	43596
Completeness (%)	99.4 (91.3)
Multiplicity	10.5 (2.8)
$\langle I/\sigma(I)\rangle$	26.6 (2.2)
R_{meas}	0.08 (0.47)