Reduction of small-angle scattering profiles to finite sets of structural invariants

Jérôme Houdayer^{a,*} and Frédéric Poitevin^{b,c}

^aInstitut de Physique Théorique, Université Paris Saclay, CEA, UMR 3681 du CNRS, Gif-sur-Yvette, France, ^bDepartment of Structural Biology, Stanford, CA 94305, USA, and ^cStanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. *Correspondence e-mail: jerome.houdayer@cea.fr

Received 12 October 2016

Accepted 21 March 2017

Edited by D. A. Keen, STFC Rutherford Appleton Laboratory, UK

Keywords: small-angle scattering; structural invariants; data reduction.

This paper shows how small-angle scattering (SAS) curves can be decomposed in a simple sum using a set of invariant parameters called K_n which are related to the shape of the object of study. These K_n , together with a radius R , give a complete theoretical description of the SAS curve. Adding an overall constant, these parameters are easily fitted against experimental data giving a concise comprehensive description of the data. The pair distance distribution function is also entirely described by this invariant set and the $D_{\max}$ parameter can be measured. In addition to the understanding they bring, these invariants can be used to reliably estimate structural moments beyond the radius of gyration, thereby rigorously expanding the actual set of model-free quantities one can extract from experimental SAS data, and possibly paving the way to designing new shape reconstruction strategies.

1. Introduction

The technique of small-angle scattering in solution (SAS) allows the recording of structural information about virtually any macromolecular object in a wide array of conditions. Since these recordings are rotational averages of the object's scattering, the actual information contained within the SAS curve is limited (Moore, 1980); their usefulness has nonetheless been shown over the years (Doniach, 2001; Yang, 2014). Furthermore, ongoing developments in X-ray sources (Liu & Spence, 2016) have opened the way to the technique of fast solution scattering (FSS), aka correlated X-ray scattering (CXS), which gives access to correlations in the fluctuating scattering (Kam, 1977), thereby increasing the information content (Kirian *et al.*, 2011; Mendez *et al.*, 2014, 2016; Donatelli *et al.*, 2015). Current theoretical developments in FSS/CXS have renewed our interest in the fundamentals of SAS.

A possible approach for relating the SAS curve to the underlying structure of the scattering object consists of comparing the experimental data to a theoretical SAS curve derived from a tentative structural model, in order to assess the extent to which the latter is representative in solution. This *model-to-data* approach has been popularized by the widely used program *CRY SOL* (Svergun *et al.*, 1995), followed by many others (Merzel & Smith, 2002; Zuo *et al.*, 2006; Park *et al.*, 2009; Grishaev *et al.*, 2010; Schneidman-Duhovny *et al.*, 2010; Poitevin *et al.*, 2011; Liu, Morris *et al.*, 2012; Chen & Hub, 2014). The quality of the resulting fitted profile can be used as a score which is minimized as the initial structural model is deformed (Panjkovich & Svergun, 2016), or as one exhaustively samples the structural ensemble around it (Rózycki *et al.*, 2011; Chen & Hub, 2014). *CRY SOL* belongs to the program suite *ATSAS* (Petoukhov *et al.*, 2012) which also

contains programs that address the inverse problem of shape reconstruction directly from the data (Chacón *et al.*, 1998; Svergun, 1999; Walther *et al.*, 2000). In a nutshell, the strategy adopted consists of threading beads on a predefined grid, whose union defines a scattering volume, and displacing them using a trial-and-error process aimed at minimizing the difference between the theoretical curve calculated from the union of beads, and the actual SAS curve. Being able to solve that inverse *data-to-model* problem appears much more desirable than the previously mentioned *model-to-data* approach, for it should allow a model-free description of the data together with a measure of their information content. However, despite their success, grid-based approaches do not explicitly exploit the relationship between the structure of the data and that of the underlying object itself, but rather provide a practical real-space discretization procedure not so different from the structural refinement procedure found in the *model-to-data* approach.

Most recently, another suite of programs, *SASTBX* (Liu, Hexemer & Zwart, 2012), has been made available to the community. Like *ATSAS*, this suite offers both approaches to relate the experimental data to structural models. However, the correspondence between the experimental data and the underlying structure of the probed object is made more explicit, introducing the Zernike formalism. The density function of the object is expanded on a three-dimensional Zernike polynomials basis (Mak *et al.*, 2008), resulting in a hierarchical expansion of the corresponding small-angle X-ray scattering (SAXS) profile. In fact, including higher-order terms in the density expansion results in including finer details that affect regions of higher q in the SAS curve. We were stricken by how natural this expansion appeared to be for SAS data [see *e.g.* Fig. 2 in Liu, Morris *et al.* (2012)]. *SASTBX* offers two answers to the *data-to-model* approach, a shape determination and a shape retrieval procedure. The first has interestingly been developed for FSS data but is very similar to the grid-based approach of *ATSAS* (Liu *et al.*, 2013). The other one, *shapeup* (Liu, Hexemer & Zwart, 2012), is more exciting; the underlying idea of *shapeup* is that SAXS curves can be reduced to a small set of parameters related to the Zernike moments, which are directly related to the shape of the object. Comparing a new set of parameters to an existing database thus allows one to retrieve a set of shapes similar to that of the object studied.

In §2, building on ideas from Zwart and colleagues, we derive an expansion of the SAS curves stemming from a structural description of the scattering object with three-dimensional Zernike polynomials. In §3, we show how this expansion can be useful in practice for the reduction of experimental data in terms of a finite set of structural invariant parameters. We further demonstrate how a hierarchical set of structural moments – starting with the lowest-order radius of gyration and increasing in order to finer shape descriptors until the data disappear in the noise – can be derived from these structural invariants. In the last section, §4, we recapitulate the main results of our derivation and recall their relevance for describing SAS experimental data; we finally

discuss how they could be used to effectively provide information on the shape of the probed object, the expected outcome of SAS experiments.

2. Decomposing the SAS curve

2.1. Mathematical framework and invariances

The mathematical problem we are addressing here is the computation of

$$I_f(q) = \int_S \left| \int_{\mathbb{R}^3} f(\mathbf{r}) \exp(i\mathbf{q} \cdot \mathbf{r}) d\mathbf{r} \right|^2 d\hat{\mathbf{q}}. \quad (1)$$

The real-space function f represents the object of study, in our case the average solvent-subtracted electronic density of the molecule. The integral over space in equation (1) is the Fourier transform of f , and it corresponds to the wave scattered by f evaluated at the wavevector $\mathbf{q}$ in reciprocal space – where we denote q its norm and $\hat{\mathbf{q}}$ its direction. Its norm squared corresponds to the intensity collected at the same wavevector value. Averaging over the unit sphere S is equivalent to performing the average over the randomly oriented ensemble of molecules in the scattering volume, and yields the isotropic intensity function $I_f(q)$ which now only depends on the modulus of the wavevector. As detailed in Park *et al.* (2009), equation (1) is valid in the q range where interparticle interference (low q) and correlation in the fluctuations around f (high q) can be neglected. In other words, it is valid in the limit of infinite dilution and vanishing thermal noise.

This equation presents three invariances which we are going to use to transform it into a more exploitable form: (i) rotational invariance around the origin: clearly a rotation of f around the origin is absorbed by the outer integral of equation (1) and leaves it unchanged, (ii) translational invariance: a spatial translation of f gives a phase in the inner integral which is absorbed by the norm, (iii) scale invariance: if we choose a length R and define

$$g(\mathbf{x}) = f(\mathbf{r}), \quad \text{with} \quad \mathbf{x} = \frac{\mathbf{r}}{R},$$

we then obtain the following scaling equation:

$$I_f(q) = R^6 I_g(t) \quad \text{with} \quad t = qR. \quad (2)$$

All values related to g such as $\mathbf{x}$, t and $I_g(t)$ are dimensionless. We consider the smallest sphere containing the support of f (the circumscribed sphere of the object) and we call its radius $R_{\min}$. Without loss of generality, using the translation invariance, we assume this sphere to be centred at the origin. From there on, we will choose $R \geq R_{\min}$ and the support of g is therefore contained in the unit ball centred at the origin (we call this ball B). This will allow us, later on, to use an adequate representation for g . As we will see later, the scaling length R is related to the maximum extension of the object, $D_{\max}$, and is different from its radius of gyration which we will denote R_1 .

2.2. Moment expansion

Before we proceed with our derivation, let us first explore a well known path. Developing the norm and integrating over S in equation (1) give

$$I_f(q) = 4\pi \int \int_{\mathbb{R}^3} f^*(\mathbf{r}_1) f(\mathbf{r}_2) \frac{\sin(qr_{12})}{qr_{12}} d\mathbf{r}_1 d\mathbf{r}_2, \quad (3)$$

where the star denotes complex conjugation and $\mathbf{r}_{12} = \mathbf{r}_2 - \mathbf{r}_1$. Using the fact that

$$\frac{\sin z}{z} = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n},$$

we get

$$I_f(q) = 4\pi \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} q^{2n} \int \int_{\mathbb{R}^3} f^*(\mathbf{r}_1) f(\mathbf{r}_2) r_{12}^{2n} d\mathbf{r}_1 d\mathbf{r}_2. \quad (4)$$

From the preceding equation, we can write

$$\frac{I_f(q)}{I_f(0)} = 2 \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} M_n(f) q^{2n}, \quad (5)$$

with the rotation- and translation-invariant moment $M_n(f)$ being defined by

$$M_n(f) = \frac{\int \int_{\mathbb{R}^3} f^*(\mathbf{r}_1) f(\mathbf{r}_2) r_{12}^{2n} d\mathbf{r}_1 d\mathbf{r}_2}{2 \left| \int_{\mathbb{R}^3} f(\mathbf{r}) d\mathbf{r} \right|^2} = R^{2n} M_n(g).$$

We now introduce a family of scales $R_n(f)$, $R_1(f)$ being the well known radius of gyration:

$$R_n(f) = \sqrt[n]{M_n(f)} = R R_n(g).$$

This finally allows us to define a set of rotation, translation and scale invariants $H_n(f)$ (for $n \geq 2$) by

$$H_n(f) = \frac{R_n(f)}{R_1(f)} = H_n(g).$$

It is clear that, together with R_1 to set the scale, these invariants allow one to fully describe $I_f(q)/I_f(0)$. We will see in §3.4 that inversely one can extract their values from numerical data on $I_f(q)$. However, it should be stressed already that the $H_n(f)$ do *not* completely characterize the shape of the object of study, since they only contain the same information as $I_f(q)$ [or $p_f(r)$, see §2.3.3 below].

