Describing small angle scattering profiles by a limited set of intensities

Thomas D. Grant

May 24, 2021

Abstract

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) [1]. The IFT approach parameterizes the real space pair distribution function ($P(r)$) of a particle using a set of orthogonal 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 [2] which used a trigonometric series to describe the basis functions, where the real space and reciprocal space basis functions are Fourier mates. We present an equation relating the Moore coefficients to the intensities of the SAS profile at specific positions. We provide a series of new equations that determine the size and shape parameters describing a particle from this distinct set of intensity values. We also derive an analytical real space regularizer to smooth the $P(r)$ curve and ameliorate systematic deviations caused by series termination, commonly used in IFT methods [1, 3] though not described in Moore's original approach, which is particularly susceptible to such effects [4, 5]. The algorithm is provided as a script, *denss.fit_data.py*, as part of the DENSS software package for SAS [6], which includes both command line and interactive graphical interfaces.

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 three-dimensional structure of the particle. The scattering of a molecule, $I(q)$, is determined by its three-dimensional electron density function, 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 three-dimensional 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 three-dimensional 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 termination of higher order scattering data to a finite resolution, uncertainties in intensity measurements, and systematic errors, direct calculation of the Fourier transform yields $P(r)$ functions with large systematic deviations [1, 2, 4, 3, 5]. One popular approach to extracting this structural information from SAS profiles is the indirect Fourier transform (IFT) proposed by Glatter [1]. In this approach, a set of basis functions with weights is used to parameterize the $P(r)$ function by optimizing the weights using the fit of the corresponding intensity function to the experimentally obtained intensity profile.

One such IFT algorithm proposed by Moore [2] leverages information theory [7] to describe a set of basis functions defined by the maximum particle dimension, D. Moore uses a trigonometric series to define $P(r)$, resulting in a set of basis functions describing $I(q)$, shown by equations 1 through 3:

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

$$I(q) = \frac{2}{\pi q} \sum_{n=1}^{\infty} a_n B_n(q) \tag{2}$$

$$B_n(q) = \frac{D}{2}\left[\frac{\sin(qD - n\pi)}{qD - n\pi} - \frac{\sin(qD + n\pi)}{qD + n\pi}\right] \tag{3}$$

where a_n are weights, henceforth referred to as the Moore coefficients, B_n are the basis functions, and D is the maximum particle dimension (Note: modest variations compared to Moore's original description of these functions by a factor of 2π are due the use of $q = 4\pi\sin(\theta)/\lambda$ rather than $s = 2\sin(\theta)/\lambda$ and using $P(r)$ and $I(q)$ rather than $Q(r) = P(r)/r$ and $U(q) = qI(q)$). Key to Moore's approach (and other IFT methods [1, 3]) is that the weights a_n define both the real space and reciprocal space profiles, using the appropriate basis functions. Least squares can be used to determine the a_n's and the associated standard errors by minimizing the χ^2 formula (equation 4):

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{I_e(q_i) - I_c(q_i)}{\sigma_i}\right)^2 \tag{4}$$

where I_e is the experimental intensity for data point i, I_c is the intensity calculated at q_i given by equation 2, σ_i is the experimental error on the intensities, and N is the total number of data points. This approach has the advantage of providing the necessary information on the variances and covariances of the

coefficients to determine the errors on each. 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 a_n's 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 [2]. 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 [5, 4], likely due to 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 [8].

In Moore's implementation of the IFT as well as others, the coefficients a_n have no physical meaning other than as weights for the basis functions. Here we extend Moore's derivation to solve the a_n's in terms of physically meaningful intensity values, I_n. The resulting equation for $I(q)$ strictly separates the mathematical description of a particle's size, defined by D, from its shape, defined by the I_n's, and thus enables a more intuitive description of a SAS profile as a function of the size and shape of a particle. We present equations for calculating a variety of commonly used SAS parameters and their associated errors from the I_n's. Additionally, we derive a modified equation for least squares minimization taking into account regularization of the $P(r)$ curve. Finally, we provide open-source software with convenient interfaces for performing all of the presented calculations and show the results using both simulated and real experimental data.

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" [9, 10, 11]. The intensities at q_n, i.e. $I_n = I(q_n)$, therefore become important values as they determine the a_n's and thus can be used to completely describe the low-resolution size and shape of a particle obtainable by SAS. It is therefore convenient to derive the mathematical relationship between I_n and a_n. Combining equations 2 and 3 yields

$$I(q) = \frac{D}{\pi} \sum_{n=1}^{\infty} \frac{a_n}{q} \left[\frac{\sin(qD - n\pi)}{qD - n\pi} - \frac{\sin(qD + n\pi)}{qD + n\pi} \right] \tag{5}$$

The intensity I_m at $q_m = m\pi/D$ is thus

$$I_m = I(q_m) = \frac{D}{\pi}\sum_{n=1}^{\infty}\frac{a_n}{m}\frac{D}{\pi}\left[\frac{\sin((n-m)\pi)}{(n-m)\pi} - \frac{\sin((n+m)\pi)}{(n+m)\pi}\right] \tag{6}$$

Since

$$\left[\frac{\sin((n-m)\pi)}{(n-m)\pi} - \frac{\sin((n+m)\pi)}{(n+m)\pi}\right] = \begin{cases} 0 : n \neq m \\ 1 : n = m \end{cases} \tag{7}$$

the sum reduces to a single term when $m = n$, resulting in

$$I_m = \left(\frac{D}{\pi}\right)^2 \frac{a_m}{m} \tag{8}$$

and therefore

$$a_m = \left(\frac{\pi}{D}\right)^2 m I_m. \tag{9}$$

Equation 9 defines a relationship between the mth Moore coefficient and the intensity at the mth Shannon point. Inserting equation 9 into equation 5 and simplifying yields a general equation for $I(q)$ as a function of the intensity values at the Shannon points:

$$I(q) = 2\sum_{n=1}^{\infty} I_n \frac{(n\pi)^2}{(n\pi)^2 - (qD)^2}\frac{\sin(qD)}{qD}(-1)^{n+1}. \tag{10}$$

Redefining the basis functions B_n as

$$B_n(q) = \frac{(n\pi)^2}{(n\pi)^2 - (qD)^2}\frac{\sin(qD)}{qD}(-1)^{n+1} \tag{11}$$

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

$$I(q) = 2\sum_{n=1}^{\infty} I_n B_n(q). \tag{12}$$

As in Moore's original approach, the B_n functions are determined by the maximum dimension of the particle, D. B_n's for $D = 50$ Å are illustrated in Figure 1.

