JOURNAL OF APPLIED CRYSTALLOGRAPHY

ISSN 1600-5767

Describing small-angle scattering profiles by a limited set of intensities

Thomas D. Grant*

Department of Structural Biology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, NY 14203, USA. *Correspondence e-mail: tdgrant@buffalo.edu

Received 29 May 2021
Accepted 24 June 2022

Edited by D. I. Svergun, European Molecular Biology Laboratory, Hamburg, Germany

Keywords: small-angle scattering; indirect Fourier transform; solution scattering; pair distribution function.

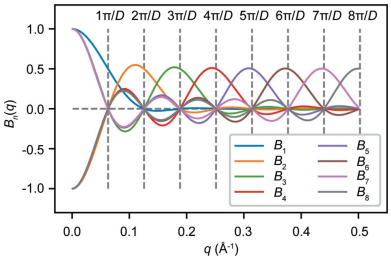
Supporting information: this article has supporting information at journals.iucr.org/j

Small-angle scattering (SAS) probes the size and shape of particles at low resolution through the analysis of the scattering of X-rays or neutrons passing through a solution of particles. One approach to extracting structural information from SAS data is the indirect Fourier transform (IFT). The IFT approach parameterizes the real-space pair distribution function [$P(r)$] of a particle using a set of basis functions, which simultaneously determines the scattering profile [$I(q)$] using corresponding reciprocal-space basis functions. This article presents an extension of an IFT algorithm proposed by Moore [*J. Appl. Cryst.* (1980), **13**, 168–175] which used a trigonometric series to describe the basis functions, where the real-space and reciprocal-space basis functions are Fourier mates. An equation is presented relating the Moore coefficients to the intensities of the SAS profile at specific positions, as well as a series of new equations that describe the size and shape parameters of a particle from this distinct set of intensity values. An analytical real-space regularizer is derived to smooth the $P(r)$ curve and ameliorate systematic deviations caused by series termination. Regularization is commonly used in IFT methods though not described in Moore's original approach, which is particularly susceptible to such effects. The algorithm is provided as a script, `denss.fit_data.py`, as part of the *DENSS* software package for SAS, which includes both command line and interactive graphical interfaces. Results of the program using experimental data show that it is as accurate as, and often more accurate than, existing tools.

1. Introduction and overview

Small-angle scattering (SAS) yields structural information at low resolution about the size and shape of particles in solution. X-rays or neutrons scattering from freely tumbling particles in solution exhibit rotational averaging in reciprocal space, resulting in isotropic scattering profiles collected on 2D detectors. This rotational averaging results in the loss of information describing the 3D structure of the particle. The scattering of a molecule $I(q)$, where q is the momentum transfer [$q = (4\pi/\lambda)\sin\theta$, where θ is half the scattering angle and λ is the wavelength of the incident radiation], is determined by its 3D scattering length density function, and thus SAS profiles can be calculated directly from known atomic structures. However, due to the spherical averaging of the intensities, the inverse problem of calculating a unique 3D structure from SAS profiles is not possible. Nonetheless, structural information describing global properties of size and shape can be obtained through analysis of the SAS profile.

While unique 3D real-space information cannot be obtained directly from a SAS profile, a Fourier transform of the reciprocal-space intensity profile yields the set of pair distances in the particle, known as the pair distribution function or $P(r)$. However, due to limitations caused by the



termination of higher-order scattering data to a finite q range, uncertainties in intensity measurements and systematic errors, direct calculation of the Fourier transform yields $P(r)$ functions with large systematic deviations (Glatter, 1977; Moore, 1980; Hansen & Pedersen, 1991; Svergun, 1992; Svergun & Pedersen, 1994). One popular approach to extracting this structural information from SAS profiles is the indirect Fourier transform (IFT) proposed by Glatter (1977). In this approach, a set of basis functions is used to parameterize the $P(r)$ function. The weights of these basis functions are then adjusted to optimize the fit of the corresponding intensity function to the experimental scattering profile.

One such IFT algorithm proposed by Moore (1980) takes advantage of information theory (Shannon, 1948) to describe a set of basis functions defined by the maximum particle dimension D. Moore uses a trigonometric series to define a function $Q(r) = P(r)/r$. This definition resulted in a convenient relationship between the real-space $Q(r)$ and the reciprocal-space $U(q) = qI(q)$, where the two are Fourier mates. Key to Moore's approach (and other IFT methods; Glatter, 1977; Svergun, 1992) is that the coefficients of the series terms define both the real-space and reciprocal-space profiles, using the appropriate basis functions. Least squares can be used to determine the coefficients and the associated standard errors by minimizing the fit to the experimental scattering profile (full details are given in Section S1 of the supporting information). This approach has the advantage of providing the necessary information on the variances and covariances of the coefficients to determine the errors on each coefficient. Moore showed, using Shannon information theory, that the number of coefficients that can be determined from the data is the number of independent pieces of information that the data are able to describe about the particle. Moore derived a series of equations relating the coefficients to commonly used SAS parameters such as the forward scattering intensity $I(0)$, the radius of gyration R_g and the average vector length $\overline{r}$, along with error estimation for each parameter. One advantage of Moore's approach over others is that a separate regularizing function is not explicitly required to smooth the $P(r)$ curve due to the use of the sine series (Moore, 1980). However, in practice with experimental data, it has been found that Moore's approach is often more susceptible to large oscillations in the $P(r)$ curve due to series termination (Svergun & Pedersen, 1994; Hansen & Pedersen, 1991), probably because of the lack of a regularizing function. Such regularizing functions have been shown to be effective at smoothing the $P(r)$ curves calculated using Moore's approach (Tully *et al.*, 2021; Rambo, 2021).

Here we extend Moore's derivation to relate the Moore coefficients to specific intensity values such that each term in the series is now weighted by a corresponding intensity, termed I_n (Section S1 in the supporting information). We present equations for calculating a variety of commonly used SAS parameters and their associated errors from the I_n values. Additionally, we derive a modified equation for least-squares minimization taking into account an analytical regularization of the $P(r)$ curve. We provide open-source software with convenient interfaces for performing all of the presented calculations, including a novel approach to estimating parameters sensitive to systematic errors. Finally, we describe the results using both simulated and real experimental data and compare with current state-of-the-art software tools.