2.3. Invariant expressions of the SAS curve

This section is the central part of our work and relies on technical computations that have been moved to the Appendices.

2.3.1. Zernike expansion and rotational invariance. We introduce three-dimensional Zernike polynomials on the unit ball B to represent g (Canterakis, 1999; Mak *et al.*, 2008). This representation presents nice properties under rotation that will help us proceed with the computation. For completeness and to fix conventions, we recall the main results about three-dimensional Zernike polynomials in Appendix A. Since the support of g is contained in the unit ball, we can write an

expansion of the following form, where the sum over l is limited to $n - l$ even,

$$g(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{l=0}^n \mathbf{c}_{nl}(g) \cdot \mathbf{Z}_{nl}(\mathbf{x}).$$

Before we proceed with the computation of $I_g(t)$, we want to better understand the rotational properties of this representation. In Appendix B, we show that we can define complex rotational invariants by

$$I_{nn'l}(g) = \mathbf{c}_{nl}^*(g) \cdot \mathbf{c}_{n'l}(g).$$

This means that $I_{nn'l}(g)$ does not depend on the orientation of g . The $I_{nn'l}$ are, in addition, insensitive to the relative orientation of the different l layers of the Zernike representation. In equation (37), we also define $\bar{I}_{nn'l}(g)$ as real counterparts to the $I_{nn'l}(g)$. Their importance comes from the following results proved in Appendix B: *any rotation-invariant quadratic form on g is a linear combination of the $I_{nn'l}(g)$ and any rotation-invariant real symmetric quadratic form on g is a linear combination of the $\bar{I}_{nn'l}(g)$.*

Using equation (34) in equation (1), one can derive the result of Liu, Morris *et al.* (2012) where $b_n(t)$ is given by equation (33):

$$I_g(t) = 16\pi^2 \sum_{n,n'=0}^{\infty} (-1)^{(n+n')/2} b_n(t) b_{n'}(t) \sum_{l=0}^{\min(n,n')} (-1)^l I_{nn'l}(g).$$

We will now work our own way from there, starting with some minor rewriting. The sums are restricted such that n , n' and l have the same parity, so $(-1)^l = (-1)^{-n'}$, and we can rewrite the indices n and n' as $n+k$ and $n-k$ with no parity constraints on k . Using equations (33) and (37), this results in

$$I_g(t) = \frac{1}{t^2} \sum_{n=0}^{\infty} \sum_{k=0}^n (-1)^k j_{n+k+1}(t) j_{n-k+1}(t) F_{nk}(g) \quad \text{with} \\ F_{nk}(g) = 16\pi^2 [(2n+3)^2 - 4k^2]^{1/2} \sum_{l=0}^{n-k} \bar{I}_{n-k,n+k,l}(g), \quad (6)$$

where j_n is the spherical Bessel function of order n and where F_{nk} is real and $F_{n0} \geq 0$. Equation (6) is our rotation-invariant expression for $I_g(t)$ which is, as expected, a linear combination of the $\bar{I}_{nn'l}$. Making it explicitly translation invariant is the key to this paper.

2.3.2. Translation-invariant expressions for $I_g(t)$. The first hint of translation invariance is equation (5) since the $M_n(f)$, being the coefficients of the q expansion of $I_f(q)/I_f(0)$, are rotation and translation invariant. We proceed with a t expansion of equation (6). To do this, we must first derive the t expansion of $j_{n-k}(t) j_{n+k}(t)$. The result is given by equation (44) in Appendix C2. Introducing it in equation (6) raises the following relationship, A and B being given by equations (45) and (46):

$$I_g(t) = \sum_{n=0}^{\infty} L_n(g) t^{2n} \text{ with} \\ L_n(g) = \sum_{l=0}^n \sum_{k=0}^l (-1)^k \frac{A_{n+1}^{l+1}}{B_{n+1}^k} F_{lk}(g). \quad (7)$$

Identifying with equation (4) gives

$$L_n(g) = 4\pi \int \int_B g^*(\mathbf{x}_1) g(\mathbf{x}_2) \frac{(-1)^n}{(2n+1)!} x_{12}^{2n} d\mathbf{x}_1 d\mathbf{x}_2 \quad (8)$$

and the following scaling rule:

$$L_n(f) = R^{6+2n} L_n(g). \quad (9)$$

The moments M_n and H_n are related to the L_n coefficients through

$$M_n(g) = \frac{(-1)^n (2n+1)!}{2} \frac{L_n(g)}{L_0(g)}, \quad (10)$$

$$H_n(g) = \sqrt{\frac{(2n+1)!}{3^n} \frac{L_0^{n-1}(g) L_n(g)}{2L_1^n(g)}}. \quad (11)$$

Equation (7) is our first rotation- and translation-invariant expression for $I_g(t)$. The practical interest of this relation is, however, somewhat limited. While it is theoretically possible to extract the L_n from numerical data, we expect it to be quite imprecise as this equation once truncated into a finite sum would only be accurate around $t = 0$ since it would diverge at large t .

The understanding of what has to be done comes from the study of translation invariants made of finite linear combinations of the F_{nk} . It seems, on small values of n , that for each n one can build a unique such invariant K_n containing F_{n0} but no other $F_{n'0}$. From equation (6) and Appendix C3, we derive

Figure 1
Functions $[j_{n+1}^2(t)]/t^2$ for $n = 0, 1, 2, 3$, normalized to a maximum of 1.

$$I_g(t) = \sum_{n=0}^{\infty} K_n(g) \frac{j_{n+1}^2(t)}{t^2} \text{ with} \\ K_n(g) = \sum_{l=0}^n \sum_{k=0}^l \beta_n^{lk} F_{lk}(g) \text{ and} \\ \beta_n^{lk} = (-1)^k \alpha_{n+1}^{l+1,k} = (-1)^k \sum_{p=l+1}^{n+1} \frac{C_{n+1}^p A_p^{l+1} B_p^0}{B_p^k}. \quad (12)$$

Since $\alpha_p^{n0} = \delta_{pn}$, then $\beta_n^{l0} = \delta_{nl}$ and we recover our starting intuition:

$$K_n(g) = F_{n0}(g) + \sum_{l=1}^n \sum_{k=1}^l \beta_n^{lk} F_{lk}(g). \quad (13)$$

We can easily relate the $K_n(g)$ to the $L_n(g)$ by t expanding equation (12) using equation (44) and identifying, and inverting using equation (48):

$$L_n(g) = \frac{1}{B_{n+1}^0} \sum_{l=0}^n A_{n+1}^{l+1} K_l(g), \\ K_n(g) = \sum_{l=0}^n C_{n+1}^{l+1} B_{l+1}^0 L_l(g). \quad (14)$$

Finally, introducing equation (8) in this last equation, we can relate the K_n to $U_n(x)$, even polynomials of degree $2n$:

$$K_n(g) = 4\pi \int \int_B g^*(\mathbf{x}_1) g(\mathbf{x}_2) U_n(\mathbf{x}_{12}) d\mathbf{x}_1 d\mathbf{x}_2, \\ U_n(x) = \sum_{l=0}^n \frac{(-1)^l}{(2l+1)!} C_{n+1}^{l+1} B_{l+1}^0 x^{2l}. \quad (15)$$

Equation (12) is our final explicitly rotation- and translation-invariant expression for $I_g(t)$, and the $K_n(g)$ are rotational and translational invariants. We believe the K_n generate all translational and rotational invariants which are *finite* linear combinations of the $\tilde{I}_{nn'l}$. Note that there exist *infinite* invariants which are *not* generated by the K_n , as exemplified by equations (38) and (39).

As we will see in §3, equation (12) is particularly well adapted to fitting numerical data, since the functions $j_{n+1}^2(t)/t^2$ are bell shaped with well separated maxima as can be seen in Fig. 1. But inspecting equations (14) and (9), it appears that there are no scaling rules for the $K_n(g)$. We thus *cannot* naturally define $K_n(f)$, but we can use the $K_n(g)$ to compute the $H_n(f)$, which are a better choice to describe the shape of the object of study. Indeed, from R and $K_0 \dots K_m$ one can compute L_n , M_n and H_n for $n \leq m$ using equations (10), (11) and (14).

2.3.3. Pair distance distribution function. It is usual (see e.g. Moore, 1980; Liu & Zwart, 2012; Köfinger & Hummer, 2013) when studying $I_f(q)$ to introduce the pair distance distribution function:

$$p_f(r) = \int \int_{\mathbb{R}^3} f^*(\mathbf{r}_1) f(\mathbf{r}_2) \delta(r - r_{12}) d\mathbf{r}_1 d\mathbf{r}_2. \quad (16)$$

The support of $p_f(r)$ is $[0, D_{\max}]$ where $D_{\max}$ is the diameter of the support of f (namely the largest distance between two points with a non-zero f value). A regular tetrahedron mini-

mizes $D_{\max}/R_{\min}$, where $R_{\min}$ has been introduced in §2.1, and thus

$$\frac{1}{2}D_{\max} \leq R_{\min} \leq \frac{1}{2}\left(\frac{3}{2}\right)^{1/2} D_{\max}. \quad (17)$$

From equations (3) and (16) and using equation (43) for j_0 , it is clear and well known that $p_f(r)$ and $I_f(q)$ are related by

$$I_f(q) = 4\pi \int_0^\infty p_f(r) \frac{\sin(qr)}{qr} dr \quad (18)$$

$$p_f(r) = \frac{1}{2\pi^2} \int_0^\infty qr I_f(q) \sin(qr) dq. \quad (19)$$

Interestingly, one can compute $M_n(f)$ knowing $p_f(r)$:

$$M_n(f) = \frac{\int_0^\infty p_f(r) r^{2n} dr}{2 \int_0^\infty p_f(r) dr}. \quad (20)$$

So one can see the H_n as scale-invariant moments of p_f .