The $P(r)$ function can be determined from the continuous $I(q)$ according to equation 13:

$$P(r) = \frac{1}{2\pi^2}\int_{q=0}^{\infty} I(q)\frac{\sin(qr)}{qr}dq \tag{13}$$

$P(r)$ can be represented using I_n values by inserting equation 9 into equation 1, resulting in equation 14:

$$P(r) = \frac{r}{2D^2}\sum_{n=1}^{\infty} I_n n \sin\left(\frac{n\pi r}{D}\right) \tag{14}$$

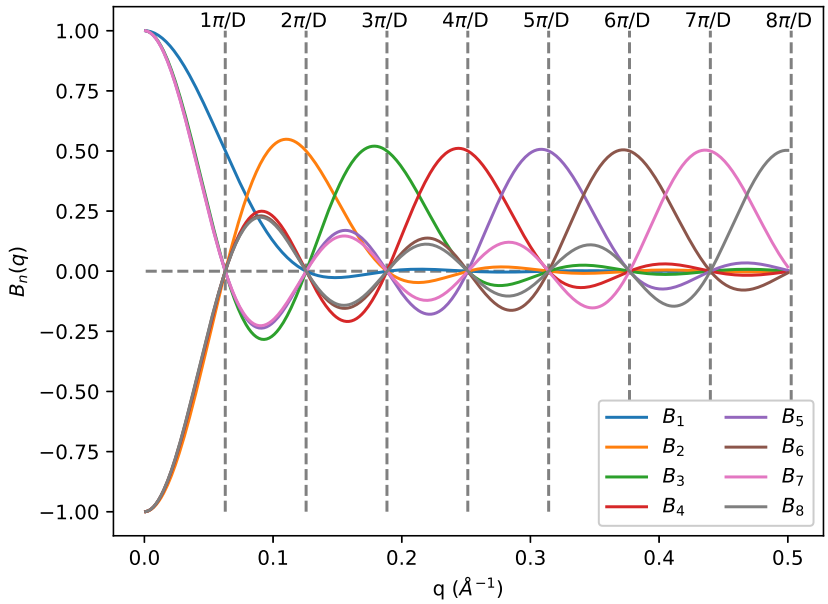

Figure 1: Plot of reciprocal space basis functions, B_n, for a particle of size $D = 50$ Å. Vertical dashed lines show the locations of the Shannon points, q_n.

or by defining real space basis functions S_n as follows:

$$P(r) = \sum_{n=1}^{\infty} I_n S_n(r) \tag{15}$$

$$S_n(r) = \frac{nr}{2D^2} \sin\left(\frac{n\pi r}{D}\right) \tag{16}$$

As intensity values measured precisely at q_n are typically not collected during experiment, least squares minimization of χ^2 can be used to determine optimal I_n's from the oversampled SAS profile, resulting in greatly increased precision for I_n values compared to measured intensities. To determine I_n, let

$$\chi^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \left(I_e(q_i) - 2 \sum_{n=1}^{\infty} I_n B_n(q_i) \right)^2 \tag{17}$$

Values for each I_m are sought which minimize χ^2, i.e. where $\delta\chi^2/\delta I_m = 0$, yielding

$$\sum_i \frac{1}{\sigma_i^2} I_e(q_i) B_m(q_i) = 2 \sum_i \sum_n \frac{1}{\sigma_i^2} I_n B_m(q_i) B_n(q_i). \tag{18}$$

Let

$$Y_m = \sum_i \frac{1}{\sigma_i^2} I_e(q_i) B_m(q_i) \tag{19}$$

and

$$C_{mn} = 2 \sum_i \frac{1}{\sigma_i^2} B_m(q_i) B_n(q_i). \tag{20}$$

then $\boldsymbol{Y} = \boldsymbol{I}\boldsymbol{C}$ and $\boldsymbol{I} = \boldsymbol{Y}\boldsymbol{C}^{-1}$, where $\boldsymbol{I}$ and $\boldsymbol{Y}$ denote the arrays of I_m and Y_m values, respectively, and $\boldsymbol{C}$ is a matrix whose elements are C_{mn}. The I_m's can now be used with equation 12 to provide a smoothed representation of $I(q)$ as well as $P(r)$ using equation 15. The matrix $\boldsymbol{C}^{-1}$ contains all the information on the variances and covariances of $\boldsymbol{I}$. The standard deviation for I_m can thus be calculated from $\boldsymbol{C}$ as

$$\sigma_m = \left[C_{mm}^{-1}\right]^{1/2}. \tag{21}$$

Furthermore the errors on the calculated $I_c(q)$ curve can be calculated as:

$$\sigma_{I_c(q)} = 2 \left[\sum_n \sum_m B_n(q) B_m(q) C_{nm}^{-1} \right]^{1/2} \tag{22}$$

and the errors in $P(r)$ are

$$\sigma_{P(r)} = \left[\sum_n \sum_m S_n(r) S_m(r) C_{nm}^{-1} \right]^{1/2}. \tag{23}$$

The maximum particle dimension, D, is required for determining the q_n values associated with the I_n's. Estimates for the true value of D that are too small will result in B_n's that lack sufficiently high frequencies to adequately reconstruct $I(q)$. Additionally, 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 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 [2]. 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 based on where $P(r)$ gradually falls to zero.

2.2 Derivation of Parameters from I_n's

Similar to what Moore described for the a_n coefficients, since the I_n's 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's. For example, to determine the forward scattering intensity, $I(0)$, we take the limit of equation 10 as q approaches zero to yield

$$I(0) = 2 \sum_{n=1}^{\infty} I_n (-1)^{n+1}. \tag{24}$$

Since the matrix $\boldsymbol{C}^{-1}$ contains all the information on the variances and covariances of the I_n's, the uncertainties in each parameter can be derived using error propagation. The error in $I(0)$ is then found to be

$$\sigma_{I(0)} = 2\left[\sum_n \sum_m (-1)^{n+m} C^{-1}_{nm}\right]^{1/2}. \tag{25}$$

Equation 24 demonstrates a simple relationship between the forward scattering of a particle and the I_n's; note that the particle dimension, D, is not explicitly present in equation 24. Figure 2 illustrates the relationship of the I_n's and $I(0)$. The forward scattering of a particle is not directly measured in an experiment

