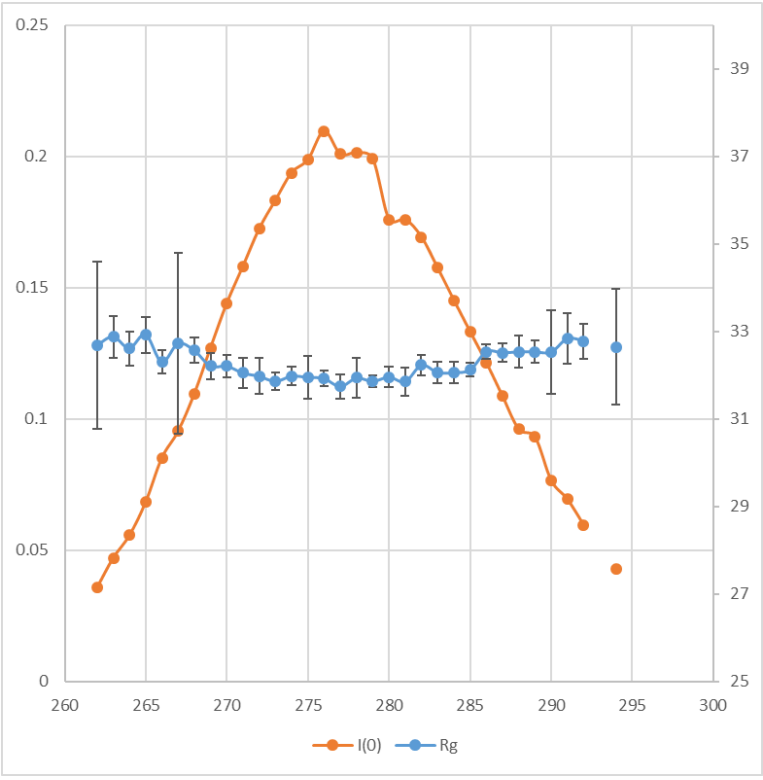
Data Reduction and preliminary analysis

The first step of the practical is to determine *I*(0) and *R*g for all of the provided files. The suggested way to do this was to use AutoRg from the command line (this did cause some issues during the actual practical), but the Chromix program from the ATSAS suite, or BioXTAS Raw are able to do this part more interactively. Using AutoRg, you will find that Guinier analysis is performed only on frames around the elution peak. This analysis is shown in the attached file GI.xlsx.

Q1: Qualitatively describe the relationship between Rg and I(0)? Why do I(0) and Rg change in this way? HINT: Is there evidence of an attractive or repulsive interaction between molecules in solution at higher concentrations?

A1: What the plot shows (reproduced below) is that when *I*(0) is low, *R*g is higher, and when *I*(0) is high, *R*g is lower. For a monodisperse protein, *I*(0) is a proxy for concentration, hence, when concentration is low, *R*g is higher and vice versa. This is typical behaviour for a repulsive interaction between particles in solution. At the peak concentration, the solution is not in the dilute regime, so data needs to be averaged over frames at lower concentrations where Rg plateaus at higher values, and the solution is deemed dilute.



Performing Guinier analysis on each of the 3 provided data sets yields interesting results. The *R*g values are: 32.7 ± 0.1 (leading edge, frames 262-268); 31.6 ± 0.1 (peak, frames 274-280); 32.6 ± 0.1 (trailing edge, frames 286-292). With respect to estimated uncertainty, the *R*g at high concentration is significantly lower than for the other samples. Importantly, the *q*.*R*g range (or *s*.*R*g range in the ATSAS software) used for the analysis is larger for the low concentration samples (*q*.*R*g < 1.3), and smaller for the high concentration sample (*q*.*R*g < 1.1 – if you increase the *q*.*R*g range, the difference plot shows clear systematic deviation of the Guinier fit and the experimental data.). Thus, the structure factor term has an observable impact on the Guinier region in the high concentration sample.

Q2: Are the Rg values significantly different for each of the data ranges (within estimated uncertainty)? Which range would you choose for further analysis (Frames 262-268, 274-280, 286-292)? Why?