Recalling the scaling notations defined in equation (2), we can easily derive the scaling rule, namely $p_f(r) = R^5 p_g(x)$, and we can now apply our previous results. First using equation (17) and $R \geq R_{\min}$, we can conclude that the support of p_g is included in $[0, 2]$. If we use equation (12) in equation (19), we get

$$\begin{aligned} p_g(x) &= \sum_{n=0}^\infty K_n(g) T_n(x), \text{ with} \\ T_n(x) &= \frac{x}{2\pi^2} \int_0^\infty \frac{j_{n+1}^2(t)}{t} \sin(tx) dt \\ &= \frac{1}{4\pi(2n+3)} \left[x^2 - \frac{1}{2} \sum_{l=0}^{n+1} \frac{(-1)^l}{(2l+3)!} C_{n+1}^l B_l^1 x^{2l+3} \right], \end{aligned} \quad (21)$$

for $x \leq 2$ and $T_n(x) = 0$ for $x \geq 2$. We show in Appendix D2 how the T_n were computed. Their shape is shown in Fig. 2.

2.3.4. Porod's theory. In Porod's theory (Porod, 1951) on the scattering of globular objects (*i.e.* objects of constant density), an important quantity that is derived is the Porod's invariant Q_f , defined by

Figure 2
Polynomials $T_n(x)$ for $n = 0, 1, 2, 3$.

$$Q_f = \frac{1}{2\pi^2} \int_0^\infty q^2 I_f(q) dq.$$

For this quantity, the following scaling rule stands:

$$Q_f = R^3 Q_g,$$

so that we can apply our results, namely using equations (12) and (51),

$$\begin{aligned} Q_g &= \frac{1}{2\pi^2} \sum_{n=0}^\infty K_n(g) \int_0^\infty j_{n+1}^2(t) dt \\ &= \frac{1}{4\pi} \sum_{n=0}^\infty \frac{K_n(g)}{2n+3}. \end{aligned}$$

There are three important quantities that can be derived from this: the so-called Porod's volume which is the volume of the object, its surface and a correlation length. They are related to Q_f by

$$\begin{aligned} V_f &= \frac{I_f(0)}{Q_f}, \\ S_f &= \frac{V_f}{2\pi Q_f} \lim_{q \rightarrow \infty} q^4 I_f(q), \\ l_f &= \frac{1}{2\pi Q} \int_0^\infty q I_f(q) dq. \end{aligned}$$

Obviously, we have $V_f = R^3 V_g$, $S_f = R^2 S_g$ and $l_f = R l_g$.

We can now compute these quantities: namely, since $I_g(0) = K_0(g)/9$, we have

$$V_g = \frac{K_0(g)}{9Q_g}.$$

For S_g , we must compute the limit. When $t \rightarrow \infty$

$$j_{n+1}^2(t) \simeq \frac{1}{2t^2} [1 + (-1)^n \cos(2t)]$$

and thus

$$t^4 I_g(t) \simeq \frac{1}{2} \left[\sum_{n=0}^\infty K_n(g) + \cos(2t) \sum_{n=0}^\infty (-1)^n K_n(g) \right].$$

The limit for S_g is hence well defined only when

$$\sum_{n=0}^\infty (-1)^n K_n(g) = 0 \text{ and } S_g = \frac{V_g}{4\pi Q_g} \sum_{n=0}^\infty K_n(g). \quad (22)$$

Finally, using equation (52),

$$\begin{aligned} l_g &= \frac{1}{2\pi Q_g} \sum_{n=0}^\infty K_n(g) \int_0^\infty \frac{j_{n+1}^2(t)}{t} dt \\ &= \frac{1}{4\pi Q_g} \sum_{n=0}^\infty \frac{K_n(g)}{(n+1)(n+2)}. \end{aligned}$$

2.3.5. Extension to $R < R_{\min}$. From equations (15) and (16), we can write

$$K_n(g) = 4\pi \int_0^2 p_g(x) U_n(x) dx. \quad (23)$$

Inserting in equation (12), and noting that the resulting relation is true for all possible p_g , we can thus identify with equation (18) and we deduce that

$$\frac{\sin(tx)}{tx} = \sum_{n=0}^{\infty} U_n(x) \frac{j_{n+1}^2(t)}{t^2}. \quad (24)$$

We have done some numerical experiments and have found the following: this identity seems to be true for any $t, x \in \mathbb{R}$ with a pointwise convergence. On the other hand, the convergence seems to be uniform in t and x only for $|x| < 2$. This is important for exchanging summation and integration, as this is justified only when the convergence of the series is uniform. So, in practice, this formula is usable only when $|x| < 2$. This relation is very interesting because it allows one to separate the variables t and x . Had it been known before, one could have deduced equation (12) directly.

Up to now, we have needed that $R \geq R_{\min}$ in order to write the Zernike expansion for g . Using equation (24) in equation (18), we can recover our K_n expansion of $I_g(t)$ [equation (12)] through equation (23). But this time the derivation is valid as long as the support of p_g is included in $[0, 2]$, that is as long as $R \geq D_{\max}/2$. We can thus conclude that all the formulae of this paper which do not mention the Zernike decomposition of g (through $c_{nl}, I_{nl}, F_{nk}, \dots$) are valid as soon as $R \geq D_{\max}/2$ and not only when $R \geq R_{\min}$ (remember that $R_{\min} \geq D_{\max}/2$).

Moreover, as H_n is constant, the relationship between H_n and g through c_{nlm} is independent of R , and thus valid for any R . When $R < R_{\min}$, this amounts to integrating out of the unit ball. Since one can deduce the K_n from the H_n as soon as $R \geq D_{\max}/2$, the relationship between K_n and g through the c_{nlm} is valid as soon as $R \geq D_{\max}/2$, including when $R < R_{\min}$. In conclusion, all our formulae including those mentioning the Zernike decomposition are also valid when $D_{\max}/2 \leq R < R_{\min}$. Our theory cannot be extended further to $R < D_{\max}/2$ as equation (21) cannot represent p_g with non-zero values for $x \geq 2$.

Observing equation (21) and Fig. 2, it is clear that a pair distance distribution function p_g reaching 0 at $x = 2$ and not before will require fewer K_n parameters to be described precisely. Hence we see that the best value for R is $D_{\max}/2$ as it optimizes the information content of the first K_n . We now see that this choice is allowed and desired.

2.3.6. Non-uniqueness of the invariant expansion. We can alternatively describe the object through the autocorrelation function of its density, also contained in the unit ball. A similar reasoning leads to an expansion on the basis of $[j_{2n+1}(2t)]/t$ functions. Both this new expansion and the one derived previously can be generalized into an infinite set of invariant expansions over the basis of $[j_{n+s}^2(t)]/t^{2s}$ or $[j_{2n+s}(2t)]/t^s$ func-

Figure 3 Illustration of the synthetic (grey) and fitted (black) SAS profiles for four different geometrical shapes. The light grey data were not used to fit the profile. The parameters given on top of each profile are derived from the procedure detailed in §3.3.

tions, respectively, where s can take any integral value. We direct the reader to Appendix D for a more in-depth derivation of these alternative expansions.

3. Application to SAXS data

In the following sections, §3.1 and §3.2, we will illustrate how our K_n expansion can efficiently fit synthetic data. We will then discuss in §3.3 a possible procedure for an optimal choice of our parameters R and N , thus estimating $D_{\max}$. In §3.4, we further describe how the radius of gyration R_1 and moments H_n can be extracted from the K_n . Finally, in §3.5, we check the applicability of our approach and try to fit experimental data taken from the SASBDB online database (Valentini *et al.*, 2015), and extract the invariant parameters from it.

3.1. Synthetic I_f

Synthetic $I_f(q)$ data were generated for four geometrical shapes fitting in a ball of radius 1 nm (hence $D_{\max} = 2$ nm): a cube (with edge $2/3^{1/2}$ nm), a cylinder (with height $4/5^{1/2}$ nm and radius $1/5^{1/2}$ nm), a torus (with outer radius 1 nm and inner radius $1/2$ nm), an ellipsoid (with semi-axes, respectively, 1, 0.8 and 0.6 nm). The data were sampled at 1000 points spanning $q = 0$ to $q = 15$ nm⁻¹, with an added noise of 5%, as shown in Fig. 3 (grey points); additional data up to $q = 20$ nm⁻¹ were produced but not used.

To fit $I_f(q)$ we have to choose a value for R and use equation (2) to compute the corresponding data for $I_g(t)$ where $t = Rq$, and fit them against a truncated version of equation (12), namely

$$I_g^{\text{exp}}(t) \simeq \sum_{n=0}^{N-1} K_n \frac{j_{n+1}^2(t)}{t^2}. \quad (25)$$

The detailed protocol used to determine the parameters R and N is given later in §3.3. For a given choice of these parameters, we applied a regular least-square fitting procedure against equation (25), using only the data for $q < 15 \text{ nm}^{-1}$, to determine the K_n . Results are shown in Fig. 3 (black line). As was shown in Fig. 1, the fitting functions are bell shaped with well separated maxima. Each K_n will thus fit a section of the curve around its maximum which is approximately around $t_n \simeq n$, so that increasing N will allow one to better fit the tail of the curve. In practice, the experimental intensity is often not normalized, and defined up to an unknown scaling factor, different from the one related to our scaling to the unit ball. This factor will be absorbed in all our fitted parameters and their value will thus be defined up to a scaling constant. This will not however affect our results for all the quantities of interest involve their ratio, not their absolute value.

3.2. Synthetic p_g

A strong test of our fits can be performed by comparing the analytical value of $p_g(x)$ and the approximation obtained from the previous fits on $I_g(t)$ through the truncated version of equation (21), recalling our scaling notations $x = r/R$ and $t = Rq$,

$$p_g(x) \simeq \sum_{n=0}^{N-1} K_n T_n(x),$$

with the K_n coming from the fit of $I_g(t)$ [equation (25)]. The corresponding pair distance distribution functions are shown in Fig. 4.