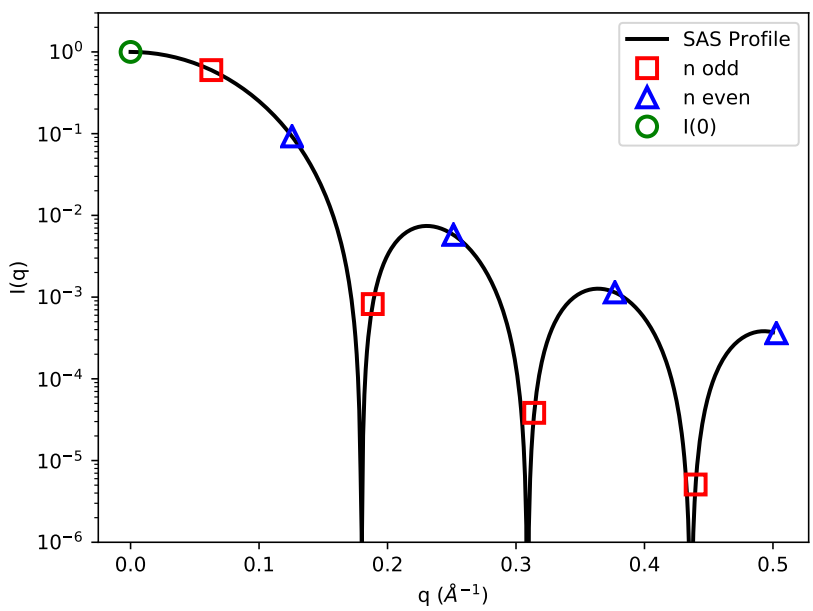

Figure 2: Plot of example scattering profile showing the relationship of I_n's and $I(0)$. Odd I_n's are shown as red squares, while even I_n's (which have a multiplication factor of -1 in equation 24) are shown as blue triangles. Equation 24 says that twice the total sum of the red squares and (negative) blue triangles is equal to the forward scattering, $I(0)$, shown as a green circle.

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 24 provides an alternative method to measuring the forward scattering of a particle directly from the data through the sum of the I_n's. While equation 24 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 24 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 24 should provide an accurate estimate of the forward scattering for most particles and experimental setups.

Other parameters can similarly be derived. For example, R_g can be calculated from the $P(r)$ curve according to the following equation:

$$R_g^2 = \frac{\int r^2 P(r) dr}{2 \int P(r) dr}. \tag{26}$$

Substituting equation 14 into equation 26 yields an estimate for R_g from the I_n's:

$$R_g^2 = \frac{D^2}{I(0)} \sum_{n=1}^{\infty} I_n F_n. \tag{27}$$

where

$$F_n = \left(1 - \frac{6}{(n\pi)^2}\right)(-1)^{n+1}. \tag{28}$$

The error in R_g is then

$$\sigma_{R_g} = \frac{D^2}{2I(0)R_g} \left[\sum_n \sum_m F_n F_m C_{nm}^{-1}\right]^{1/2}. \tag{29}$$

In equation 27 it can be seen that R_g has a non-linear dependence on I_n, and thus the error on R_g is dependent on R_g itself. Another parameter describing particle size is the average vector length in the particle, $\overline{r}$, defined as

$$\overline{r} = \frac{\int r P(r) dr}{\int P(r) dr}. \tag{30}$$

$\overline{r}$ can therefore be estimated from the I_n's as

$$\overline{r} = \frac{4D}{I(0)} \sum_{n=1}^{\infty} I_n E_n \tag{31}$$

where

$$E_n = \left(\frac{(-1)^n - 1}{(n\pi)^2} - \frac{(-1)^n}{2}\right). \tag{32}$$

The error in $\overline{r}$ is

$$\sigma_{\overline{r}} = \frac{4D}{I(0)} \left[\sum_n \sum_m E_n E_m C_{nm}^{-1}\right]^{1/2}. \tag{33}$$

The Porod invariant, Q, is defined as the integrated area under the Kratky plot [12]

$$Q = \int_{q=0}^{\infty} q^2 I(q) dq \tag{34}$$

which, by plugging in equation 12, can be described in terms of the I_n's as

$$Q = \left(\frac{\pi}{D}\right)^3 \sum_{n=1}^{\infty} I_n n^2 \tag{35}$$

with associated error

$$\sigma_Q = \left(\frac{\pi}{D}\right)^3 \left[\sum_n \sum_m (nm)^2 C_{nm}^{-1}\right]^{1/2}. \tag{36}$$

The Porod volume can then be calculated, using its definition containing the Porod invariant [12], by the following equation

$$V_p = \frac{2\pi^2 I(0)}{Q} \tag{37}$$

with error

$$\sigma_{V_p} = \frac{2\pi^2 I(0)}{Q^2}\sigma_Q. \tag{38}$$

The Porod volume is commonly used to estimate molecular weight for globular biological macromolecules. More recently Rambo et. al. [11] derived a new SAS invariant termed the Volume of Correlation defined as

$$V_c = \frac{I(0)}{\int qI(q)dq} = \frac{V_p}{2\pi\ell_c} \tag{39}$$

where ℓ_c is the length of correlation [12]. V_c can be used to estimate molecular weight from macromolecules that may be either globular or flexible [11]. V_c can be estimated from the I_n's as

$$V_c = \frac{D^2 I(0)}{2\pi}\left[\sum_{n=1}^{\infty} I_n n Si(n\pi)\right]^{-1} \tag{40}$$

where $Si(n\pi)$ is the Sine Integral. The error in V_c is then

$$\sigma_{V_c} = \frac{2\pi V_c^2}{D^2 I(0)}\left[\sum_n \sum_m nmSi(n\pi)Si(m\pi)C_{nm}^{-1}\right]^{1/2} \tag{41}$$

The length of correlation can be found from equation 39 or by combining equations 34, 37 and 39 it can be shown that

$$\ell_c = \pi \frac{\int qI(q)dq}{\int q^2 I(q)dq} \tag{42}$$

and therefore

$$\ell_c = \frac{2D}{\pi}\frac{\sum_{n=1}^{\infty} I_n n Si(n\pi)}{\sum_{n=1}^{\infty} I_n n^2} \tag{43}$$

with uncertainty estimated from equation 39

$$\sigma_{\ell_c} = \frac{1}{2\pi V_c}\left[\sigma_{V_p}^2 + \left(\frac{V_p}{V_c}\right)^2 \sigma_{V_c}^2\right]^{1/2} \tag{44}$$

2.3 Regularization of $P(r)$

The original IFT proposed by Glatter [1] as well as other IFTs [3, 13], makes use of regularization of the $P(r)$ curve similar to the general method of Tikhonov regularization [14]. The goal is to use the knowledge that $P(r)$ functions for most particle shapes are smooth 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 45:

$$T = \chi^2 + \alpha S \tag{45}$$

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 by Moore has been shown to be more susceptible than other IFT methods to oscillations in the $P(r)$ curve [4, 5], most likely due to the lack of a regularizing function.