2. Theoretical background

2.1. Extension of Moore's IFT

Moore's use of Shannon information theory to define $I(q)$ resulted in a selection of q values, namely $q_n = n\pi/D$, termed 'Shannon channels' (Feigin & Svergun, 1987; Svergun & Koch, 2003; Rambo & Tainer, 2013). The intensities at q_n, *i.e.* $I_n = I(q_n)$, therefore become important values as they determine the Moore coefficients a_n and thus similarly can be used to describe completely the low-resolution size and shape of a particle obtainable by SAS. In Section S1 we derive the mathematical relationship between I_n and a_n which results in the following general equation for $I(q)$ as a function of the intensity values at the Shannon points:

$$I(q) = 2\sum_{n=1}^{\infty} I_n \frac{(n\pi)^2}{(n\pi)^2 - (qD)^2} \frac{\sin(qD)}{qD} (-1)^{n+1}. \quad (1)$$

Defining basis functions B_n as

$$B_n(q) = \frac{(n\pi)^2}{(n\pi)^2 - (qD)^2} \frac{\sin(qD)}{qD} (-1)^{n+1}, \quad (2)$$

$I(q)$ can now be expressed as a sum of the basis functions B_n weighted by the intensity values at q_n,

$$I(q) = 2\sum_{n=1}^{\infty} I_n B_n(q). \quad (3)$$

As in Moore's original approach, the B_n functions are determined by the maximum dimension of the particle D. B_n values for $D = 50$ Å are illustrated in Fig. 1. The $P(r)$ function can be represented using the series of I_n values as

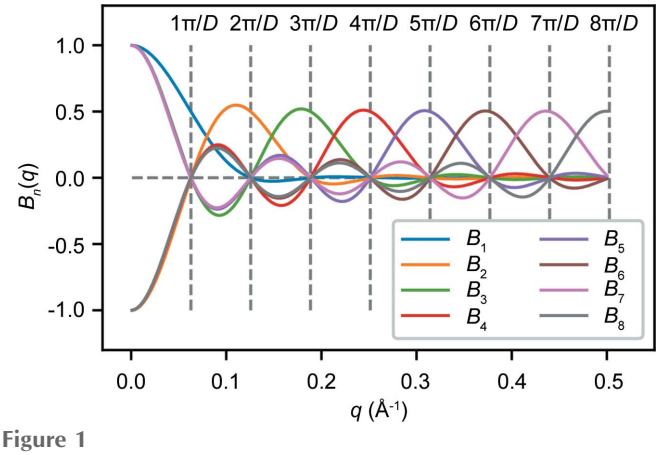


Figure 1
A plot of reciprocal-space basis functions B_n for any particle of size D = 50 Å. Vertical dashed lines show the locations of the Shannon points q_n.

$$P(r) = \frac{r}{2D^2}\sum_{n=1}^{\infty} I_n n \sin\left(\frac{n\pi r}{D}\right) \quad (4)$$

(Section S1) or by defining real-space basis functions S_n as follows:

$$P(r) = \sum_{n=1}^{\infty} I_n S_n(r), \quad (5)$$

$$S_n(r) = \frac{nr}{2D^2}\sin\left(\frac{n\pi r}{D}\right). \quad (6)$$

Least squares can be used to determine optimal values for each I_n from the oversampled experimental SAS profile, along with error estimates for each, taking into account the variances and covariances of the coefficients. These terms can then be used to calculate the corresponding $I(q)$ and $P(r)$ curves using equations (1) and (4) and the associated errors (Section S1).

The maximum particle dimension D is required for determining the q_n values associated with the I_n values. Estimates for the true value of D that are too small will result in B_n values that lack sufficiently high frequencies for the adequate reconstruction of $I(q)$. Estimates of D that are too large will result in overfitting the data. Moore found that testing increasing values of D yielded improved fits to the experimental $I(q)$ function and used χ^2 (Section S1) to estimate the true value of D by selecting the smallest D value that minimizes χ^2 while avoiding larger D values that result in overfitting (Moore, 1980). An alternative method is to estimate D from the $P(r)$ curve by first guessing a reasonable value for D, such as $3.5R_g$ or larger, fit $I(q)$ and calculate the $P(r)$ curve, and then estimate the true value of D on the basis of where $P(r)$ gradually falls to zero.

2.2. Derivation of parameters from I_n values

Similarly to what Moore described for the a_n coefficients, since the I_n values contain all the information present in $I(q)$, quantities that can be derived from $I(q)$ can also be derived directly from the I_n values. For example, to determine the forward scattering intensity $I(0)$, we take the limit of equation (1) as q approaches zero to yield

$$I(0) = 2\sum_{n=1}^{\infty} I_n(-1)^{n+1}. \quad (7)$$

Equation (7) demonstrates a simple relationship between the forward scattering of a particle and the I_n values. Note that the particle dimension D is not explicitly present in equation (7). Fig. 2 illustrates the relationship between the I_n values and $I(0)$.

The forward scattering of a particle is not directly measured in an experiment due to its coincidence with the incident beam and is thus typically estimated as an extrapolated value from low-q data points or by integration of the $P(r)$ function. Equation (7) provides an alternative method of measuring the forward scattering of a particle directly from the data through the sum of the I_n values. While equation (7) is defined as a sum from $n = 1$ to infinity, typical experimental setups only provide data for the first 10–30 Shannon channels, depending on the size of the particle. Thus in practice equation (7) yields an estimate of the forward scattering rather than an exact measurement. However, since the vast majority of the scattering intensity present in the profile occurs within these 10–30 Shannon channels, equation (7) should provide an accurate estimate of the forward scattering for most particles and experimental setups.