The case of the torus is interesting. Cutting the data to [0:15] yields a large discrepancy in the [15:20] q range, as well as symptomatic oscillations in p_g . This might be a sign that there are not enough data to faithfully reproduce the torus shape,

even with some approximation. Interestingly enough, if the whole [0:20] q range is considered, the situation improves dramatically (data not shown).

We now detail how N and R are determined.

3.3. Choice of R and N , measure of $D_{\max}$

The choice of R and N is an important task. As argued above, theoretically the best choice for R is $R = D_{\max}/2$. First, recall that, in order for our theory to hold and in particular to have equation (12), we need $R \geq D_{\max}/2$. Now, it is easy to argue that given this constraint, smaller R is better. Indeed, as R increases, the support of g shrinks and the data range for I_g increases. In this case, information is being spread to larger t and hence larger n , and thus it is more difficult to gather with the same number of parameters.

Of course, $D_{\max}$ is usually not known beforehand. A good way to measure it is to look at the quality of the fit for a given value of N . On one hand, there is no reason for our fit [equation (25)] to work when $R < D_{\max}/2$; we thus expect a low quality of the fit in this region. On the other hand, as R increases, the spread of the information to larger t values will lead to a decreasing quality of the fit as N is fixed. In conclusion, we expect the best fit quality around $R = D_{\max}/2$.

To measure fit quality, we use the following score:

$$\chi^2 = \frac{1}{N_{\text{exp}} - N - 1} \sum_{i=1}^{N_{\text{exp}}} \left[\frac{d_i - I_g^{\text{fit}}(t_i)}{\sigma_i} \right]^2, \quad (26)$$

where d_i is the data point at $t = t_i$ and σ_i the associated error. This is the quantity we are minimizing for the fit. It should be around 1 for a good fit and larger otherwise.

We show in Fig. 5 the goodness of fit as a function of N and R in the case of the cube mentioned before. The result is quite typical and illustrates our argument. First, it is clear that a

Figure 4

Illustration of the pair distance distribution functions corresponding to Fig. 3. The black line is our estimation and the grey dots are exact values.

curve for a given N is necessarily below the curve for a smaller N [slight crossings may happen due to the overall denominator in equation (26)]. But those curves are constrained by a minimum value which is essentially around $\chi^2 = 1$. As argued before, χ^2 is large at small and large R and reaches a minimum around $R = D_{\max}/2$. At small N , the curves have a well defined minimum; as N increases the curves go down and to the right (to smaller χ^2 and larger R) and eventually reach the minimum possible value around $\chi^2 = 1$. At this stage the bottom part of the curve flattens against this minimum and extends towards larger R values. In the cube case this critical N value where the first plateau appears is $N = 9$. From this behaviour, we can conclude that N should be chosen larger or equal to this critical value to get a good fit and not much larger. Indeed, if we note

Figure 5
Determination of $D_{\max}$ from synthetic data for a cube. Goodness of fit traced as a function of N and R .

($q_{\min}$, $q_{\max}$) the data range for I_f , once R is fixed the data range for I_g is ($Rq_{\min}$, $Rq_{\max}$). Since the n th fitting function has its maximum around $t_n \simeq n$, this gives a maximal value $N_{\max} \simeq Rq_{\max}$ for N ; going further would lead to meaningless values of the corresponding K_n . Actually, $N_{\max}$ is usually too large a choice for N as we start overfitting the data at much lower N which leads to widely fluctuating fit parameters.

Once N is chosen, R should be chosen at the minimum of the corresponding χ^2 curve. This minimum can easily be obtained numerically from the data, even if it is visually not obvious where this minimum is. This gives an estimation of $D_{\max}/2$. Here at $N = 9$, this gives $R = 0.87$ nm and $D_{\max} \simeq 1.74$ nm.

The value of $D_{\max}$ determined for the cube is smaller than the expected exact value of 2 nm. This can be explained by the fact that angular features of the cube cannot be obtained without a very large number of terms in the expansion, beyond what is present in the data: at low resolution, the cube appears eroded, its $D_{\max}$ appears to be smaller and it fits in a sphere smaller than that of the ideal cube. In fact, even with this choice of R the reconstructed p_f is very good for the cube (see Fig. 4), as the exact p_f is extremely flat around $r = 2$ nm.

Although objects with edges as sharp as the cube are unlikely to be found within the pool of biological objects, this observation is noteworthy: the actual value of $D_{\max}$ can only be approached to an extent dictated by the resolution achieved in the available experimental data. Moreover, flat tails in p_f are likely to be missed.

3.4. Measure of R_1 and H_n

The Guinier approximation (Guinier & Fournet, 1955) is the most widely used method to measure the radius of gyration from the data. In this approximation, only data near $q = 0$ are taken into account, leading to experimental complications as this region is particularly sensitive to interparticle interferences. Other methods have thus been proposed that take the full data into account. Glatter (1977) showed that it could be extracted from the approximated pair distance distribution function derived from the data through indirect Fourier

transform methods, leading to a more precise estimation compared with the Guinier approximation. Our method follows a similar approach, as the whole profile is directly used to derive R_1 . Indeed, from the measure of K_0 and K_1 , we obtain a measure of the radius of gyration R_1 through

$$R_1 = R \left[\frac{3}{5} \left(1 - \frac{K_1}{5K_0} \right) \right]^{1/2}. \quad (27)$$

We can likewise compute the H_n up to $n = N - 1$ using equations (11) and (14).

To assess the error made, we measure the second moment H_2 of the cube, for 20 realizations of the added noise. We show in Fig. 6 the deviation to the ‘true’ value as a function of the same parameters N and R . We observe that the variance of the result increases with N , and one gets closer to the exact value when the goodness of fit is smaller. This confirms that one has to choose the smallest value of N for which the goodness of fit plateaus, and pick the value of R at the global minimum of the plateau. We get the same behaviour for the other shapes and other observables (R_1 , H_3 , ...). Here, with $N = 9$ and $R = 0.87$ nm, we determine the value of H_2 with 0.01% precision.

In practice, to compute an estimation of the error, one cannot average on noise. One could use the confidence intervals given by the linear fit method on the K_n parameters. This happens to give much too large error estimations probably due to the fact that our observables, being invariant, have much smaller fluctuations than the K_n from which they are computed.

We suggest the following procedure to estimate errors. We split the available data in a few subsets (we use five) and compute our observable for each subset and extract an error estimation from the variation of the observable between different subsets. The error coming from the imprecision on R is usually negligible, as soon as the χ^2 plateau is sufficiently long. This is important as we thus do not need a precise determination of $D_{\max}$ to get precise determinations of the other observables. We show in Table 1 the results obtained for

Figure 6
Determination of H_2 from synthetic data for the cube. H_2 traced as a function of N and R . The exact solution is shown as a solid horizontal line. Error bars correspond to statistics on 20 realizations of the noise.

Table 1

Results for our synthetic data.

The numbers in parentheses are the errors on the last digit, computed as explained in the text.

Shape	N	R (nm)	R_1 (nm)	H_2	H_3	H_4
Cube	9	0.87	0.5776 (10)	1.3086 (6)	1.4846 (17)	1.6113 (31)
Cube (exact)		1	0.5774	1.3087	1.4848	1.6120
Cylinder	11	0.93	0.6052 (9)	1.3754 (8)	1.6092 (16)	1.7772 (23)
Cylinder (exact)		1	0.6055	1.3755	1.6096	1.7781
Torus	12	0.94	0.7911 (16)	1.3199 (3)	1.4848 (5)	1.5916 (8)
Torus (exact)		1	0.7906	1.3206	1.4862	1.5939
Ellipsoid	11	0.94	0.6324 (12)	1.3200 (7)	1.5079 (18)	1.6449 (30)
Ellipsoid (exact)		1	0.6325	1.3197	1.5073	1.6441

Table 2Examples chosen from the SASBDB database (Valentini *et al.*, 2015).Reference is the SASBDB website inner reference, R_1 is the radius of gyration (in nm) according to the database and likewise for $D_{\max}$.

Short name	Full name	Reference	R_1	$D_{\max}$
Lyz	Lysozyme	SASDAC2	1.5	4.0
Bam	Beta-amylase	SASDA62	4.2	12.7
Apo	Apoferitin	SASDA82	6.8	12.8
Cyt	Cytochrome/adrenodoxin	SASDAP5	2.9	9.5

our four shapes together with the exact values. The estimated errors still seem a little large but these error estimations fluctuate quite a lot from sample to sample and should be handled with care.

These results show that from good data one can extract the radius of gyration R_1 and the H_n invariants with a very good precision. Here overall we achieve a 0.1% precision or better. $D_{\max}$ is more difficult to measure and its determination varies significantly between different realizations of the noise.

3.5. Application to experimental data

We now present four actual experimental examples, for which data for $I_f(q)$ and $p_f(r)$ were available. We tried to take them as representative as possible (see Table 2 and Figs. 7 and 8). We add a global constant K to equation (25), the background correction (Ciccariello *et al.*, 1988), a global shift in the experimental data compared with the theory. Without this global constant, we observe the following: (i) we need a much larger N to obtain the same quality, (ii) the fitted K_n are not stable as we increase N , (iii) the reconstructed p_g is not correct, displaying large oscillations and negative parts. Adding such a constant is quite a generalized practice when fitting against experimental data. We hence have $N + 1$ fitting parameters: one constant K and N parameters K_n .

We use the same technique as before but we have ignored all data with $q < q_{\text{cutoff}}$, where we have chosen q_{cutoff} for each sample to exclude data not well fitted by our model. This is justified as our model does not account for the effective structure factor arising from interparticle interference (Stradner *et al.*, 2004). Ignoring these data is important for our procedure. The quality of the fits is quite good with relatively few fitting parameters.