To enable the regularization of the $P(r)$ curve for the derivation described above, S has been chosen to take the commonly used form of equation 46:

$$S = \int_0^D \left[P''(r)\right]^2 dr \tag{46}$$

where $P''(r)$ is the second derivative of $P(r)$ with respect to r. This representation allows for an analytical solution to the problem of regularization of the $P(r)$ curve. To determine the optimal I_n values that yield the best fit of $I(q)$ to the experimental data and a smooth $P(r)$ function, least squares minimization can be used with equation 45. The derivative of S with respect to I_m can be found by combining equation 14 and equation 46 to yield equation 47:

$$\frac{\delta S}{\delta I_m} = \sum_n I_n G_{mn} \tag{47}$$

where

$$G_{mn} = \begin{cases} \dfrac{\pi^2}{2D^5}(mn)^2 \dfrac{m^4+n^4}{(m^2-n^2)^2}(-1)^{m+n} & : m \neq n \\ \dfrac{\pi^2}{48D^5} n^4 \left(2n^2\pi^2 + 33\right) & : m = n \end{cases} . \tag{48}$$

Following the same procedure outlined in equations 17 through 20 and now including 47, equation 45 can now be solved by least squares minimization to yield the optimal I_n values while accounting for the regularizing function S according to the following modified equations:

$$\sum_i \frac{1}{\sigma_i^2} I_e(q_i) B_m(q_i) = \sum_{n=1}^{\infty} I_n \left[\alpha G_{mn} + 2 \sum_i \frac{1}{\sigma_i^2} B_m(q_i) B_n(q_i) \right]. \tag{49}$$

Once again, letting

$$Y_m = \sum_i \frac{1}{\sigma_i^2} I_e(q_i) B_m(q_i) \tag{50}$$

and now letting

$$C_{mn} = \alpha G_{mn} + 2 \sum_i \frac{1}{\sigma_i^2} B_m(q_i) B_n(q_i) \tag{51}$$

then the matrix of I_n values, $\boldsymbol{I}$, can be found in a similar fashion as before:

$$\boldsymbol{I} = \boldsymbol{Y} \boldsymbol{C}^{-\mathbf{1}}. \tag{52}$$

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 [1], Bayesian methods [13] and using perceptual criteria [3].

Equation 49 assumes a sum from $n = 1$ to infinity. However, data are only collected to the maximum n value allowed by experiment. The lack of data for $n > n_{max}$ implicitly corresponds to setting the intensities to zero for those data points where $q > q_{max}$. 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 $q > q_{max}$ intensities towards zero. In order to remove this bias and allow for the I_n values at $n > n_{max}$ to be unrestrained, intensities for $q > q_{max}$ are artificially simulated, with the important step of setting error bars so large as to allow for the I_n values to be effectively unrestrained. As the scattering profile is unknown at these high q values, we use the scattering equation of a sphere whose R_g is equal to that obtained from the experimental data. To ensure the artificial scattering intensities do not bias the fit, we set the error bars for those q values to be several orders of magnitude larger than the corresponding intensities. This allows the I_n values for $n > n_{max}$ to be effectively unrestrained, removing the implicit bias towards zero. To ensure sufficient bias is removed, we simulate the data to $3q_{max}$. As shown below, this extrapolation of the data, coupled with the regularization described above, allows for simultaneously obtaining a smooth $P(r)$ curve while accurately fitting the experimental $I(q)$ profile, even at high q. It is important to note that this extrapolation to $q > q_{max}$ does not necessarily imply that the calculated fit reflects actual intensities at such unmeasured q values, as the resolution of the data are largely dictated by the quality of the data given by the number of Shannon channels that can be reliably extracted from the data, as described in [15].

2.4 Implementation

Tools for performing the least squares fitting of I_n's to experimental data, calculation of parameters and errors, and regularization of $P(r)$, have been developed using Python, NumPy, and SciPy and are provided open source through the DENSS suite of SAS tools [6] (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) has been developed using the Matplotlib package (Figure 3).

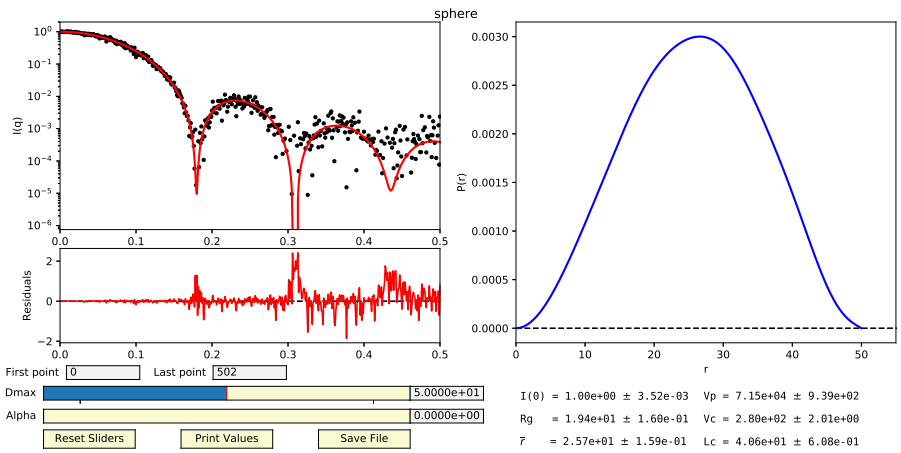

Figure 3: Interactive GUI display from the *denss.fit_data.py* script. The upper left panel shows the experimental $I(q)$ curve as black circles and the fitted $I_c(q)$ calculated from the I_n's as a red curve on a semilog plot. The residuals of the experimental and calculated intensity curves are shown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

To assist users, upon initialization of the script, the experimental data are loaded and an estimate of D is automatically calculated. To automatically estimate D, 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 [16, 17]. An initial rough estimate of R_g is first calculated using Guinier's equation [18] with the first twenty data points. The I_n's are then calculated from the experimental data using equation 52, setting $\alpha = 0$ to optimize the fit to the data. After the initial I_n's are calculated, the corresponding $P(r)$ function often suffers from severe ripples caused by Fourier termination effects due to the limited resolution of

the data detected, as described above, making it difficult to estimate D where $P(r)$ falls to zero. To alleviate this effect, a Hann filter 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 $0.01P_{max}$, where P_{max} is the maximum value of the filtered $P(r)$. This new D value is then used to recalculate the I_n's with equation 52. In addition to automatically estimating D directly from the data, users can manually enter an initial estimate of D to begin.

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's at the experimental q (Figure 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, $log(I_e(q)) - log(I_c(q))$, 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's is also displayed. In addition to manually entering new D and α values in input text boxes 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. Several of the parameters described above, including $I(0)$, R_g, $\overline{r}$, V_p, V_c, and ℓ_c, along with associated uncertainties, are calculated from the I_n's and displayed on the GUI. These parameters are updated interactively whenever D or α are changed.