Other parameters can be similarly derived (Section S2). For example, R_g can be estimated from the I_n values as

$$R_g^2 = \frac{D^2}{I(0)}\sum_{n=1}^{\infty} I_n F_n, \quad (8)$$

where

$$F_n = \left[1 - \frac{6}{(n\pi)^2}\right](-1)^{n+1}. \quad (9)$$

Another parameter describing particle size is the average vector length in the particle $\bar{r}$, which can be estimated from the I_n values as

$$\bar{r} = \frac{4D}{I(0)}\sum_{n=1}^{\infty} I_n E_n, \quad (10)$$

where

$$E_n = \left[\frac{(-1)^n - 1}{(n\pi)^2} - \frac{(-1)^n}{2}\right]. \quad (11)$$

The Porod invariant Q is defined as the integrated area under the Kratky plot (Porod, 1982), which can be described in terms of the I_n values as

$$Q = \left(\frac{\pi}{D}\right)^3\sum_{n=1}^{\infty} I_n n^2. \quad (12)$$

The Porod volume can then be calculated using the Porod invariant (Section S2) (Porod, 1982). The Porod volume is commonly used to estimate molecular weight for globular

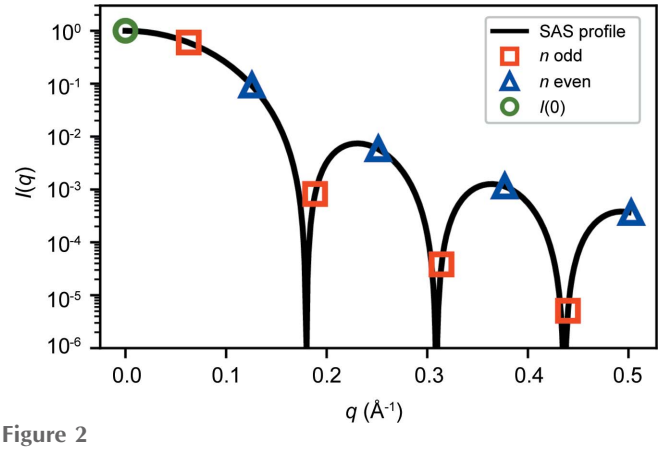


Figure 2
A plot of an example scattering profile, showing the relationship of I_n values and $I(0)$. Odd I_n values are shown as red squares, while even I_n values [which have a multiplication factor of −1 in equation (7)] are shown as blue triangles. Equation (7) states that twice the total sum of the red squares and (negative) blue triangles is equal to the forward scattering $I(0)$, shown as a green circle.

biological macromolecules. More recently, Rambo & Tainer (2013) derived a new SAS invariant termed the volume of correlation, V_c, with units of length^2 and which is related to the correlation length of the particle ℓ_c. V_c can be used to estimate the molecular weight for macromolecules that may be either globular or flexible (Rambo & Tainer, 2013). V_c can be estimated from the I_n values as

$$V_c = \frac{D^2 I(0)}{2\pi}\left[\sum_{n=1}^{\infty} I_n n\,\mathrm{Si}(n\pi)\right]^{-1}, \tag{13}$$

where $\mathrm{Si}(n\pi)$ is the Sine integral. The correlation length can similarly be calculated as

$$\ell_c = \frac{2D}{\pi}\frac{\sum_{n=1}^{\infty} I_n n\,\mathrm{Si}(n\pi)}{\sum_{n=1}^{\infty} I_n n^2}. \tag{14}$$

Since the variances and covariances of the I_n values are known from the least-squares minimization, error propagation can be used to determine the associated uncertainties for each of the parameters described above (Section S2).

2.3. Regularization of $P(r)$

The original IFT proposed by Glatter (1977) and other IFTs (Svergun, 1992; Vestergaard & Hansen, 2006) make use of regularization of the $P(r)$ curve, similar to the general method of Tikhonov regularization (Tikhonov & Arsenin, 1977). The goal is to use the knowledge that $P(r)$ functions are smooth for most particle shapes to generate curves that are free of strong oscillations from series termination and are relatively stable to statistical errors. Rather than minimize χ^2 directly, a new function T is minimized, taking into account the smoothness of the $P(r)$ curve according to equation (15):

$$T = \chi^2 + \alpha S, \tag{15}$$

where S is the regularizing function, which can take different forms, and α is a Lagrange multiplier that acts as a weight to determine the strength of the smoothing. Larger α leads to a smoother $P(r)$ function but may result in a worse fit of $I(q)$ to the experimental data. The IFT method used by Moore has been shown to be more susceptible than other IFT methods to oscillations in the $P(r)$ curve (Hansen & Pedersen, 1991; Svergun & Pedersen, 1994), most likely due to the lack of a regularizing function. We provide a detailed derivation of an analytical regularization of $P(r)$ using I_n values in Section S3.

As for other similar IFT methods utilizing regularization, a suitable choice of α must be found to optimize the smoothness of the $P(r)$ curve and the fit to the experimental data. Various methods for selecting the optimal value for α have been proposed, including via point of inflection (Glatter, 1977), Bayesian methods (Vestergaard & Hansen, 2006) and using perceptual criteria (Svergun, 1992). We describe our approach in Section 2.4 below.

Equation (3) assumes a sum from $n = 1$ to infinity. However, data are only collected to the maximum q value allowed by the experiment, q_{max}. The lack of data for $q > q_{max}$ implicitly corresponds to setting the I_n values to zero for those data points where $n > n_{max}$ [where $n_{max} = \mathrm{int}(q_{max}D/\pi)$, *i.e.* the largest index in the series]. The regularization often results in poorer fits of the intensity profile at higher experimental q values with increasing α due to this implicit bias of I_n values for $n > n_{max}$ towards zero. In order to remove this bias and allow for the I_n values at $n > n_{max}$ to be unrestrained, I_n values for $n > n_{max}$ are allowed to float (calculated up to $3n_{max}$). Note that the number of Shannon channels that can be reliably extracted from the data is dictated largely by the quality of the data in addition to the q range, as described by Konarev & Svergun (2015).

2.4. Implementation

Tools for performing the least-squares fitting of I_n values to experimental data, calculation of parameters and errors, and regularization of $P(r)$ have been developed using Python, *NumPy* and *SciPy* (Harris *et al.*, 2020; Virtanen *et al.*, 2020) and are provided open source through the *DENSS* suite of SAS tools (Grant, 2018; https://github.com/tdgrant1/denss). The primary interface to use this algorithm is the `denss.fit_data.py` Python script. To enable ease of use, in addition to the command line interface, an interactive graphical user interface (GUI) (Fig. 3) has been developed using the *Matplotlib* package (Hunter, 2007).