On the curves of Fig. 8, we can distinguish three parts. (i) The leftmost part around $t = 0$ which is not always well fitted and for which changing N does not help. This is precisely the part that we decided to ignore by setting q_{cutoff} accordingly. (ii) The central part which is covered by the bell-shaped part of the fitting functions is perfectly well fitted by our equation. The n th fitting function is centred around $t_n \simeq n$, so this corresponds roughly to the region $t \lesssim N$. (iii) The tail of the curve, which is not covered by the central part of the fitting functions, may or may not be well fitted by their tails depending on the sample. One can increase N to try to improve that

part. In particular, for our samples, increasing N does visibly help for Apo (figure not shown); for Lyz, Bam and Cyt the tail is already quite well fitted with the current N .

We show the results for p_f corresponding to Fig. 8 in Fig. 9. There is a good agreement even though we are not doing new fits.

A closer look at Fig. 9 reveals three interesting parts on the curves. (i) A central part which is in good agreement with SASBDB. (ii) The part around $x = 0$ that should behave as x^2 for three-dimensional objects [see *e.g.* Köfinger & Hummer (2013)]. This feature seems to be lacking on SASBDB data at least for Lyz and Cyt (which look quite linear at small x). (iii) The right part around $x = 2$ is the tricky part. It tends to oscillate where it should not, and this part is particularly sensitive to the chosen parameters (q_{cutoff} , N and R).

Fig. 10 is analogous to Fig. 5 for Bam data.

Following the procedure we described before, we find in the Bam case $N = 10$ and $R = 5.98$ nm, which gives the estimation $D_{\max} = 11.96$ nm. Proceeding as for the synthetic data, we can compute the radius of gyration and the invariants H_n . We show the results for our four experimental samples in Table 3. In the case of actual data, the goodness-of-fit plot is not always as clearcut as the one for Bam; this was, in particular, the case for Lyz and Cyt where choosing N was not obvious. This may be the signal that the data have some additional experimental features not taken into account in our theory which may or may not influence the measures of the invariants. The error estimations should in any case be handled with care.

4. Discussion

The matter of reducing the SAS curve I_f to a finite set of invariant parameters calls for a discussion about the information content of I_f . A common route (Moore, 1980) consists

Figure 7

The structural models corresponding to the examples discussed here, as found on <https://www.sasbdb.org/>. See also Table 2.

Figure 8
Fits of $I_g(t)$ for various samples. The dots are experimental data and the line is our best fit for the chosen R and N . The fits have $N + 1$ adjustable parameters. The light grey dots were not used for the fits.

Figure 9
Comparison of our estimation of $p_g(x)$ to the ones from SASBDB. The line is our estimation and the dots are data from SASBDB. This is not a fit.

of expanding I_f as a Fourier sine series with finite support $[q_{\min} : q_{\max}]$, and recognizing the pair distance distribution function p_f as its discrete Fourier transform with finite support $[0 : D_{\max}]$. Doing so effectively maps continuous functions to discrete series, and for this mapping to preserve all the initial information the Nyquist–Shannon sampling theorem states that the number of terms in the series must be close to $N_s = D_{\max}(q_{\max} - q_{\min})/\pi$ (Shannon, 1949; Slepian & Pollak, 1961). Further considerations about the effect of noise on the actual information content in the data have been discussed

robust method that allows one to retrieve all the hierarchical set of structural moments H_n that are accessible in a given data set. Starting with the radius of gyration, we show how one can avoid complications linked to the Guinier approximation and propose a way to estimate it based on all the accessible data. Going further, we leave for future work the careful study of how the H_n values relate to typical shapes, in the vein of what has been done by Guinier & Fournet (1955) and Pedersen (1997), for example. For a simple shape like a ball of uniform density, using equation (20) with $p_g \propto T_0$, we easily find that

extensively in the fields of optics (Frieden, 1971), and more recently of SAS (Vestergaard & Hansen, 2006; Pedersen *et al.*, 2014).

Although the sine series expansion is appealing, we believe that the expansion proposed here in equation (12) is more natural for SAS data. In fact, there exists g_n such that $K_n(g_m) = \delta_{nm}$, for example

$$g_n = \frac{1}{4\pi[2(2n+3)]^{1/2}} [Z_{nn}^{-n} + (-1)^n Z_{nn}^n]$$

which means that the associated fitting functions $f_{n+1}^2(t)/t^2$ have the shape of a possible output. In particular, they are positive and have the right decreasing at infinity (namely in t^{-4}). It is the only expansion in our sets with such properties.

Furthermore, unlike the approach presented in this paper, the traditional route does not provide a direct relationship between the SAS data and the underlying scattering object. We claim that the invariant expansion proposed in this work provides both a fair estimate of the number N of degrees of freedom available in the scattering data and a fair estimate of the object's maximum diameter $D_{\max}$, and provides us with a set H_n of model-free parameters beyond the radius of gyration R_1 that reflects the distribution of electrons in the corresponding scattering object.

Recently, Rambo & Tainer (2013) have proposed that the volume of correlation, actually homogeneous to a surface as it is defined as the ratio between the Porod's volume and the correlation length, together with the radius of gyration are a very stringent set of model-free parameters for relating the SAS curve to the overall shape of the scattering object. We presented here a rigorous generalization of that intuition, by deriving a

Table 3

Parameters and results for our samples.

q_{cutoff} is the cutoff (in nm^{-1}) chosen for ignoring data near $q = 0$ and N is the number of K_n parameters in the fit. N_s is the Shannon number. D_{max} (in nm) is measured using the procedure described in the text using figures such as Fig. 10. χ^2 is the fit quality for the chosen R and N . R_1 is the radius of gyration (in nm) and H_n is dimensionless. The numbers in parentheses are estimations of the errors on the last digits.

Sample	q_{cutoff}	N	N_s	D_{max}	χ^2	R_1	H_2	H_3	H_4
Lyz	0.6	5	8	4.08	0.82	1.4706 (21)	1.3257 (5)	1.5182 (10)	1.6575 (14)
Bam	0.4	10	22	11.96	0.96	4.0444 (39)	1.3395 (3)	1.5388 (6)	1.6829 (10)
Apo	0.4	8	23	13.08	1.98	5.2211 (9)	1.2910 (1)	1.4429 (1)	1.5441 (1)
Cyt	0.5	14	14	9.12	2.04	2.8835 (72)	1.3857 (3)	1.6290 (5)	1.8091 (6)

$$H_n(\text{ball}) = 2 \left(\frac{5}{3} \right)^{1/2} \sqrt{\frac{9}{(n+2)(n+3)(2n+3)}}$$

which in particular gives $H_2 \simeq 1.300$, $H_3 \simeq 1.465$ and $H_4 \simeq 1.578$. These values are closer to those of Lyz, or even more Apo (see Table 3) in qualitative accordance with their structural models (see Fig. 7). A more general shape classification protocol might be needed to extend our ability to describe, differentiate and recognize more elaborate shapes such as those of Bam and Cyt in the same figure.

Beyond their use for comparing known shapes, one might ask how our set of invariant parameters could help design *ab initio* shape reconstruction strategies directly from the data – the *data-to-model* approach defined in §1. It would be appealing, for computational efficiency, to be able to derive analytical shapes from the data, rather than having to resort to more expensive numerical exploration strategies, which owe their success to a clever interplay between real-space and intensity-space refinement (Elser, 2011; Liu *et al.*, 2013). We recall the following result from §2.3.1 that any rotation-invariant quadratic form on g is a linear combination of the rotational invariants $I_{nl}(g)$ which are insensitive to the relative orientation of the different l layers of the Zernike representation. As such, $I_g(t)$ is not sensitive to the relative orientation of the different l layers of the Zernike representation. Strategies for the reconstruction of the original shape from $I_g(t)$ will have to overcome this insensitivity. And

Figure 10

Quality of the fit for Bam as a function of R for various N values. Small values of χ^2 around 1 represent a good fit, large ones a bad fit.

no help can come from additional quadratic constraints for example on $\int_B |g(\mathbf{x})|^2 d\mathbf{x}$ or $\int_B |\nabla g(\mathbf{x})|^2 d\mathbf{x}$ since they also suffer from the same insensitivity. Clever approaches must then be derived to bypass these theoretical limitations, *e.g.* by introducing interdependencies between the layers of the Zernike expansion.

Finally, we believe that this work could prove helpful for automatic merging of SAS data sets (Spill *et al.*, 2014) where our expansion could be used instead of the parametric prior function, and for the definition of better ways to solve the problem of overfitting and metric bias in evaluating model-data agreement (Rambo & Tainer, 2013; Franke *et al.*, 2015). We also hope to extend that work to the field of FSS or single-particle imaging (Ma & Liu, 2016) in the near future.

APPENDIX A

Three-dimensional Zernike polynomials on the unit ball

For completeness and to fix notation we recall here the main definitions and properties of three-dimensional Zernike polynomials; for more details see Canterakis (1999), Mak *et al.* (2008), Mathar (2008) and Liu, Morris *et al.* (2012).

Zernike polynomials form a polynomial basis to represent functions on the unit ball. They have the following form

$$Z_{nl}^m(\mathbf{x}) = R_{nl}(x)Y_l^m(\hat{\mathbf{x}}),$$

where $n, l \in \mathbb{N}$ and $m \in \mathbb{Z}$, with $n \geq l$, $n - l$ even and $|m| \leq l$. Y_l^m is the usual spherical harmonics,

$$Y_l^m(\theta, \varphi) = \left[\frac{(2l+1)(l-m)!}{4\pi(l+m)!} \right]^{1/2} P_l^m(\cos \theta) \exp(im\varphi),$$

where P_l^m is the usual associated Legendre polynomials. The radial part is constrained by

$$\int_0^1 x^2 R_{nl}(x) R_{n'l'}(x) dx = \delta_{nn'} \delta_{ll'} \quad (28)$$

and is given by

$$R_{nl}(x) = \sum_{\nu=0}^k q_{kl}^{\nu} x^{l+2\nu},$$

$$q_{kl}^{\nu} = (-1)^{k+\nu} (4k+2l+3)^{1/2} \frac{\binom{2k}{k} \binom{k}{\nu} \binom{2(k+l+\nu)+1}{2k}}{2^{2k} \binom{k+l+\nu}{k}}, \quad (29)$$

where we have set $k = (n - l)/2$, which is an integer. And we have

$$Z_{nl}^{-m}(\mathbf{x}) = (-1)^m Z_{nl}^{m*}(\mathbf{x}). \quad (30)$$

Thus defined, $Z_{nl}^m(\mathbf{x})$ are polynomials of total degree n in the Cartesian coordinates. They form an orthonormal basis for the functions on the unit ball.