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 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. The equation governing the scattering of a sphere of radius R is given by equation 53 [12]

$$I_{sphere}(q) = I(0)\left[3\frac{\sin(qR) - qR\cos(qR)}{(qR)^3}\right]^2 \tag{53}$$

(For simplicity, $I(0)$ is here used as a global scale factor accounting for proportionality constants related to volume, concentration, scattering length density, etc., and is set to unity in the following equations). Since the equation of scattering for a sphere is known exactly, the I_n's for a sphere can be calculated by

evaluating equation 53 at $q_n = n\pi/D = n\pi/2R$. Doing so results in equation 54

$$I_{n,sphere} = \frac{9}{2}\left(\frac{2}{n\pi}\right)^6 \left[1 + (-1)^{n+1} + \left(\frac{n\pi}{2}\right)^2 (1 + (-1)^n)\right]. \tag{54}$$

It is important to note that the radius of the sphere, R, does not enter into equation 54. This demonstrates one of the key features of using I_n's to describe scattering profiles. Using this formulation strictly separates the concept of the shape of a particle from its size, where the shape of the particle is described by the I_n's and its size is described by the B_n's, which are dependent on D. Figure 4 illustrates this point for spheres of three different radii. The I_n's are

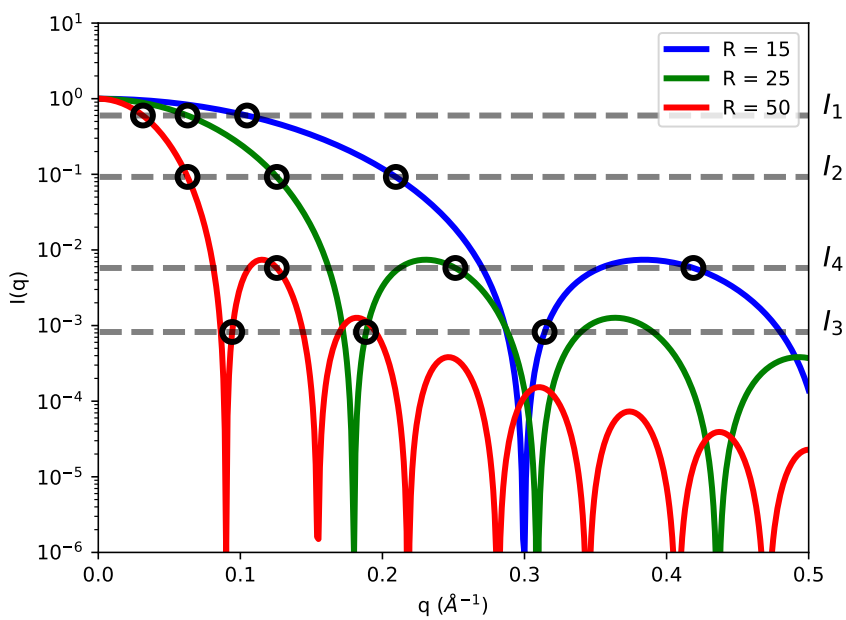

Figure 4: Spheres of different radii have identical I_n values. Calculated $I(q)$ profiles from equation 53 for three spheres of radii 15 Å, 25 Å, and 50 Å are shown as blue, green and red curves, respectively. Small black circles highlight the intensities for the first four q_n values for each curve. Gray horizontal dashed lines show the values for the first four I_n's for a sphere, labeled on the right side. Note that each equivalent q_n for all three spheres lies on the same dashed line, showing that I_n values are independent of the size of the sphere.

identical for each sphere regardless of its radius, illustrated by the dashed lines for each of the first four I_n's and that only the q_n values change, dilated by the size of the sphere. Interestingly, the odd I_n's for a sphere fall exactly as q^{-6} and the even I_n's fall exactly as q^{-4} (Figure 5). The decay of intensity at higher angles proportional to q^{-4} is a well known characteristic of scattering profiles of globular particles known as Porod's Law [12], generally an approximation

for most globular particles but here derived analytically for a sphere for even I_n's. All of the parameters derived above can be expressed for a sphere as

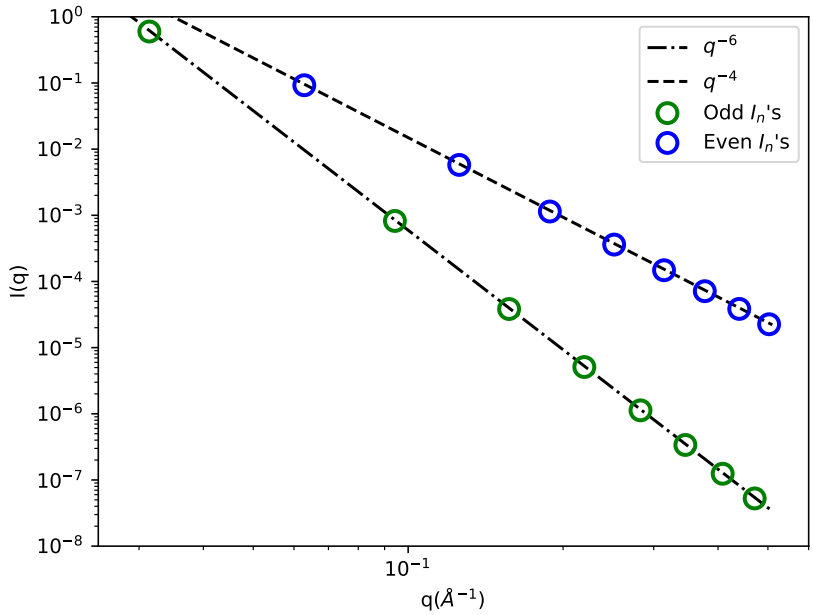

Figure 5: Log-log plot of I_n values for a sphere. Odd I_n's decay exactly as q^{-6}, whereas even I_n's decay as q^{-4}.

well, yielding well-known values for spheres of radius $R = D/2$. For example, equation 27 can be used to calculate the radius of gyration of a sphere of radius R by inserting equation 54 yielding

$$R^2_{g,sphere} = D^2 \sum_{n=1}^{\infty} \frac{9}{2} \left(\frac{2}{n\pi}\right)^6 \left[1 + (-1)^{n+1} + \left(\frac{n\pi}{2}\right)^2 (1 + (-1)^n)\right] F(n) \quad (55)$$