2.4.1. Automatic estimation of D. To assist users, upon initialization of the script the experimental data are loaded and estimates of D and α are automatically calculated. To estimate D automatically, an initial estimate of D is calculated that is likely to be significantly larger than the actual D. This subsequently enables a more accurate estimation of D where $P(r)$ falls to zero. An initial value of $D = 7R_g$ is used as this should ensure a large enough value given a variety of particle shapes (Petoukhov *et al.*, 2007; Grant *et al.*, 2015). An initial rough estimate of R_g is first calculated using the Guinier equation (Guinier *et al.*, 1955) with the first 20 data points. In cases where that estimate fails (*e.g.* due to excessive noise or a positive slope of the Guinier plot), the Guinier peak method is instead used (Putnam, 2016). The I_n values are then calculated from the experimental data using the regularized least-squares approach outlined in Section S3, setting $\alpha = 0$ to optimize the fit to the data. After the initial I_n values have been calculated, the corresponding $P(r)$ function often suffers from severe ripples caused by Fourier termination effects due to the finite range of data, as described above, making it difficult to estimate D where $P(r)$ falls to zero. To alleviate this effect, a Hann filter, which is a type of Fourier filter (Blackman & Tukey, 1958), is applied to remove the Fourier truncation ripples from $P(r)$. D is then calculated from this filtered $P(r)$ curve as the first position r where $P(r)$ falls below $0.01P_{max}$ after the maximum, where P_{max} is the maximum value of the filtered $P(r)$. This new D value is then used to recalculate the I_n values for the best fit to the experimental scattering profile. In addition to automatically estimating D directly from the data, users can manually enter an initial estimate of D to begin with.

2.4.2. Automatic estimation of α. Next, the optimal α is estimated, which yields I_n values corresponding to a smooth $P(r)$ function while still resulting in a calculated $I(q)$ curve that

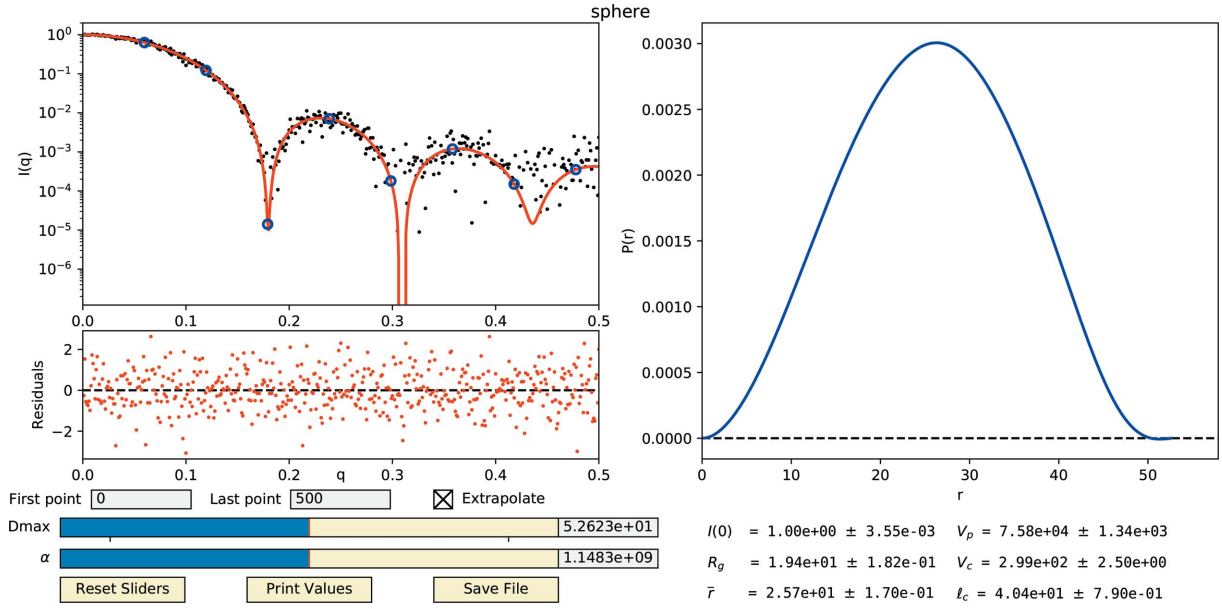


Figure 3
The interactive GUI display from the `denss.fit_data.py` script. The upper left panel shows the experimental $I(q)$ curve as black circles, fitted I_n values as blue circles and the fitted $I_c(q)$ calculated from the I_n values as a red curve, all on a semilog plot. The residuals of the experimental and calculated intensity curves are shown below the intensity plot. The panel on the right shows the $P(r)$ curve calculated from the I_n values. Input text boxes are provided at the bottom left to allow for trimming data points at the beginning or end of the curve, along with a checkbox to disable the calculation of intensities at high q values. Interactive sliders for D_{max} and α are also provided, along with corresponding input boxes for manual entry. The bottom right of the window shows size parameters calculated from the I_n values and associated uncertainties. Buttons for resetting the sliders, printing the size parameters and saving the results can be found at the bottom left.

fits the experimental data. First, the best χ^2 value possible is calculated by setting $\alpha = 0$ and using the D value estimated in the previous step. Then, various values of α are scanned, from 10^{-20} to 10^{20} in logarithmic steps of 10^1. This wide range is used to accommodate a variety of different scattering profiles covering a range of signal-to-noise values. At each step the χ^2 is calculated. The optimal α is chosen by interpolating where $\chi^2 = 1.1\chi^2_{best}$, *i.e.* where χ^2 rises to 10% above the best possible value.

2.4.3. Interface. The GUI mode of the script displays a plot of the intensities on a semilog y axis and plots the experimental data $I_e(q)$ and the initial fit $I_c(q)$, calculated from the I_n values at the experimental q (Fig. 3). The script additionally calculates $I_c(q)$ at q values extrapolated to $q = 0$. Users can alternatively provide a set of desired q values to calculate $I_c(q)$ as an ASCII text file when starting the program. The residuals, $[I_e(q_i) - I_c(q_i)]/\sigma_i$, are also displayed to assist in assessing the quality of the fit. Next to the plot of intensities, the $P(r)$ curve calculated from the I_n values is also displayed. In addition to input text boxes for manually entering new D and α values in the GUI, interactive sliders are available to change the D and α values, which automatically update the plots as they are adjusted. Users can also change the beginning and ending data points if desired, to remove outlier data points that often occur at either end of the experimental profile, or disable the calculation of intensities for $q > q_{max}$. Several of the parameters described above, including $I(0)$, R_g, $\bar{r}$, V_p, V_c and ℓ_c, along with associated uncertainties, are calculated from the I_n values and displayed in the GUI. These parameters are updated interactively whenever D or α are changed.