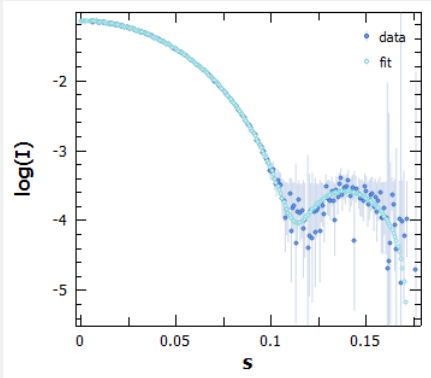
A2: The *R*g for the low concentration data (Frames 262-268 and Frames 286-292) are not significantly different, but they are both significantly different to the high concentration data. As there can often be aggregate protein eluting close to the leading edge of the main peak, you would usually use the data from the trailing edge for further analysis. But in this case, the elution peak is well separated from any other peaks and you could choose data from either the leading or trailing edge of the peak, but not from the peak itself.

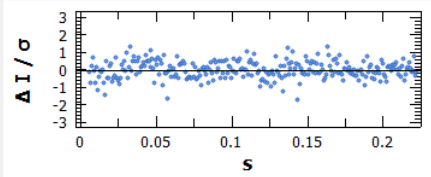
Distance distribution

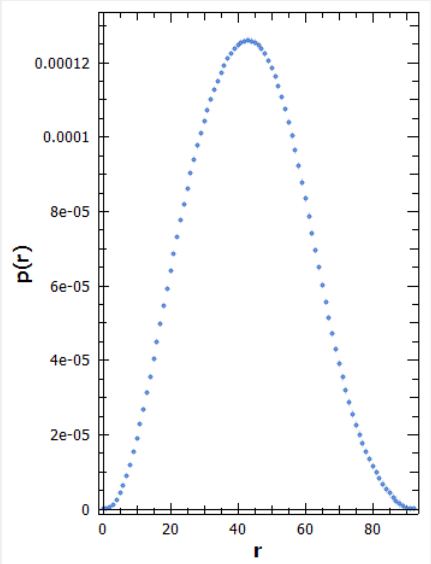
The distance distribution function is determined robustly for both the lower concentration datasets, where small changes in alpha (the smoothness parameter) do not affect the result. The *D*max value is automatically determined to be ~90 Å. Manually, *D*max could be reasonably set to as high as 95 Å, and the uncertainty is of the order of ± 5 Å. For the high concentration data set, the automatically determined *D*max is closer to 85 Å and the curve approaches the r-axis quite sharply. It is quite clear that something is not quite right. Increasing the Dmax value to 140 Å, it can be seen that the curve goes negative at large values of *r*, consistent with repulsive inter-particle interactions. If the Dmax value for the other two data sets is increased, then the curve stays very close to 0 past an r-value of ~90 Å. This behaviour is expected for data from a monodisperse solution.

Primus Distance Distribution Wizard

Distance Distribution Analysis
C:/Users/awh/Desktop/SAXS-Prac/2DsubGI_sum_0262_to_0268.dat







Total quality estimate		0.85
Guinier R_g/I(0)	32.42	0.07
p(r) R_g/I(0)	32.41	0.07
Porod Volume		229100.00

Range: 1 – 275

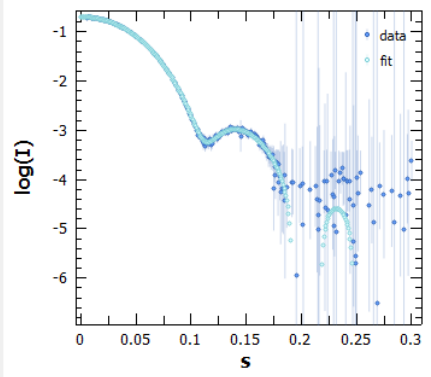
Setup: Point Collimation

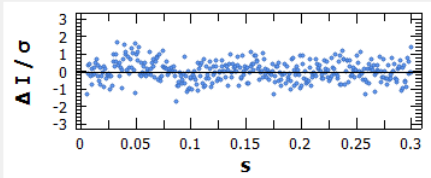
System: Arbitrary Monodisperse
Rmin: 0.00
Rmax: 92.00
☑ p(Rmin)=0 ☑ p(Rmax)=0
Points: 93
Alpha: 1.0000
Autognom