The entire infinite sum converges to $\frac{3}{20}$ and the expression simplifies to

$$R^2_{g,sphere} = \frac{3D^2}{20} = \frac{3}{5}R^2 \quad (56)$$

which is the well-known equation relating the radius of gyration of a solid sphere to its radius. Similarly, the Porod invariant Q of a sphere can be calculated according to equation 35 and shown to be

$$Q_{sphere} = \frac{12\pi}{D^3} = \frac{3\pi}{2R^3} \quad (57)$$

which can be combined with equation 37 to evaluate the Porod volume as

$$V_{p,sphere} = \frac{4\pi}{3} R^3 \tag{58}$$

which is the well-known equation for the volume of a sphere of radius R. Each of the other parameters listed in the equations above can also be solved in a similar fashion and shown to be as follows:

$$\overline{r}_{sphere} = \frac{36}{35} R \tag{59}$$

$$V_{c,sphere} = \frac{4}{9} R^2 \tag{60}$$

$$\ell_{c,sphere} = \frac{3}{2} R \tag{61}$$

In Figure 6 the scattering profile for a sphere of radius 25 Å calculated according to equation 53 with added Gaussian noise ($I_e(q)$) is shown with the fitted I_n values calculated from equation 18 and the recovered $I_c(q)$ profile from equation 12. Eight Shannon points were used to fit the data, from which size parameters

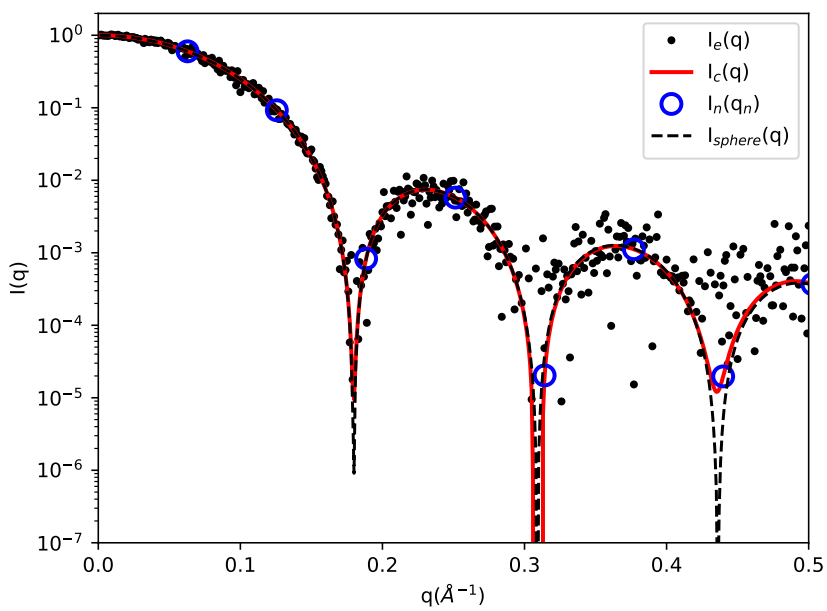

Figure 6: Plot of calculated intensity curve, $I_c(q)$, fitted to simulated noisy intensity values for a sphere of radius 25 Å.

were calculated according to the above equations using the fitted I_n values,

Parameter	Expected ($n = \infty$)	Expected ($n = 8$)	Calculated
$I(0)$	1.00	1.00	1.00 ± 0.004
R_g(Å)	19.36	19.37	19.38 ± 0.167
V_p(Å^3)	65450	70228	71450 ± 939
$\overline{r}$(Å)	25.71	25.73	25.73 ± 0.159
V_c(Å^2)	277.78	279.33	280.08 ± 2.01
l_c(Å)	37.50	40.01	40.60 ± 0.608

Table 1: Table of parameters calculated from I_n values for the sphere profile shown in Figure 6. The columns correspond to expected parameter values using an infinite number of Shannon channels, using eight Shannon channels as dictated by the simulated q_{max}, and the recovered values calculated from the fit.

shown in Table 1. The I_n values can also be used to calculate the $P(r)$ curve according to equation 14. The resulting curve, $P_c(r)$, is shown in Figure 7 along with the $P(r)$ curve for a sphere [12].

$$P_{sphere}(r) = r^2 \left(1 - \frac{3}{2}\left(\frac{r}{D}\right) + \left(\frac{1}{2}\frac{r}{D}\right)^3\right). \tag{62}$$

Data from publicly accessible databases for experimental SAS data, such as BIOISIS (https://www.bioisis.net) and SASBDB [19], are particularly useful for verification and testing of algorithms such as described here. To test *denss.fit_data.py* on experimental data sets, we downloaded two datasets from the benchmark section of the SASBDB online database, in particular SASDFN8 (apoferritin) and SASDFQ8 (bovine serum albumin) [20]. Automated estimates of D were suitable for accurate fitting and parameter estimation as indicated by the plot of residuals and comparison with the published parameter values (Figure 8). 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 (Figure 8), while increasing α to too large of a value results in successively 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 good quality fit of $I(q)$.

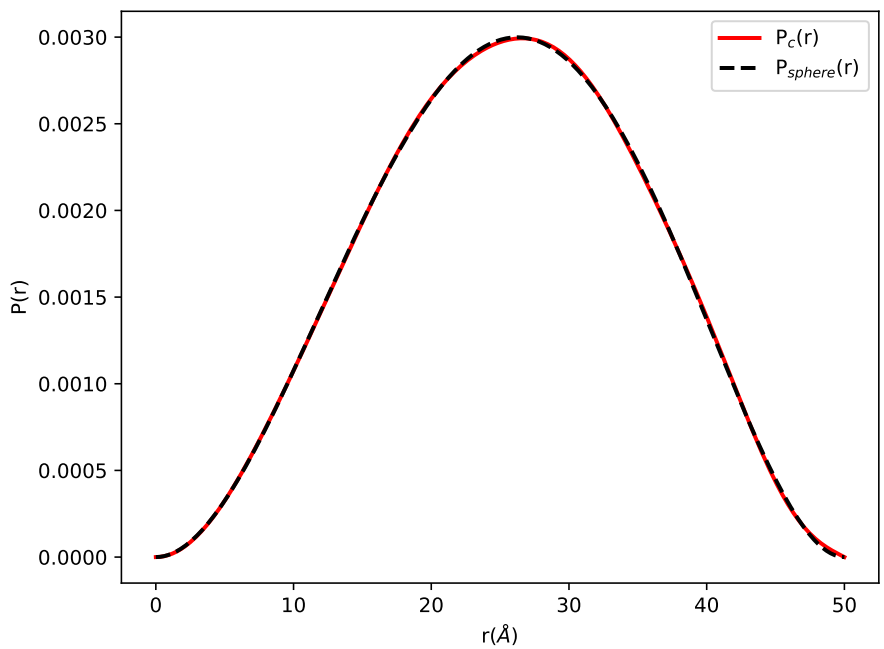

Figure 7: Plot of calculated $P_c(r)$ curve from I_n values fitted to simulated noisy intensity values for a sphere of radius 25 Å. The exact $P(r)$ curve for the sphere, $P_{sphere}(r)$ is also plotted as a dashed line.