Since the support of g is contained in the unit ball B , we can write its expansion in the Zernike basis, namely

$$g(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{l=0}^n \sum_{m=-l}^l c_{nl}^m(g) Z_{nl}^m(\mathbf{x}), \quad (31)$$

where the sum on l is limited to l such that $n - l$ is even. And the $c_{nl}^m(g)$ are given by

$$c_{nl}^m(g) = \int_B g(\mathbf{x}) Z_{nl}^{m*}(\mathbf{x}) d\mathbf{x}.$$

To simplify notations, $\mathbf{c}_{nl}$ will stand for the column vector $c_{nl}^{-l}, \dots, c_{nl}^l$ (with $2l + 1$ elements) and likewise for $\mathbf{Z}_{nl}$ and $\mathbf{Y}_l$. Equation (31) then becomes

$$g(\mathbf{x}) = \sum_{n=0}^{\infty} \sum_{l=0}^n \mathbf{c}_{nl}(g) \cdot \mathbf{Z}_{nl}(\mathbf{x}),$$

still with $n - l$ even.

We also need the Fourier transform of g : using the plane wave expansion

$$\exp(i\mathbf{t} \cdot \mathbf{x}) = 4\pi \sum_{l=0}^{\infty} i^l j_l(tx) \mathbf{Y}_l(\hat{\mathbf{t}}) \cdot \mathbf{Y}_l^*(\hat{\mathbf{x}}),$$

where j_l is the spherical Bessel function of order l , the orthonormality of the spherical harmonics and Mathar's result (Mathar, 2008)

$$\int_0^1 x^2 R_{nl}(x) j_l(tx) dx = (-1)^{(n-l)/2} b_n(t), \quad (32)$$

with

$$b_n(t) = (2n + 3)^{1/2} \frac{j_{n+1}(t)}{t}, \quad (33)$$

one can easily deduce (see Liu, Morris *et al.*, 2012) that

$$\int_B g(\mathbf{x}) \exp(i\mathbf{t} \cdot \mathbf{x}) d\mathbf{x} = 4\pi \sum_{n=0}^{\infty} b_n(t) \sum_{l=0}^n (-1)^{(n-l)/2} i^l \mathbf{c}_{nl}(g) \cdot \mathbf{Y}_l(\hat{\mathbf{t}}). \quad (34)$$

APPENDIX B

Rotational invariance of quadratic forms

B1. Effect of rotations on Zernike expansions

Since we are interested in rotational invariance, it is important to understand the behaviour of the Zernike representation under rotation. It is well known that under a rotation R around the origin, spherical harmonics behave as follows:

$$R(Y_l^m) = \sum_{p=-l}^l \rho_l^{mp}(R) Y_l^p$$

where $\rho_l(R)$ is a $(2l + 1)$ unitary matrix. We can write this more compactly as

$$R(\mathbf{Y}_l) = \rho_l(R) \cdot \mathbf{Y}_l.$$

Moreover ρ_l is an irreducible representation of $\text{SO}(3)$ of dimension $2l + 1$. From this, we can infer

$$\begin{aligned} R(\mathbf{Z}_{nl}) &= \rho_l(R) \cdot \mathbf{Z}_{nl}, \\ \mathbf{c}_{nl}(Rg) &= \rho_l^*(R) \cdot \mathbf{c}_{nl}. \end{aligned}$$

B2. Rotational invariants

We can thus define rotational invariants by

$$I_{nn'l}(g) = \mathbf{c}_{nl}^*(g) \cdot \mathbf{c}_{n'l}(g) = \sum_{m=-l}^l c_{nl}^{m*}(g) c_{n'l}^m(g),$$

with $n - l$, $n' - l$ and $n - n'$ even. From this definition, it is clear that $I_{nn'l}(g) = I_{n'n'l}^*(g)$ and $I_{nn'l}(g) \geq 0$. From equation (30), we can deduce that $I_{nn'l}(g)$ is real if g is real. Since $\rho_l(R)$ is unitary, the $I_{nn'l}$ are indeed invariants; namely we have

$$I_{nn'l}(Rg) = I_{nn'l}(g).$$

This invariance is in fact even deeper: if, for any family $\{R_s\}$ of rotations, we note

$$\{R_s\}g = \sum_{n=0}^{\infty} \sum_{l=0}^n \mathbf{c}_{nl}(R_s g) \cdot \mathbf{Z}_{nl},$$

then

$$I_{nn'l}(\{R_s\}g) = I_{nn'l}(g).$$

This means that those invariants are also insensitive to the relative orientation of the different l layers of the Zernike representation.

B3. Rotational invariance of quadratic forms

The importance of these invariants comes from the following result: *any quadratic form on g which is invariant by rotation is a linear combination of the $I_{nn'l}$.*

Indeed, such a form is written

$$\begin{aligned} Q(c) &= {}^t c^* Q c \\ &= \sum_{l=0}^{\infty} \sum_{n,n'=l}^{\infty} {}^t \mathbf{c}_{nl}^* Q_{nn'l} \mathbf{c}_{n'l} \\ &\quad + \sum_{l=0}^{\infty} \sum_{n=l}^{\infty} \sum_{l'=0, l' \neq l}^{\infty} \sum_{n'=l'}^{\infty} {}^t \mathbf{c}_{nl}^* Q_{nn'l'} \mathbf{c}_{n'l'}, \end{aligned} \quad (35)$$

where the $Q_{nn'l}$ and $Q_{nn'l'}$ are the blocks of the Q matrix. If Q is invariant by rotation then $Q(Rc) = Q(c)$ which implies that $Q_{nn'l} \rho_l(R) = \rho_l(R) Q_{nn'l}$ and $Q_{nn'l'} \rho_{l'}(R) = \rho_{l'}(R) Q_{nn'l'}$. Irreducibility of the ρ_l and Schur's lemma give that $Q_{nn'l} = q_{nn'l} \text{Id}$ for some constant $q_{nn'l}$ and $Q_{nn'l'} = 0$ and hence

$$Q(c) = \sum_{l=0}^{\infty} \sum_{n,n'=l}^{\infty} q_{nn'l} I_{nn'l}(c), \quad (36)$$

as claimed.

A corollary to this is that *any rotation-invariant quadratic form on g* is insensitive to the relative orientation of the l layers of the Zernike representation.

B4. Real symmetric quadratic forms

For real symmetric quadratic forms, $q_{nn'l}$ is real and equation (36) reduces to

$$Q(c) = \sum_{l=0}^{\infty} \sum_{n=l}^{\infty} \sum_{n'=n}^{\infty} q_{nn'l} \bar{I}_{nn'l}(c),$$

where

$$\bar{I}_{nn'l} = \begin{cases} I_{nnl} & \text{if } n = n', \\ I_{nn'l} + I_{n'n'l} = 2\text{Re}(I_{nn'l}) & \text{otherwise.} \end{cases} \quad (37)$$

Thus, *any real symmetric rotation-invariant quadratic form is a linear combination of the $\bar{I}_{nn'l}$* .

Here is the expression for two common real symmetric quadratic forms:

$$\int_B |g(\mathbf{x})|^2 d\mathbf{x} = \sum_{l=0}^{\infty} \sum_{n=0}^{\infty} \bar{I}_{nnl}(g), \quad (38)$$

$$\int_B |\nabla g(\mathbf{x})|^2 d\mathbf{x} = \sum_{l=0}^{\infty} \sum_{n=l}^{\infty} \sum_{n'=n}^{\infty} S_{nn'l} \bar{I}_{nn'l}(g) \quad (39)$$

with as always n, n' and l with the same parity, and

$$S_{nn'l} = \frac{1}{2} [n(n+3) - l(l+1)] [(2n+3)(2n'+3)]^{1/2}.$$

Note that actually computing $S_{nn'l}$ generally is a challenge, but we have inferred the formula for small n and n' ; it is correct at least for $n, n' \leq 20$ (and thus probably always correct).

APPENDIX C

Properties of spherical Bessel functions

C1. Useful relations

The spherical Bessel function j_n is related to the usual Bessel function through

$$j_n(t) = \left(\frac{\pi}{2t}\right)^{1/2} J_{n+1/2}(t).$$

In particular

$$j_0(t) = \frac{\sin(t)}{t}.$$

It also verifies the following recursion relations:

$$\frac{2n+1}{t} j_n(t) = j_{n-1}(t) + j_{n+1}(t), \quad (40)$$

$$(2n+1) \frac{dj_n(t)}{dt} = nj_{n-1}(t) - (n+1)j_{n+1}(t). \quad (41)$$

It has the following series expansion:

$$j_n(t) = \sum_{l=0}^{\infty} (-1)^l X_n^l t^{n+2l} \quad \text{with} \quad X_n^l = \frac{1}{[2(n+l)+1]!!(2l)!!} \quad (42)$$

where $!!$ denotes the double factorial [$n!! = n(n-2)\dots$ and $0!! = 1$]. Finally, it obeys orthogonality relations, which can be formally written as

$$\int_0^{\infty} t^2 j_n(at) j_n(bt) dt = \frac{\pi}{2a^2} \delta(a-b). \quad (43)$$

C2. t Expansion

We want to find the t expansion for $j_{n+k}(t)j_{n-k}(t)$, so we start from equation (42). We thus get

$$j_{n+k}(t)j_{n-k}(t) = \sum_{s=0}^{\infty} t^{2(s+n)} (-1)^s W_{nks},$$

with

$$W_{nks} = \sum_{l=0}^s X_{n+k}^l X_{n-k}^{s-l}.$$

We have not been able to go further analytically but a close scrutiny of the first values of W_{nks} led us to

$$W_{nks} = \frac{\binom{2(n+s)+1}{s}}{[2(n-k+s)+1]!![2(n+k+s)+1]!!}.$$

We have checked this expression systematically for $n, k, s \leq 100$, so that we are certain it is correct. From this, we deduce the miraculously simple formula

$$j_{n-k}(t)j_{n+k}(t) = \sum_{l=n}^{\infty} \frac{A_l^n}{B_l^k} t^{2l}, \quad (44)$$

with

$$A_l^n = (-1)^{l-n} \binom{2l+1}{l-n}, \quad (45)$$

$$B_l^k = [2(l-k)+1]!![2(l+k)+1]!!. \quad (46)$$

Note that A_l^n does not depend on k and B_l^k does not depend on n .