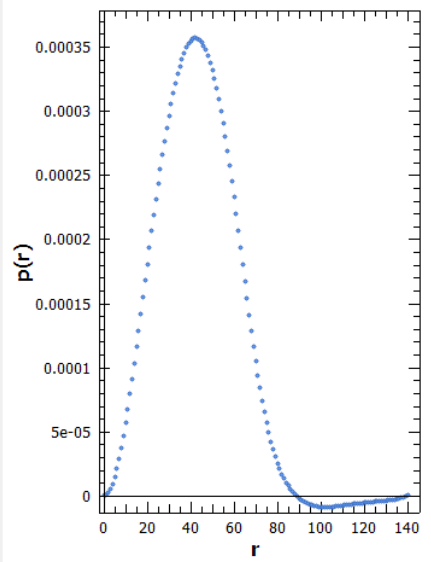
Save | < Back | Finish

Primus Distance Distribution Wizard

Distance Distribution Analysis
C:/Users/awh/Desktop/SAXS-Prac/2DsubGI_sum_0274_to_0280.dat







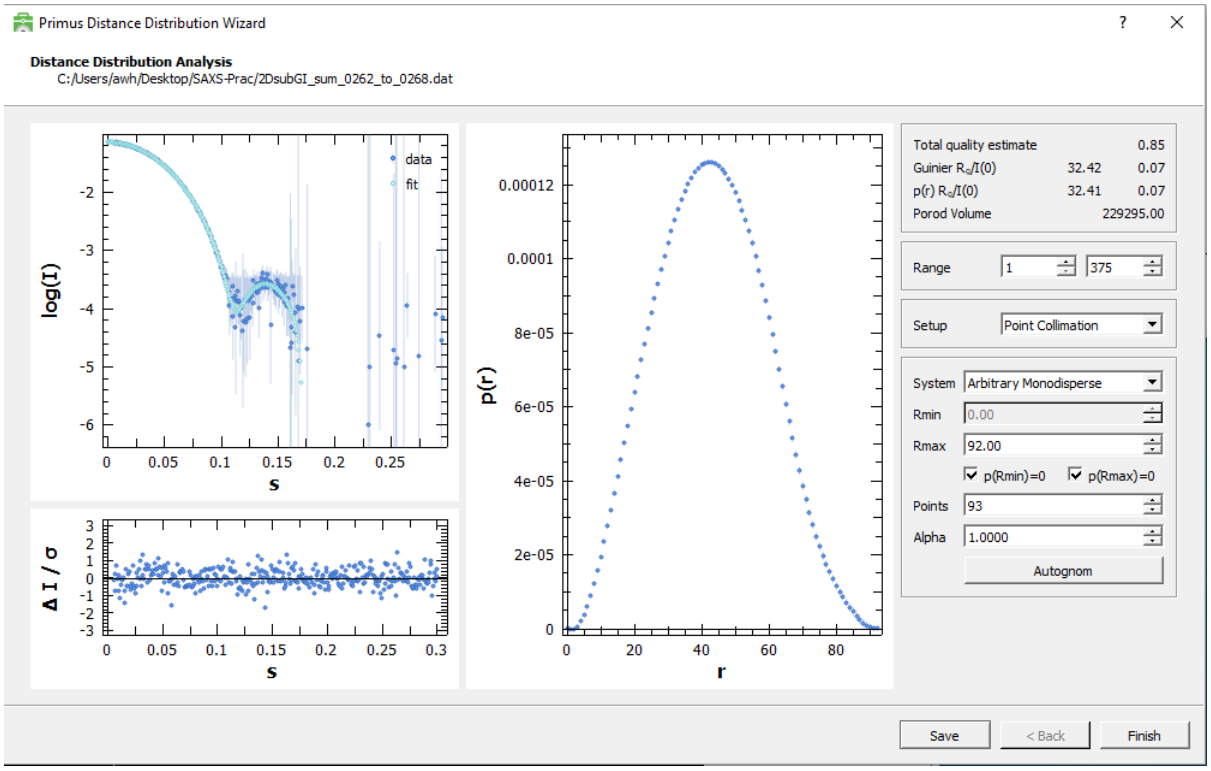
Total quality estimate		0.66
Guinier R_g/I(0)	30.56	0.20
p(r) R_g/I(0)	30.53	0.20
Porod Volume		235341.00

Range: 1 – 375

Setup: Point Collimation

System: Arbitrary Monodisperse
Rmin: 0.00
Rmax: 140.00
☑ p(Rmin)=0 ☑ p(Rmax)=0
Points: 141
Alpha: 14.5800
Autognom

Save | < Back | Finish



Q3: Letting PrimusQt determine the smoothness parameter, what Dmax value yields a p(r) function that approaches the r-axis asymptotically for each data range? Does the p(r) remain positive at all distances?