3.1 Discussion

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 meaningful 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. There are key advantages of this new approach, however, that ultimately make this mathematical description of the SAS profile more intuitive and may help to understand the relationship between reciprocal space intensities and particle size and shape in real space.

One of the more intuitive aspects of this new approach is the independence of the I_n values on particle size estimator D, resulting in a mathematical distinction of a particle's size from its shape. The distinction between a particle's size and shape is often described qualitatively. R_g is a single descriptor dependent on both the size and shape of a particle, as is average vector length, volume, etc. This fact can be seen in each parameter's corresponding equation above, where both D and I_n are used in the evaluation. A distinction can be seen, for example, when examining the equation for the forward scattering, $I(0)$, where

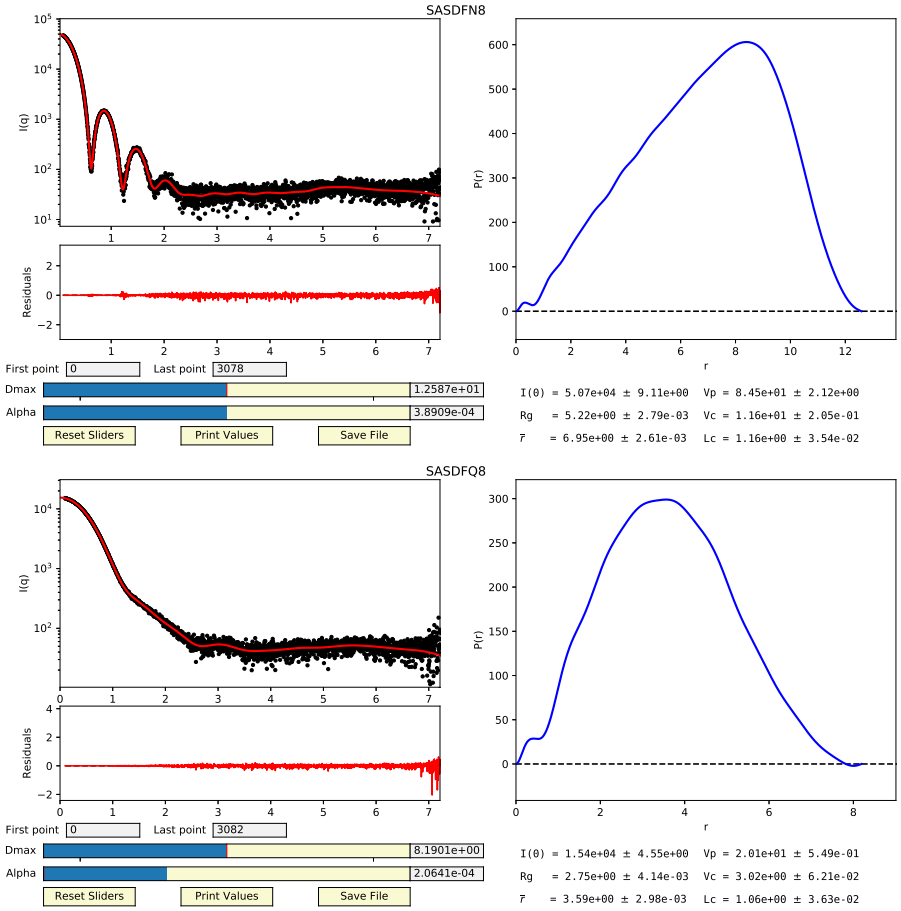

Figure 8: Fitting of I_n values to real experimental datasets while using regularization results in good quality fits, smooth $P(r)$ curves, and accurate parameter estimation. GUI displays of SASDFN8 (top) and SASDFQ8 (bottom). Note that the experimental q values were given in nanometers, resulting in nanometer units for parameters displayed.

the particle size estimator D is noticeably absent, only being dependent on the alternating sum over the I_n's. This reflects the fact that $I(0)$ is dependent only on the excess scattering density of the particle, i.e. the number of electrons scattering X-rays, which can be rearranged arbitrarily in space and have no effect on the forward scattering. Thus while the I_n's may change, the sum of the I_n's according to equation 24 must remain the same.

The mathematical separation of size from shape may facilitate algorithmic developments for looking up particle shapes in databases. Particles with similar shapes but different sizes could be found by looking for similar I_n values (or alternatively relationships between I_n values), rather than complete scattering profiles, and then simply dilating the found shape by the appropriate ratio of

sizes. Similar approaches have been employed using R_g to describe the size of the particle rather than D, by multiplying the q values by R_g and plotting $I(qR_g)$ rather than $I(q)$. This has been used in previous studies to compare scattering profiles from different particles using overall shape characteristics independent of particle size, which has been useful for new analytical methods such as flexibility assessment [21], particle shape matching [22], machine learning [23] and assessing shape ambiguity [24]. In each of those cases, R_g is used as a metric to normalize the size of the particle such that shapes can be directly compared. However, R_g itself is dependent on both the size and shape of the particle. In contrast, the derivation described above mathematically separates particle size from shape using D as the size descriptor, and may be useful in comparing scattering profiles from particles of different size. In practice R_g has proven to be an excellent metric for such purposes as its determination from experimental data is far more precise and robust than D, which often exhibits errors of 10% or more, compared to R_g with errors as little as $1-2\%$. Future research will be required to assess whether or not such large uncertainties in D can be tolerated within algorithms that attempt to normalize for particle size. As mentioned by Moore [2], it would be advantageous to err on the side of slightly overestimating D at the risk of overfitting rather than underestimating D and risk underfitting.

As in Moore's original approach, the use of least squares minimization for the derivation above of a series of SAS parameters directly from the I_n values 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's. Typical experimental data sets have on the order of 500 data points covering approximately ten Shannon channels [15], depending on the size of the particle, instrument geometry, and data quality, oversampling the data some 50-fold. The estimated uncertainty of the intensity at each data point typically ranges from 1% to 100% or more. However, by reducing the data to the unique set of I_n values and their corresponding uncertainties, an estimate of the signal to noise ratio can be obtained for the fitted values, which can be upwards of 20 to 30-fold greater than the measured values as in the example of the sphere given above. Additionally, the analytical regularization derived above simultaneously enables smooth $P(r)$ curves and accurate fits to experimental data, all while providing error estimates for the I_n's and associated parameter calculations accounting for covariances in the data.