C3. j_n^2 Expansion

Here we prove the following relation used to derive equation (12):

$$j_{n-k}(t)j_{n+k}(t) = \sum_{l=n}^{\infty} \alpha_l^{nk} j_l^2(t). \quad (47)$$

We can constrain the α_l^{nk} using equation (44) on both sides:

$$\begin{aligned} j_{n-k}(t)j_{n+k}(t) &= \sum_{l=n}^{\infty} \alpha_l^{nk} j_l^2(t) \\ &= \sum_{l=n}^{\infty} \alpha_l^{nk} \sum_{p=l}^{\infty} \frac{A_p^l}{B_p^0} t^{2p} \\ &= \sum_{p=n}^{\infty} t^{2p} \sum_{l=n}^p \frac{A_p^l}{B_p^0} \alpha_l^{nk}. \end{aligned}$$

Identifying the coefficients, we then get

$$\sum_{l=n}^p A_p^l \alpha_l^{nk} = \frac{A_p^n B_p^0}{B_p^k}.$$

A mathematical miracle happens here again, as this formula can be inverted using the following fact: for any n and m

$$\sum_{l=0}^n C_n^l A_l^m = \delta_{nm}, \quad (48)$$

where δ is the Kronecker symbol and

$$C_n^l = \frac{2n+1}{2l+1} \binom{n+l}{2l}. \quad (49)$$

We thus obtain

$$\alpha_l^{nk} = \sum_{p=n}^l \frac{C_l^p A_p^n B_p^0}{B_p^k}. \quad (50)$$

Using these values for α_l^{nk} , equation (47) then formally holds. To prove equation (47), one needs to ensure that the series converges. We have $\max_l j_l^2(t) = O(1/l)$; it thus suffices that $\alpha_l^{nk} = O(1/l)$ in order to have equation (47) converge. To show it analytically seems difficult, but numerical tests indicate that it is indeed the case, so equation (47) stands.

C4. Some useful integrals

(i) It is known that for $-2\nu < \alpha < 2$, we have

$$\int_0^\infty t^{\alpha-1} j_\nu^2(t) dt = \frac{\pi^{1/2} \Gamma(1 - \frac{\alpha}{2}) \Gamma(\frac{\alpha}{2} + \nu)}{4 \Gamma(\frac{3}{2} - \frac{\alpha}{2}) \Gamma(-\frac{\alpha}{2} + \nu + 2)}.$$

In particular, for $n \geq 0$, we have

$$\int_0^\infty j_n^2(t) dt = \frac{\pi}{2(2n+1)}, \quad (51)$$

$$\int_0^\infty \frac{j_{n+1}^2(t)}{t} dt = \frac{1}{2(n+1)(n+2)}, \quad (52)$$

$$\int_0^\infty \frac{j_{n+1}^2(t)}{t^2} dt = \frac{\pi}{(2n+1)(2n+3)(2n+5)} \quad (53)$$

$$\int_0^\infty \frac{j_{n+s}^2(t)}{t^{2s-2}} dt = \frac{2^{s-2} \pi (2s-3)!! (2n+1)!!}{(s-1)!(2n+4s-1)!!}. \quad (54)$$

(ii) If we consider equation (32) for $l = 0$ (and therefore n even) and set $y = 2x$, we get

$$4\pi \int_0^\infty y^2 R_{2n,0}(y/2) \frac{\sin(ty)}{ty} dy = (-1)^n 32\pi b_{2n}(2t).$$

Now, using equation (43) (for j_0), we can write for $0 < y < 2$

$$y^2 R_{2n,0}(y/2) = \frac{1}{2\pi^2} \int_0^\infty ty (-1)^n 32\pi b_{2n}(2t) \sin(ty) dt.$$

And finally, we have for $0 < x < 2$

$$\int_0^\infty j_{2n+1}(2t) \sin(tx) dt = \frac{(-1)^n \pi}{8(4n+3)^{1/2}} x R_{2n,0}(x/2). \quad (55)$$

APPENDIX D

Non-uniqueness of the invariant expansion

D1. Using the autocorrelation function

Here we present another equivalent invariant expansion for $I_g(t)$, which could be interesting. If we introduce the autocorrelation function by

$$\varepsilon_g(\mathbf{x}) = \int g(\mathbf{y}) g^*(\mathbf{y} - 2\mathbf{x}) d\mathbf{y}, \quad (56)$$

it is clear that the support of ε_g is included in the unit ball B as soon as $R \geq D_{\max}/2$, and we can write a Zernike expansion for ε_g . Starting from equation (1), we can then write

$$\begin{aligned} I_g(t) &= \int_S d\hat{\mathbf{t}} \left| \int g(\mathbf{x}) \exp(it \cdot \mathbf{x}) d\mathbf{x} \right|^2 \\ &= \int_S d\hat{\mathbf{t}} \int g^*(\mathbf{x}_1) g(\mathbf{x}_2) \exp(it \cdot \mathbf{x}_{12}) d\mathbf{x}_1 d\mathbf{x}_2 \\ &= 8 \int_S d\hat{\mathbf{t}} \int_B \varepsilon_g(\mathbf{x}) \exp(2it \cdot \mathbf{x}) d\mathbf{x} \\ &= 32\pi \sum_{n=0}^\infty b_n(2t) \sum_{l=0}^n (-1)^{(n-l)/2} i^l c_{nl}(\varepsilon_g) \cdot \int_S \mathbf{Y}_l(\hat{\mathbf{t}}) d\hat{\mathbf{t}}, \end{aligned}$$

where we have used equation (34) for ε_g in the last line. The last integral is null except for $l = 0$ (and therefore n even) and we thus obtain

$$I_g(t) = 64\pi^{3/2} \sum_{n=0}^\infty (-1)^n b_{2n}(2t) c_{2n,0,0}(\varepsilon_g).$$

We can finally rewrite this as

$$I_g(t) = \sum_{n=0}^\infty Q_n(g) \frac{j_{2n+1}(2t)}{t}, \quad (57)$$

with

$$Q_n(g) = (-1)^n 32\pi^{3/2} (4n+3)^{1/2} c_{2n,0,0}(\varepsilon_g).$$

As for K_n , we can relate Q_n to L_n by t expanding equation (57) and identifying with equation (7). We thus get

$$L_n(g) = \sum_{l=0}^n \frac{2^{2n+1} (-1)^{n-l}}{(2n+2l+3)!! [2(n-l)]!!} Q_l(g),$$

$$Q_n(g) = (4n+3) \sum_{l=0}^n \frac{(2n+2l+1)!!}{2^{n+l+1} (n-l)!} L_l(g).$$

We can also define the analogue of U_n ,

$$Q_n(g) = 4\pi \int_B g^*(\mathbf{x}_1) g(\mathbf{x}_2) V_n(x_{12}) d\mathbf{x}_1 d\mathbf{x}_2,$$

$$V_n(x) = (4n+3) \sum_{l=0}^n \frac{(-1)^l (2n+2l+1)!!}{2^{n+l+1} (2l+1)! (n-l)!} x^{2l},$$

and write the pair distance distribution function corresponding to this other expansion as

$$p_g(x) = \sum_{n=0}^{\infty} Q_n(g) S_n(x),$$

where, using equations (55) and (29),

$$S_n(x) = \frac{(-1)^n}{16\pi(4n+3)^{1/2}} x^2 R_{2n,0}(x/2).$$

D2. Two infinite sets of invariant expansions for $I_g(t)$

We will now see that both our invariant expansions [equations (12) and (57)] can each be generalized into an infinite set of invariant expansions, valid as soon as $R \geq D_{\max}/2$.

Starting from equation (12) and using equations (40) and (47), we can build another expansion, namely

$$I_g(t) = \sum_{n=0}^{\infty} K_n^0(g) j_n^2(t).$$

Proceeding as before, we can write

$$U_n^0(x) = \sum_{l=0}^n \frac{(-1)^l}{(2l+1)!} C_n^l B_l^0 x^{2l},$$

$$\frac{\sin(tx)}{tx} = \sum_{n=0}^{\infty} U_n^0(x) j_n^2(t).$$

We can now easily prove, by recursion on s , the following generalization of this formula, namely

$$\frac{\sin(tx)}{tx} = \sum_{n=0}^{\infty} U_n^s(x) \frac{j_{n+s}^2(t)}{t^{2s}}, \quad (58)$$

with

$$U_n^s(x) = (-1)^s \frac{1}{x} \frac{d^{2s}}{dx^{2s}} [x U_{n+s}^0(x)]. \quad (59)$$

Moreover, all these identities seem to have the same convergence properties as our first one. Proceeding as before, we can

Figure 11
Functions $[j_{n+s}^2(t)]/t^{2s}$ for $n = 1$ and $s = 0, 1, 2, 3$, normalized to a maximum of 1.

then derive the following, where the case $s = 1$ corresponds to our original K_n , U_n and T_n :

$$I_g(t) = \sum_{n=0}^{\infty} K_n^s(g) \frac{j_{n+s}^2(t)}{t^{2s}}, \quad (60)$$

$$p_g(x) = \sum_{n=0}^{\infty} K_n^s(g) T_n^s(x), \quad (61)$$

$$K_n^s(g) = \sum_{l=0}^n C_{n+s}^{l+s} B_{l+s}^0 L_l(g),$$

$$L_n(g) = \frac{1}{B_{n+s}^0} \sum_{l=0}^n A_{n+s}^{l+s} K_l^s(g),$$

$$K_n^s(g) = 4\pi \int_B g^*(\mathbf{x}_1) g(\mathbf{x}_2) U_n^s(x_{12}) d\mathbf{x}_1 d\mathbf{x}_2,$$

$$U_n^s(x) = \sum_{l=0}^n \frac{(-1)^l}{(2l+1)!} C_{n+s}^{l+s} B_{l+s}^0 x^{2l},$$

$$T_n^s(x) = \frac{x}{2\pi^2} \mu_n^s(x),$$

$$\mu_n^s(x) = \int_0^\infty \frac{j_{n+s}^2(t)}{t^{2s-1}} \sin(tx) dt.$$

Fig. 11 illustrates how the first function spreads with increasing s . At the same time, oscillations dampen at higher t .