A3: The *D*max for the low concentration samples is 93 ± 5 Å (or thereabouts). The *D*max for the high concentration sample is ~140 Å, but the curve goes negative ~85 Å, which is indicative of repulsive inter-particle interference – consistent with the previous findings.

Q4: Looking at the Rg and Dmax values, would you change your opinion which dataset you would choose for further analysis?

A4: Data averaged over frames 262-268 or frames 286-292 could be used for further analysis. Both give almost identical *p*(*r*) functions, *R*g values, and *D*max values.

Estimating molecular mass

Calculation of molecular mass for this example is quite complicated. The contrast, partial specific volume, and extinction coefficient need to be estimated from the sequence. Corrections need to be made for sample path length (the sample thickness in co-flow mode is less than the diameter of the

capillary). Corrections to the A280 values are also required as the absorbance cell does not have a 1 cm path length, and the data aren't automatically corrected for thickness. Lastly, the average sample concentration over the measurement time needs to be determined.

Entering the sequence into MULCh, we get the following:

Tabulated scattering length densities and contrasts

	ρ (10^{10}cm^{-2})			$\Delta\rho$ (10^{10}cm^{-2})		
	1	**2**	**Solvent**	**1**	**2**	**Total**
X-RAY	12.358	12.358	9.403	2.955	2.955	2.955
NEUTRON						
0.0	1.968	1.968	-0.560	2.528	2.528	2.528
0.1	2.095	2.095	0.135	1.960	1.960	1.960
0.2	2.222	2.222	0.829	1.393	1.393	1.393
0.3	2.349	2.349	1.524	0.825	0.825	0.825
0.4	2.476	2.476	2.219	0.257	0.257	0.257
0.5	2.603	2.603	2.914	-0.311	-0.311	-0.311
0.6	2.730	2.730	3.609	-0.879	-0.879	-0.879
0.7	2.857	2.857	4.304	-1.446	-1.446	-1.446
0.8	2.985	2.985	4.999	-2.014	-2.014	-2.014
0.9	3.112	3.112	5.694	-2.582	-2.582	-2.582
1.0	3.239	3.239	6.388	-3.150	-3.150	-3.150
Calculated match-point (f_{D_2O})				0.445	0.445	0.445

Other related quantities

The density (in g.cm^{-3} @ 20°C) of the solvent at any D_2O fraction is:

- Density = $0.998 + 0.107 f_{D_2O}$

The molecular mass (kDa) at any D_2O fraction is:

- $\text{Mass}_1 = 43.078 + 2.257 f_{D,1} + 0.643 f_{\text{AccessH},1} f_{D_2O} = 43.078 + 0.611 f_{D_2O}$
- $\text{Mass}_2 = 43.078 + 2.257 f_{D,2} + 0.643 f_{\text{AccessH},2} f_{D_2O} = 43.078 + 0.611 f_{D_2O}$

The partial specific volume (cm^{3}g^{-1}, fully proteated) is:

- $v_1 = 0.732$
- $v_2 = 0.732$
- $v = 0.732$

The X-ray scattering length per unit mass (10^{10}cm.g^{-1}) is:

- $\Delta\rho_{M,1} = 9.050$
- $\Delta\rho_{M,2} = 9.050$
- $\Delta\rho_{M} = 9.050$

Q5: From the calculated sequence: What is the partial specific volume of Glucose Isomerase? What is the contrast of Glucose Isomerase (don't forget the 10^{10})? What is the mass of glucose isomerase protein based on the sequence?

A5: Excluding the first methionine (which are usually removed by the MAP enzyme in the cell), the partial specific volume of Glucose Isomerase (residues 2-388) is 0.732 cm^3/g, and the contrast is 2.955×10^{10} cm^{-2}. The mass derived from the primary sequence of Glucose Isomerase is 43 kDa.

Q6: Why would we want to use the Abs 0.1% instead of the molar extinction coefficient?

A6: Because the equation to determine particle mass requires the concentration be in mass based units ($g.cm^{-3}$), it is easier to use the Abs 0.1% value. It is possible to determine the molar concentration and then convert it to a w/v concentration, but using the Abs 0.1% value makes the task slightly easier.