2.4.4. Calculation of V_p, V_c and ℓ_c. Particular care must be taken when estimating parameters that are sensitive to systematic errors in high-q data points, such as V_p, V_c and ℓ_c. In practice, direct estimation of these parameters using the equations described above may yield unstable results, even with regularization. Porod's law is based on the assumption that all scattering comes from the surface of a particle, resulting in an asymptotic intensity decay proportional to q^{-4} (Porod, 1982), giving rise to the ability to estimate values such as the Porod volume V_p. In practice, shape scattering contributes significantly (Rambo & Tainer, 2011), as do systematic errors caused by inaccurate background subtraction (Manalastas-Cantos *et al.*, 2021), resulting in poor estimation of these parameters without correction. To deal with this, many algorithms impose an artificial constant subtraction to force the Porod decay, which has proven effective at providing accurate estimates of particle volume (Manalastas-Cantos *et al.*, 2021). However, different algorithms have different methods for calculating the constant to subtract and for determining the fitting region where these calculations are performed, and there is often subjectivity involved in selecting the appropriate 'Porod region' (Rambo & Tainer, 2011; de Oliveira Neto *et al.*, 2021). To avoid such issues with constant subtraction altogether, we have developed a different approach.

In our approach, we take advantage of the regularization provided above by intentionally oversmoothing using a large α. Oversmoothing has the effect of removing shape scattering while simultaneously enforcing a decay similar to Porod's law of q^{-4}, making the resulting scattering profile more consistent with the assumptions of the Porod law. To do this, we multiply

α by a factor of 10, which in our tests with experimental data resulted in the most accurate and robust results (see *Results* section below). We also limit the q range to $8/R_g$, which has previously been shown to be a reasonable cutoff for calculating Porod volume (Manalastas-Cantos *et al.*, 2021; de Oliveira Neto *et al.*, 2021). Note that this oversmoothing is only applied for calculation of the three parameters mentioned above and their associated errors and does not affect the actual fit of the scattering profile, $P(r)$ curve or other parameters.

2.4.5. Output. Finally, upon exiting the script, the experimental data and calculated fit of the intensities are saved in a file, with the calculated parameter values saved in the header. The corresponding $P(r)$ curve is also saved.

In addition to providing the `denss.fit_data.py` script as an interface to the algorithm described above, other scripts in the *DENSS* package also utilize this algorithm, including `denss.py` and `denss.all.py`, to allow automatic fitting of the data and estimation of D and α when using these programs for *ab initio* 3D density reconstructions.

3. Results

One of the few shapes for which an analytical scattering equation has been derived is the solid sphere (Rayleigh, 1910; Porod, 1982). Since the equation of scattering for a sphere is known exactly, the I_n values for a sphere can be calculated directly (Section S4), resulting in equation (16),

$$I_{n,\,\mathrm{sphere}} = \frac{9}{2}\left(\frac{2}{n\pi}\right)^6 \left\{1 + (-1)^{n+1} + \left(\frac{n\pi}{2}\right)^2 [1 + (-1)^n]\right\}. \quad (16)$$

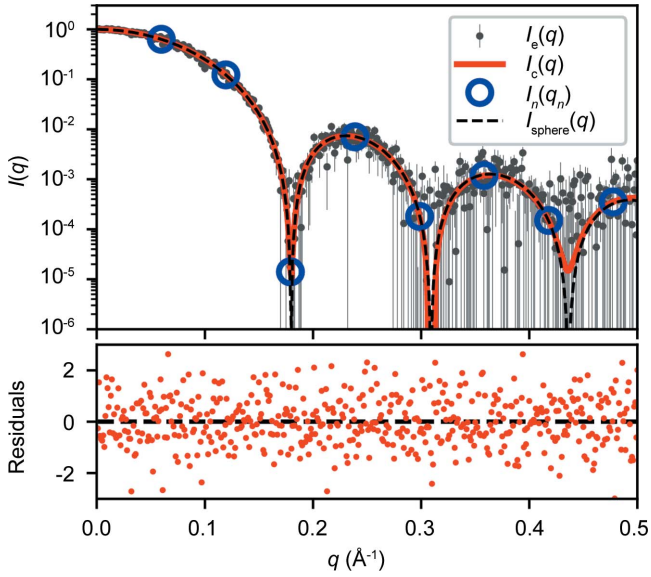


Figure 4
A plot of a calculated intensity curve $I_c(q)$ (red line) fitted to simulated noisy intensity values $I_e(q)$ (grey dots with error bars) for a sphere of radius 25 Å. The blue circles show the Shannon intensities $I_n(q_n)$ and the black dashed line shows the exact scattering profile of the sphere $I_{sphere}(q)$. The bottom plot shows the residuals of the experimental data with respect to the calculated profile.

Table 1
Parameters calculated from I_n values for the sphere profile shown in Fig. 4.

The columns correspond to expected parameter values using an infinite number of Shannon channels and the recovered values calculated from the fit.

Parameter	Expected	Calculated
$I(0)$	1.00	1.00 ± 0.004
R_g (Å)	19.36	19.38 ± 0.18
$\bar{r}$ (Å)	25.71	25.74 ± 0.17
V_p (Å^3)	65450	75838 ± 1338
V_c (Å^2)	277.78	298.52 ± 2.50
ℓ_c (Å)	37.50	40.43 ± 0.79

Note that the radius R of the sphere does not enter into equation (16). Interestingly, the odd I_n values for a sphere decay exactly as q^{-6} and the even I_n values decay exactly as q^{-4}. The decay of intensity at higher angles proportional to q^{-4} is described by Porod's law as mentioned above, generally an approximation for most globular particles but here derived analytically for a sphere for even I_n values.