The reduction of the SAS profile to the set of I_n values may accelerate computational algorithms. Modeling software must often calculate the scattering profiles of tens or hundreds of thousands of candidate models before converging to a solution. Such calculations are often computationally expensive when using the standard Debye equation [25], requiring a double sum over all atoms and an additional evaluation over all measured q values. Spherical harmonics [26], coarse-grained representations [27, 28], and various other approximations [29] can be used to calculate the scattering profile to simplify the double sum,

but each approach is ultimately a trade-off of accuracy and speed. Using the equations above, significant increases in speed can be realized without reducing accuracy. Rather than calculating intensities at each q value measured by experiment, the modeling algorithm can calculate intensity at just the few I_n values, then use equation 12 to quickly expand $I(q)$ to each of the experimental q values to measure the fit. By reducing the number of intensities needed for measuring the best fit to the experimental data, these calculations can be accelerated by the rate of oversampling, which can be upwards of 50- to 100-fold or more with modern instruments. This may enable not only faster programs but also more advanced algorithmic developments by searching larger pools of models to find more accurate or complex solutions.

3.2 Acknowledgments

Support for this research was provided by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01GM133998 and by the National Science Foundation through the BioXFEL Science and Technology Center under award number 1231306.

References

[1] O. Glatter. "A new method for the evaluation of small-angle scattering data". *Journal of Applied Crystallography* 10 (5) (1977), pp. 415–421.

[2] P. B. Moore. "Small-angle scattering. Information content and error analysis". *Journal of Applied Crystallography* 13 (2) (1980), pp. 168–175.

[3] D. I. Svergun. "Determination of the regularization parameter in indirect-transform methods using perceptual criteria". *Journal of Applied Crystallography* 25 (4) (1992), pp. 495–503.

[4] S. Hansen and J. S. Pedersen. "A comparison of three different methods for analysing small-angle scattering data". *Journal of Applied Crystallography* 24 (5) (1991), pp. 541–548.

[5] D. I. Svergun and J. S. Pedersen. "Propagating errors in small-angle scattering data treatment". *Journal of Applied Crystallography* 27 (3) (1994), pp. 241–248.

[6] T. D. Grant. "*Ab initio* electron density determination directly from solution scattering data". *Nature Methods* 15 (3) (2018), pp. 191–193.

[7] C. E. Shannon. "A mathematical theory of communication". *The Bell System Technical Journal* 27 (3) (1948), pp. 379–423.

[8] M. D. Tully, N. Tarbouriech, R. P. Rambo, and S. Hutin. "Analysis of SEC-SAXS data via EFA deconvolution and Scatter". *Journal of Visualized Experiments* (167) (2021).

[9] L. A. Feigin and D. I. Svergun. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*. 1st ed. New York: Plenum Press, 1987, p. 350.

[10] D. I. Svergun and M. H. J. Koch. "Small-angle scattering studies of biological macromolecules in solution". *Reports on Progress in Physics* 66 (10) (2003), pp. 1735–1782.

[11] R. P. Rambo and J. A. Tainer. "Accurate assessment of mass, models and resolution by small-angle scattering". eng. *Nature* 496 (7446) (2013), pp. 477–481.

[12] G. Porod. *Small angle x-ray scattering / edited by O. Glatter and O. Kratky.* eng. London ; Academic Press, 1982, pp. 17–51.

[13] B. Vestergaard and S. Hansen. "Application of Bayesian analysis to indirect Fourier transformation in small-angle scattering". *Journal of Applied Crystallography* 39 (6) (2006), pp. 797–804.

[14] A. N. Tikhonov and V. Y. Arsenin. "Solutions of ill-posed problems". *New York* 1 (1977), p. 30.

[15] P. V. Konarev and D. I. Svergun. "*A posteriori* determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems". *IUCrJ* 2 (3) (2015), pp. 352–360.

[16] M. V. Petoukhov, P. V. Konarev, A. G. Kikhney, and D. I. Svergun. "*ATSAS* 2.1 – towards automated and web-supported small-angle scattering data analysis". *Journal of Applied Crystallography* 40 (s1) (2007), s223–s228.

[17] T. D. Grant, J. R. Luft, L. G. Carter, T. Matsui, T. M. Weiss, A. Martel, and E. H. Snell. "The accurate assessment of small-angle X-ray scattering data". *Acta Crystallographica Section D* 71 (1) (2015), pp. 45–56.

[18] A. Guinier, G. Fournet, C. Walker, and K. Yudowitch. *Small-angle Scattering of X-rays.* Structure of matter series. Wiley, 1955.

[19] E. Valentini, A. G. Kikhney, G. Previtali, C. M. Jeffries, and D. I. Svergun. "SASBDB, a repository for biological small-angle scattering data". *Nucleic Acids Research* 43 (D1) (2014), pp. D357–D363.

[20] M. A. Graewert, S. D. Vela, T. W. Gräwert, D. S. Molodenskiy, C. E. Blanchet, D. I. Svergun, and C. M. Jeffries. "Adding Size Exclusion Chromatography (SEC) and Light Scattering (LS) Devices to Obtain High-Quality Small Angle X-Ray Scattering (SAXS) Data". *Crystals* 10 (11) (2020), p. 975.

[21] D. Durand, C. Vivès, D. Cannella, J. Pérez, E. Pebay-Peyroula, P. Vachette, and F. Fieschi. "NADPH oxidase activator p67phox behaves in solution as a multidomain protein with semi-flexible linkers". *Journal of Structural Biology* 169 (1) (2010), pp. 45–53.

[22] H. Liu, A. Hexemer, and P. H. Zwart. "The *Small Angle Scattering ToolBox* (*SASTBX*): an open-source software for biomolecular small-angle scattering". *Journal of Applied Crystallography* 45 (3) (2012), pp. 587–593.

[23] D. Franke, C. M. Jeffries, and D. I. Svergun. "Machine Learning Methods for X-Ray Scattering Data Analysis from Biomacromolecular Solutions". *Biophysical Journal* 114 (11) (2018), pp. 2485–2492.

[24] M. V. Petoukhov and D. I. Svergun. "Ambiguity assessment of small-angle scattering curves from monodisperse systems". *Acta Crystallographica Section D* 71 (5) (2015), pp. 1051–1058.

[25] P. Debye. "Zerstreuung von Röntgenstrahlen". *Annalen der Physik* 351 (6) (1915), pp. 809–823.

[26] D. Svergun, C. Barberato, and M. H. J. Koch. "CRYSOL– a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates". *Journal of Applied Crystallography* 28 (6) (1995), pp. 768–773.

[27] S. Yang, S. Park, L. Makowski, and B. Roux. "A Rapid Coarse Residue-Based Computational Method for X-Ray Solution Scattering Characterization of Protein Folds and Multiple Conformational States of Large Protein Complexes". *Biophysical Journal* 96 (11) (2009), pp. 4449–4463.

[28] D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, and A. Sali. "Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments". *Biophysical Journal* 105 (4) (2013), pp. 962–974.

[29] D. Schneidman-Duhovny, S. Kim, and A. Sali. "Integrative structural modeling with small angle X-ray scattering profiles". *BMC Structural Biology* 12 (1) (2012), p. 17.