The next step involves the absorbance data, and determining the time offset such that the absorbance values approximate the absorbance during measurement. There are automated and quantitative ways to align the A280 curve to the *I*(0) curve, but this can be done manually quite easily. Aligning the 5 measurements around the peak, the time offset is approximately 1.5 s. The absorbance values are then converted to concentrations by correcting for the cell path length (0.247 cm) and the Abs 0.1% value (1.077). Another issue is that the absorbance measurement intervals don't align exactly with the measured data. There are ways to deal with this, but here, a trapezoidal integration over the best matching time interval, divided by that time interval should give a reasonable approximation to the average concentration. Calculations are shown in GI.xlsx

Q7: What is the average concentration (in g/cm^3) of protein for each of the datasets (Frames 262-268, Frames 274-280, and Frames 286-292)?

A7: The average concentration of protein between frames: 262-268 is 1.1 mg/mL; 274-280 is 3.2 mg/mL; 286-292 is 1.4 mg/mL.

The last step is to correct the *I*(0) for path length due to the co-flow setup (multiply by 2.05) and use equation (2) to determine the mass.

Q8: Using equation (2), what is the apparent molecular mass of the Glucose Isomerase in solution for each dataset? Does this value differ from the value calculated from the primary sequence? If so why? HINT: look at the ratio between the mass calculated using equation (X) and that calculated from the primary sequence.

A8: The particle mass determined from equation (2) is: 170 kDa for frames 262-268; 160 kDa for frames 274-280; and 170 kDa for frames 286-292. This value is much higher than the value estimated from the sequence (43 kDa). Taking the ratio, we get a value close to 4, indicating that the particle is a homotetramer (composed of 4 individual glucose isomerase molecules).

Q9: What are possible sources of error when determining the molecular mass of the protein?

A9: The precision of *I*(0) is relatively high, thus, providing the solution is monodisperse, the determination of *I*(0) is not a significant source of error. The main sources of error come from the other parameters. Contrast, specific volume, and extinction coefficients are all estimated from the primary sequence, and can differ significantly from actual values. From a practical point of view, baseline correction accuracy of the A280 data can very important for a protein with a low extinction coefficient. With all these potential sources of error, mass estimates can deviate by 10% -20% from the actual value.

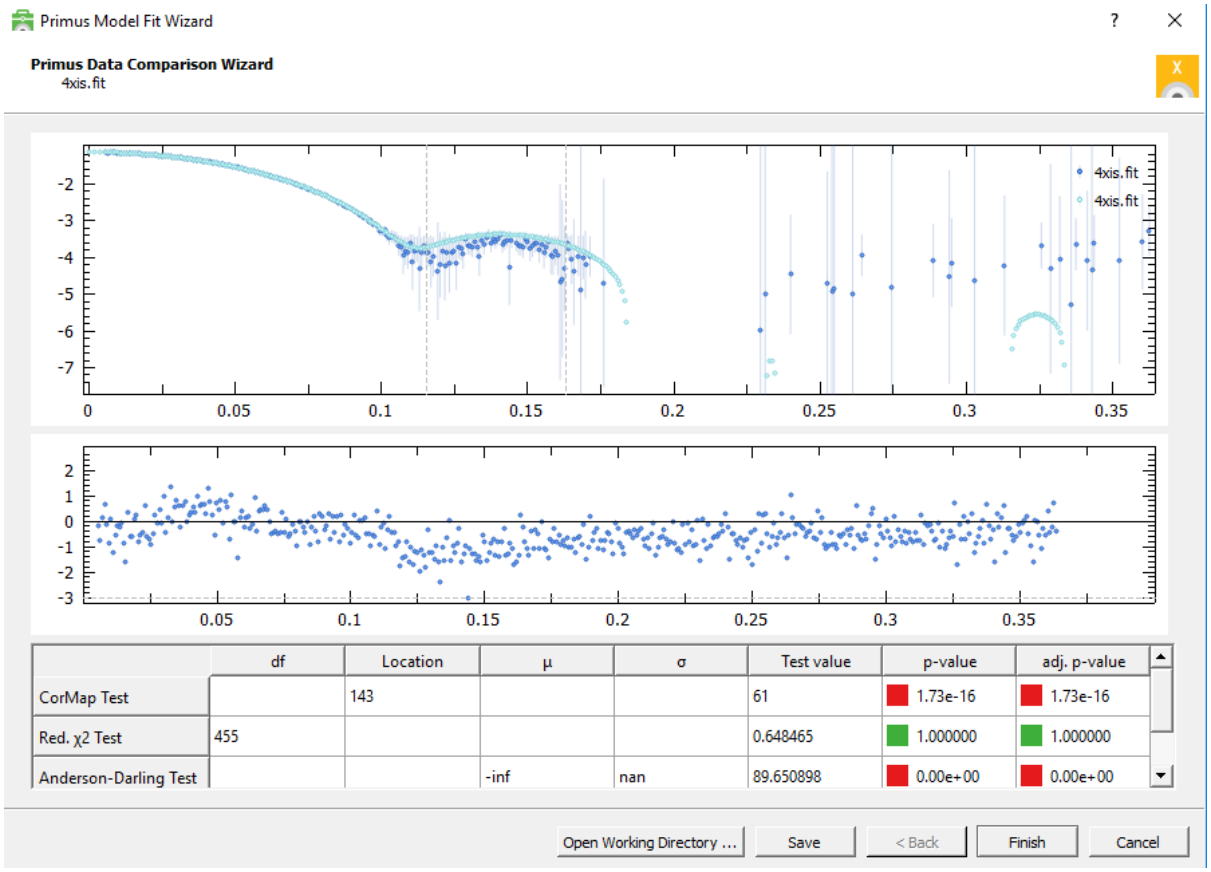
Prediction of scattering profiles from crystal structures