All parameters outlined above, including R_g, volume *etc.*, can be calculated analytically using equation (16), resulting in well known equations for solid spheres (Section S4). In Fig. 4 the scattering profile for a sphere of radius 25 Å with added Gaussian noise [$I_e(q)$] is shown with the fitted I_n values and the recovered $I_c(q)$ profile. Eight Shannon points were used to fit the data, from which size parameters were calculated using the fitted I_n values, shown in Table 1. The I_n values can also be used to calculate the $P(r)$ curve $P_c(r)$, shown in Fig. 5 along with the exact $P(r)$ curve for a sphere (Porod, 1982) (Section S4).

Data from publicly accessible databases for experimental SAS data, such as BIOISIS (https://www.bioisis.net) and SASBDB (Valentini *et al.*, 2014), are particularly useful for verification and testing of algorithms such as that described here. To test `denss.fit_data.py` on experimental data sets, we downloaded two data sets from the benchmark section of the SASBDB online database, in particular SASDFN8 (apoferritin) and SASDFQ8 (bovine serum albumin) (Graewert *et al.*, 2020). Automated estimates of D and α were

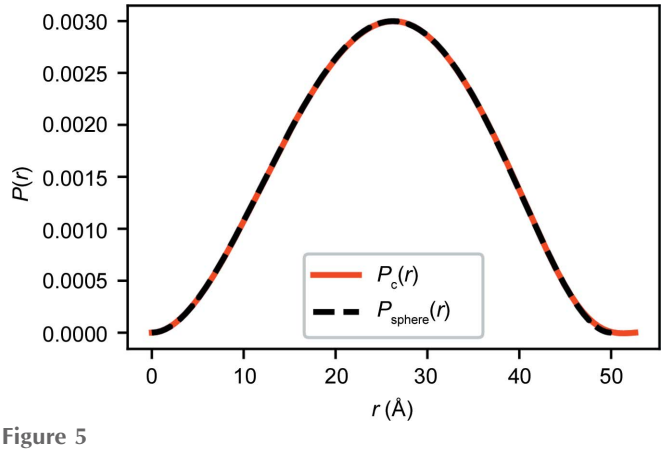


Figure 5
A plot of a calculated $P_c(r)$ curve from I_n values fitted to simulated noisy intensity values for a sphere of radius 25 Å. The exact $P(r)$ curve for the sphere, $P_{sphere}(r)$, is also plotted as a dashed line.

suitable for accurate fitting and parameter estimation, as indicated by the plot of residuals and comparison with the published parameter values (Fig. 6). Best fits are achieved when setting $\alpha = 0$, as expected, and increasing α results in smoother $P(r)$ plots. High-quality fits and smooth $P(r)$ curves can be obtained simultaneously with an appropriate α (Fig. 6), while setting α to too large a value results in poorer fits to the intensity profile. Similar to other IFT methods, a balance must be struck to select the optimal α value resulting in the smoothest $P(r)$ function possible while still enabling a good quality fit of $I(q)$.

To compare the parameter estimates with other software, we used *DATGNOM* from the *ATSAS 3.0* package to estimate R_g and $I(0)$, *DATPOROD* to estimate V_p, and *DATVC* to estimate V_c from these two data sets (Manalastas-Cantos *et al.*, 2021). A comparison of parameter values calculated by *DATGNOM*/*DATPOROD*/*DATVC* and `denss.fit_data.py` is shown in Table 2. Overall, and very importantly for community standards, the values are similar for the two different methods [~0.1% difference for R_g and $I(0)$, and ~3% difference for V_p and V_c]. To verify that the error bounds are estimated correctly, we followed the protocol outlined by Manalastas-Cantos *et al.* (2021) to use the *DATRESAMPLE* program to generate 1000 resampled scattering profiles from the two SASBDB data sets. This allows the calculation of parameters from each resampled profile and subsequently an