Finally, we can also find a simpler expression for the U_n^s and compute T_n^s . This becomes possible when one remarks that

$$U_n^0(x) = (2n+1) P_n \left(1 - \frac{x^2}{2}\right),$$

where P_n is the Legendre polynomial of order n . Indeed, we can rewrite $U_n^0(x)$ as

$$U_n^0(x) = (2n+1) \sum_{l=0}^n (-1)^l \binom{n}{l} \binom{n+l}{l} \left(\frac{x}{2}\right)^{2l}$$

and identify with the following known formula for P_n :

$$P_n(x) = \sum_{l=0}^n \binom{n}{l} \binom{-n-1}{l} \left(\frac{1-x}{2}\right)^l,$$

where we can use the fact that $\binom{-n-1}{l} = (-1)^l \binom{n+1}{l}$. We hence have

$$U_n^s(x) = (-1)^s [2(n+s)+1] \frac{1}{x} \frac{d^{2s}}{dx^{2s}} \left[x P_{n+s} \left(1 - \frac{x^2}{2}\right) \right].$$

We can also compute $\mu_n^s(x)$ by recursion on s , by noting that

$$\mu_n^s(0) = 0,$$

$$\frac{d\mu_n^s}{dx}(0) = \int_0^\infty \frac{j_{n+s}^2(t)}{t^{2s-2}} dt,$$

$$\frac{d^2\mu_n^s}{dx^2}(x) = -\mu_{n+1}^{s-1}(x),$$

the integral being given by equation (54). To compute μ_n^0 , we insert equation (21) into equation (23) to get

$$\int_0^2 U_n^0(x) T_m^0(x) dx = \frac{1}{4\pi} \delta_{nm},$$

which is to be compared to the orthogonality relation for Legendre polynomials,

$$\int_{-1}^1 P_n(x) P_m(x) dx = \frac{2}{2n+1} \delta_{nm}.$$

We can identify them after a simple change of variable and we get for $0 < x < 2$

$$\mu_n^0(x) = \frac{\pi}{4} P_n\left(1 - \frac{x^2}{2}\right).$$

This gives in particular

$$\mu_n^1(x) = \frac{\pi}{2(2n+3)} \left[x - \frac{1}{2} \sum_{l=0}^{n+1} \frac{(-1)^l}{(2l+3)!} C_{n+1}^l B_l^1 x^{2l+2} \right].$$

Starting from our other expansion [equation (57)], we can get in a similar manner

$$\frac{\sin(tx)}{tx} = \sum_{n=0}^{\infty} V_n^s(x) \frac{j_{2n+s}(2t)}{t^s}, \quad (62)$$

from which we get

$$I_g(t) = \sum_{n=0}^{\infty} Q_n^s(g) \frac{j_{2n+s}(2t)}{t^s}. \quad (63)$$

We leave to the interested reader the details of the derivation of these last lines and the computation of the other associated quantities. Finally, it is interesting to note that we have derived equations (60) and (63) for integral s values, but they also make sense for s fractional. It is possible that one could also extend equations (58) and (62) to s fractional.

Acknowledgements

The authors would like to thank Emmanuel Guitter for his help with binomial coefficients, as well as Marc Delarue and Yannick Spill for careful reading of the manuscript and insightful remarks. F. Poitevin acknowledges support from the National Institutes of Health (NIH), grant No. 1-R01-GM115749-01 (PI: Michael Levitt).

References

- Canterakis, N. (1999). In *11th Scandinavian Conference on Image Analysis*. Kangerlussuaq, Greenland: Dansk Selskab for Automatisk Genkendelse af Mønstre.
- Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
- Chen, P.-C. & Hub, J. S. (2014). *Biophys. J.* **107**, 435–447.
- Ciccariello, S., Goodisman, J. & Brumberger, H. (1988). *J. Appl. Cryst.* **21**, 117–128.
- Donatelli, J. J., Zwart, P. H. & Sethian, J. A. (2015). *Proc. Natl Acad. Sci. USA*, **112**, 10286–10291.
- Doniach, S. (2001). *Chem. Rev.* **101**, 1763–1778.
- Elser, V. (2011). *New J. Phys.* **13**, 123014.
- Franke, D., Jeffries, C. M. & Svergun, D. I. (2015). *Nat. Methods*, **12**, 419–422.

- Frieden, B. R. (1971). *Prog. Opt.* **9**, 311–407.
- Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.
- Grishaev, A., Guo, L., Irving, T. & Bax, A. (2010). *J. Am. Chem. Soc.* **132**, 15484–15486.
- Guinier, A. & Fournet, G. (1955). *Small Angle X-ray Scattering*. New York: John Wiley & Sons Inc.
- Kam, Z. (1977). *Macromolecules*, **10**, 927–934.
- Kirian, R. A., Schmidt, K. E., Wang, X., Doak, R. B. & Spence, J. C. H. (2011). *Phys. Rev. E*, **84**, 011921.
- Köfinger, J. & Hummer, G. (2013). *Phys. Rev. E*, **87**, 052712.
- Liu, H., Hexemer, A. & Zwart, P. H. (2012). *J. Appl. Cryst.* **45**, 587–593.
- Liu, H., Morris, R. J., Hexemer, A., Grandison, S. & Zwart, P. H. (2012). *Acta Cryst.* **A68**, 278–285.
- Liu, H., Poon, B. K., Saldin, D. K., Spence, J. C. H. & Zwart, P. H. (2013). *Acta Cryst.* **A69**, 365–373.
- Liu, H. & Spence, J. C. H. (2016). *Quant. Biol.* pp. 1–18.
- Liu, H. & Zwart, P. H. (2012). *J. Struct. Biol.* **180**, 226–234.
- Ma, S. & Liu, H. (2016). *J. Appl. Cryst.* **49**, 665–671.
- Mak, L., Grandison, S. & Morris, R. J. (2008). *J. Mol. Graph. Model.* **26**, 1035–1045.
- Mathar, R. J. (2008). *Baltic Astron.* **17**, 383–398.
- Mendez, D., Lane, T. J., Sung, J., Sellberg, J., Levard, C., Watkins, H., Cohen, A. E., Soltis, M., Sutton, S., Spudich, J., Pande, V., Ratner, D. & Doniach, S. (2014). *Philos. Trans. R. Soc. London B Biol. Sci.* **369**, 20130315.
- Mendez, D., Watkins, H., Qiao, S., Raines, K. S., Lane, T. J., Schenk, G., Nelson, G., Subramanian, G., Tono, K., Joti, Y., Yabashi, M., Ratner, D. & Doniach, S. (2016). *IUCrJ*, **3**, 420–429.
- Merzel, F. & Smith, J. C. (2002). *Acta Cryst.* **D58**, 242–249.
- Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
- Panjikovich, A. & Svergun, D. I. (2016). *Phys. Chem. Chem. Phys.* **18**, 5707–5719.
- Park, S., Bardhan, J. P., Roux, B. & Makowski, L. (2009). *J. Chem. Phys.* **130**, 134114.
- Pedersen, J. S. (1997). *Adv. Colloid Interface Sci.* **70**, 171–210.
- Pedersen, M. C., Hansen, S. L., Markussen, B., Arleth, L. & Mortensen, K. (2014). *J. Appl. Cryst.* **47**, 2000–2010.
- Petoukhov, M. V., Franke, D., Shkumatov, A. V., Tria, G., Kikhney, A. G., Gajda, M., Gorba, C., Mertens, H. D. T., Konarev, P. V. & Svergun, D. I. (2012). *J. Appl. Cryst.* **45**, 342–350.
- Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. (2011). *Nucleic Acids Res.* **39**(suppl. 2), W184–W189.
- Porod, G. (1951). *Kolloid Z.* **124**, 83–114.
- Rambo, R. P. & Tainer, J. A. (2013). *Nature*, **496**, 477–481.
- Różycki, B., Kim, Y. C. & Hummer, G. (2011). *Structure*, **19**, 109–116.
- Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Res.* **38**(suppl. 2), W540–W544.
- Shannon, C. E. (1949). *Proc. IEEE*, **37**, 10–21.
- Slepian, D. & Pollak, H. O. (1961). *Bell Syst. Tech. J.* **40**, 43–63.
- Spill, Y. G., Kim, S. J., Schneidman-Duhovny, D., Russel, D., Webb, B., Sali, A. & Nilges, M. (2014). *J. Synchrotron Rad.* **21**, 203–208.
- Stradner, A., Sedgwick, H., Cardinaux, F., Poon, W. C. K., Egelhaaf, S. U. & Schurtenberger, P. (2004). *Nature*, **432**, 492–495.
- Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
- Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
- Valentini, E., Kikhney, A. G., Previtali, G., Jeffries, C. M. & Svergun, D. I. (2015). *Nucleic Acids Res.* **43**(D1), D357–D363.
- Vestergaard, B. & Hansen, S. (2006). *J. Appl. Cryst.* **39**, 797–804.
- Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.
- Yang, S. (2014). *Adv. Mater.* **26**, 7902–7910.
- Zuo, X., Cui, G., Merz, K. M. Jr, Zhang, L., Lewis, F. D. & Tiede, D. M. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 3534–3539.