Going to the PDB entry for 4XIS.

Q10: Each protein that forms the molecule is coloured differently. How many individual proteins make up the functional glucose isomerase molecule? Is this consistent with molecular mass calculated from I(0)?

A10: From the crystal structure 4XIS, we can see that there are 4 individual protein subunits that make up the Glucose Isomerase particle. This is consistent with the mass estimates determined earlier.

Using Crysol within Primus Qt for the leading edge data.



Q11: What is the χ^2 statistic for comparison of the scattering curve calculated from 4XIS and the experimental data? Try with and without a background correction – does this improve the fit? Which dataset(s) agree best with the crystal structure? Is this the dataset you chose earlier for further analysis?

A11: The χ^2 statistic for the data from the leading edge is 0.65 and for the data from the trailing edge is 1.38 (with background corrections as it improves the fit in both cases). Thus, the data that provides the best fit to the data is that averaged over frames 262-268, which was stated earlier as being one of the datasets that could be used for further analysis. From the plot above it can be seen that the model scattering curve does deviate from the scattering data, but the features of the curve are similar below 0.15 $Å^{-1}$. A final note is that χ^2 should be close to 1 for a good model, however, here the χ^2 is much less than 1, indicating that the estimated uncertainty of the intensity measurements are over estimated. Thus, χ^2 cannot be used here can be used to discriminate between two different data sets (only the fit of different models to the same data set).

Q12: What does the background correction tell you about the reduced data? Is the background scattering high, low, correct? Were the frames chosen for solvent scattering representative of the solvent present as the protein eluted? If not, what may have been the issue?

A12: Past 0.15 $Å^{-1}$ it becomes clear that the data is over-subtracted, and this is the reason why a background correction improves the fit to the data. The reason for the over-subtraction is likely due to the SEC column not being properly equilibrated, such that the solvent scattering varied subtly over the course of the elution, making it difficult to average frames that are representative of the solvent scattering.

Ab initio modelling

Q13: What point group does glucose isomerase possess? Check your answer against the "Global Symmetry" on the PDB page: http://www.rcsb.org/structure/4XIS

A13: From the structure 4XIS, three perpendicular two fold axes can be identified. This means it belongs to the point group D_2 or 222 (depending on the notation). This is consistent with the information given in the PDB entry for this protein which states the point group as "Dihedral D_2".

Dummy atom modelling (1 single run) was performed on data averaged over frames 286-292. The dummy atom model is shown below, with the crystal structure overlayed on the next page.

Q14: What is the χ^2 value for the fit? Take a look at the averaged dummy atom model. Is the shape and dimensions consistent with the crystal structure?

A14: The χ^2 value for the fit to the trailing edge data (taken from the .fir output file) is 1.049. When the crystal structure is overlayed on the dummy atom model, the size and shape appears to be consistent (see image on next page for a view in one orientation).