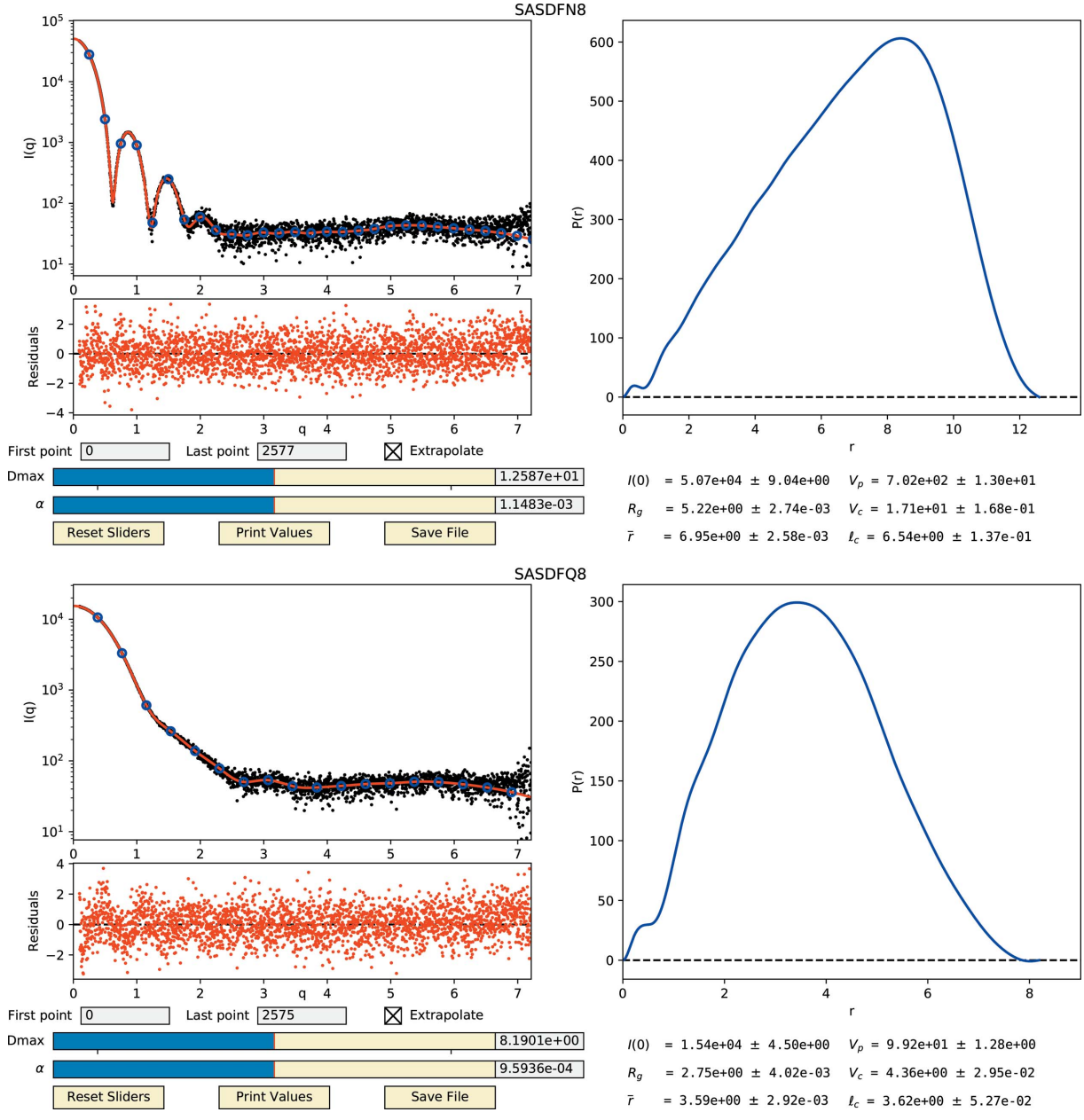


Figure 6
Fitting of I_n values to real experimental data sets while using regularization results in good quality fits, smooth $P(r)$ curves and accurate parameter estimation. GUI displays are given for SASDFN8 (top) and SASDFQ8 (bottom). Note that the experimental q values were given in nanometres, resulting in nanometre units for parameters displayed.

Table 2
Comparison of parameter values calculated from experimental data sets SASDFN8 (DFN8) and SASDFQ8 (DFQ8) using *DATGNOM*/*DATPOROD*/*DATVC* or `denss.fit_data.py`.

Columns correspond to the value calculated for each parameter (Value) and either the estimated [± (Est.)] or statistical [± (Stat.)] errors as described in the text.

		DATGNOM/DATPOROD/DATVC			`denss.fit_data.py`		
Data set	Parameter	Value	± (Est.)	± (Stat.)	Value	± (Est.)	± (Stat.)
DFN8	R_g (nm)	5.216	5.451×10^{-4}	2.931×10^{-3}	5.222	2.835×10^{-3}	5.933×10^{-4}
DFN8	$I(0)$	5.063×10^{4}	9.836	2.306×10^{1}	5.070×10^{4}	9.173	9.061
DFN8	$\bar{r}$ (nm)	N/A	N/A	N/A	6.950	2.644×10^{-3}	7.212×10^{-4}
DFN8	V_p (nm^3)	6.704×10^{2}	N/A	1.905	7.020×10^{2}	1.303×10^{1}	3.752
DFN8	V_c (nm^2)	1.714×10^{1}	N/A	8.717×10^{-3}	1.709×10^{1}	1.678×10^{-1}	2.065×10^{-2}
DFN8	ℓ_c (nm)	N/A	N/A	N/A	6.540	1.374×10^{-1}	2.735×10^{-2}
DFQ8	R_g (nm)	2.745	1.442×10^{-3}	4.122×10^{-3}	2.748	4.199×10^{-3}	1.135×10^{-3}
DFQ8	$I(0)$	1.542×10^{4}	5.665	1.017×10^{1}	1.542×10^{4}	4.570	4.596
DFQ8	$\bar{r}$ (nm)	N/A	N/A	N/A	3.594	3.010×10^{-3}	1.164×10^{-3}
DFQ8	V_p (nm^3)	9.769×10^{1}	N/A	3.367×10^{-1}	9.917×10^{1}	1.278	4.373×10^{-1}
DFQ8	V_c (nm^2)	4.638	N/A	2.890×10^{-3}	4.355	2.949×10^{-2}	4.579×10^{-3}
DFQ8	ℓ_c (nm)	N/A	N/A	N/A	3.624	5.274×10^{-2}	1.227×10^{-2}

estimate of the statistical errors based on the standard deviation of the parameter values, for comparison with the errors estimated by the programs. The results of this analysis are also shown in Table 2. The analysis shows that `denss.fit_data.py` produces similar or smaller statistical errors compared with the estimated errors, suggesting the estimated errors should be considered an upper bound and the statistical errors probably less, whereas the statistical errors appear to be underestimated by *DATGNOM* [note that only R_g and $I(0)$ have estimated errors reported]. It is noteworthy that the statistical errors on R_g and $I(0)$ are smaller from `denss.fit_data.py` (two- to fivefold smaller) than from *DATGNOM*, while the statistical errors on V_p and V_c are about twofold smaller from *DATGNOM*/*DATPOROD*/*DATVC*.

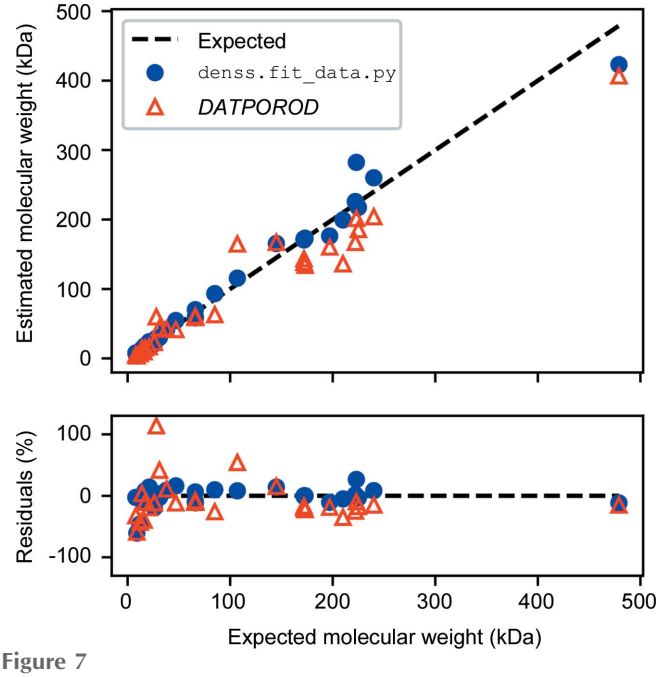


Figure 7
Comparison of expected molecular weight with values calculated using V_p estimated from `denss.fit_data.py` and *DATPOROD*.

The statistical errors described here are only based on resampling the scattering profile and do not account for systematic error that is likely to dominate. As discussed above, V_p, V_c and ℓ_c are particularly sensitive to systematic deviation. To test the algorithm for accuracy with experimental data, we calculated V_p values for 29 data sets from the *Benchmark* section of the SASBDB and used V_p to estimate the molecular weight (MW) of the particle (where MW = V_p/1.6). Fig. 7 shows a comparison of molecular weight values calculated using V_p estimates from `denss.fit_data.py` and *DATPOROD* with their expected values. Here, the expected value is taken from the expected molecular weight in the SASBDB entries calculated from the amino acid sequence. The median error from `denss.fit_data.py` is 8.7% and from *DATPOROD* is 18.0%. As expected, these real errors are in practice significantly larger than the <2% statistical or estimated errors in Table 2, confirming that systematic deviations dominate actual estimates of Porod volume from experimental data.

4. Discussion and conclusions

The approach outlined above is an extension of Moore's original description of SAS profiles using a trigonometric series with the advantage of replacing the nondescript Moore coefficients with specific intensity values. As such, this derivation is subject to all of the same requirements as Moore's, including the need for accurate intensity measurements for at least the first three Shannon channels to obtain reliable estimates of parameter values. We have described a derivation for performing regularization of the real-space $P(r)$ curve analytically, and procedures for the automatic estimation of D and α values. We also present a novel approach for estimating parameters that are particularly sensitive to systematic deviations at high q values, such as V_p.

As in Moore's original approach, the use of least-squares minimization for the derivation given here of a series of SAS parameters directly from the I_n values has enabled the estimation of uncertainties through error propagation while

accounting for covariances in the data. The oversampling of the information content in the SAS profile effectively increases the signal-to-noise ratio of each of the unique observations in the data, *i.e.* the I_n values. Additionally, the analytical regularization derived here simultaneously enables smooth $P(r)$ curves and accurate fits to experimental data, all while providing error estimates for the I_n values and associated parameter calculations, accounting for covariances in the data. Using simulated and experimental data, we have shown that these methods yield parameter values describing the size and shape of particles that are as accurate as, and often more accurate than, existing tools.

The algorithm has been made available open source as a script called `denss.fit_data.py`, accessible on GitHub at https://github.com/tdgrant1/denss. The software can be run either from the command line or as an interactive GUI.

5. Related literature

The following additional references are cited in the supporting information: Fubini (1907); Tonelli (1909).

Acknowledgements

The author thanks Drs Stephen Meisburger, Kushol Gupta and Robert Rambo for testing the software and for useful discussions.

Funding information

Support for this research was provided by the National Institute of General Medical Sciences of the National Institutes of Health (award No. R01GM133998) and by the National Science Foundation through the BioXFEL Science and Technology Center (award No. 1231306).

References

Blackman, R. B. & Tukey, J. W. (1958). *Bell Syst. Tech. J.* **37**, 185–282.
Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering*, 1st ed. New York: Plenum Press.
Fubini, G. (1907). *Rom. Acc. L. R. (5)*, **16**, 608–614.
Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.
Graewert, M. A., Da Vela, S., Gräwert, T. W., Molodenskiy, D. S., Blanchet, C. E., Svergun, D. I. & Jeffries, C. M. (2020). *Crystals*, **10**, 975.
Grant, T. D. (2018). *Nat. Methods*, **15**, 191–193.
Grant, T. D., Luft, J. R., Carter, L. G., Matsui, T., Weiss, T. M., Martel, A. & Snell, E. H. (2015). *Acta Cryst.* D**71**, 45–56.
Guinier, A., Fournet, G., Walker, C. & Yudowitch, K. (1955). *Small-Angle Scattering of X-rays.* Chichester: Wiley.
Hansen, S. & Pedersen, J. S. (1991). *J. Appl. Cryst.* **24**, 541–548.
Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020). *Nature*, **585**, 357–362.
Hunter, J. D. (2007). *Comput. Sci. Eng.* **9**, 90–95.
Konarev, P. V. & Svergun, D. I. (2015). *IUCrJ*, **2**, 352–360.
Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., Panjkovich, A., Mertens, H. D. T., Gruzinov, A., Borges, C., Jeffries, C. M., Svergun, D. I. & Franke, D. (2021). *J. Appl. Cryst.* **54**, 343–355.
Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
Oliveira Neto, M. de, de Freitas Fernandes, A., Piiadov, V., Craievich, A. F., de Araújo, E. A. & Polikarpov, I. (2022). *Protein Sci.* **31**, 251–258.
Petoukhov, M. V., Konarev, P. V., Kikhney, A. G. & Svergun, D. I. (2007). *J. Appl. Cryst.* **40**(s1), s223–s228.
Porod, G. (1982). *Small-Angle X-ray Scattering*, edited by O. Glatter & O. Kratky. London: Academic Press.
Putnam, C. D. (2016). *J. Appl. Cryst.* **49**, 1412–1419.
Rambo, R. (2021). *ScatterIV – New Code Base for Scatter*, https://github.com/rambor/scatterIV.
Rambo, R. P. & Tainer, J. A. (2011). *Biopolymers*, **95**, 559–571.
Rambo, R. P. & Tainer, J. A. (2013). *Nature*, **496**, 477–481.
Rayleigh, Lord (1910). *Proc. R. Soc. London Ser. A*, **84**, 25–46.
Shannon, C. E. (1948). *Bell Syst. Tech. J.* **27**, 379–423.
Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
Svergun, D. I. & Koch, M. H. J. (2003). *Rep. Prog. Phys.* **66**, 1735–1782.
Svergun, D. I. & Pedersen, J. S. (1994). *J. Appl. Cryst.* **27**, 241–248.
Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems.* New York: Winston.
Tonelli, L. (1909). *Rom. Acc. L. R. (5)*, **18**, 246–253.
Tully, M. D., Tarbouriech, N., Rambo, R. P. & Hutin, S. (2021). *J. Vis. Exp.* e61578.
Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2014). *Nucleic Acids Res.* **43**(D1), D357–D363.
Vestergaard, B. & Hansen, S. (2006). *J. Appl. Cryst.* **39**, 797–804.
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. & Vázquez-Baeza, Y. (2020). *Nat. Methods*, **17**, 261–272.