# SAXS Merge : an automated statistical method to merge SAXS profiles using Gaussian processes

Yannick G. Spill, Seung Joong Kim, Dina Schneidman-Duhovny, Daniel Russel, Ben Webb, Andrej Sali, Michael Nilges



Journal of
Synchrotron
Radiation

ISSN 1600-5775

Received 28 March 2013
Accepted 3 November 2013

# *SAXS Merge*: an automated statistical method to merge SAXS profiles using Gaussian processes

**Yannick G. Spill,[a,b,c] Seung Joong Kim,[c] Dina Schneidman-Duhovny,[c] Daniel Russel,[c] Ben Webb,[c] Andrej Sali[c] and Michael Nilges[a]***

[a]Unité de Bioinformatique Structurale, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France, [b]Université Paris Diderot, Paris 7, Paris Rive Gauche, 5 rue Thomas Mann, 75013 Paris, France, and [c]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, California Institute for Quantitative Biosciences, Byers Hall, 1700 4th Street, Suite 503 B, University of California San Francisco, San Francisco, CA 94158, USA. *E-mail: nilges@pasteur.fr

Small-angle X-ray scattering (SAXS) is an experimental technique that allows structural information on biomolecules in solution to be gathered. High-quality SAXS profiles have typically been obtained by manual merging of scattering profiles from different concentrations and exposure times. This procedure is very subjective and results vary from user to user. Up to now, no robust automatic procedure has been published to perform this step, preventing the application of SAXS to high-throughput projects. Here, *SAXS Merge*, a fully automated statistical method for merging SAXS profiles using Gaussian processes, is presented. This method requires only the buffer-subtracted SAXS profiles in a specific order. At the heart of its formulation is non-linear interpolation using Gaussian processes, which provides a statement of the problem that accounts for correlation in the data.

**Keywords: SAXS; SANS; data curation; Gaussian process; merging.**

## 1. Introduction

Small-angle X-ray scattering (SAXS) is a popular experiment that allows low-resolution structural information on biomolecules in solution to be gathered (Jacques & Trewhella, 2010; Rambo & Tainer, 2010; Svergun, 2010). The SAXS experiment allows for a wide variety of solution conditions and a wide range of molecular sizes. Data collection usually takes between seconds and minutes in a synchrotron facility, or up to a few hours in an in-house X-ray source (Hura *et al.*, 2009).

The SAXS profile of a biomolecule is the subtraction of the SAXS profile of the biomolecule in solution minus the SAXS profile of the matching buffer. SAXS can be used to study a wide variety of biomolecules, such as proteins, RNA or DNA, and their complexes (Lipfert *et al.*, 2009; Rambo & Tainer, 2010), under a variety of experimental conditions. Once this profile is obtained, it can be used for a variety of modeling tasks (Jacques & Trewhella, 2010; Rambo & Tainer, 2010; Svergun, 2010; Schneidman-Duhovny *et al.*, 2012). It is essential to perform the radial averaging and buffer subtraction steps with high accuracy, as an error at that stage would propagate later on.

The SAXS profile consists of a collection of momentum transfer values (scattering vector) $q$, mean intensities $I(q)$ and standard deviations $s(q)$. Data collection for a given sample is often repeated a number of times $N$ to reduce the noise (or standard deviation) in the SAXS profile by averaging. We consider $N$ as the number of points entering the calculation of $I$ and $s$, because the variation between repetitions is much greater than that due to radial averaging of a single experiment. Additionally, we suppose that the SAXS profiles were collected at several sample concentrations and X-ray exposure times. Both higher concentration and longer X-ray exposure times can provide more information at higher scattering angles. However, both conditions influence the resulting SAXS profile. At higher concentrations, particle–particle interactions can affect the slope of the very low angle part of the SAXS profile (Glatter & Kratky, 1982). At longer exposures, radiation damage can perturb any region of the SAXS profile (Kuwamoto *et al.*, 2004). To remove these unwanted effects it is thus necessary to merge datasets from different experimental conditions. It is the purpose of this method to show that it is possible to do so automatically with minimal user manipulation.

In this article we present the method behind the *SAXS Merge* webserver, a tool presented by Spill *et al.* (2014) which merges SAXS profiles in a robust, completely automated and statistically principled way. While the method was tested on SAXS datasets, it can also be applied for merging small-angle neutron scattering (SANS) datasets, because the basic equations and methods are similar for the two techniques (Svergun,

2010). *SAXS Merge* consists of five steps: data clean-up, profile fitting using Gaussian processes, rescaling of each profile to a common reference, classification and detection of incompatible regions, and merging of the remaining data points. The resulting object is a probability distribution function describing the merged SAXS profile. Resulting data consist of the experimental points that are compatible with the distribution, a maximum posterior estimate of the SAXS profile across all experiments along with a credible interval, and estimates of a number of parameters such as the radius of gyration and the Porod exponent.

## 2. Five steps for SAXS merging

*SAXS Merge* consists of five sequential steps: (i) data clean-up, (ii) profile fitting using Gaussian processes, (iii) rescaling of each profile to a common reference, (iv) classification and detection of incompatible regions, and (v) merging of the remaining data points. The first three steps are performed separately on all input SAXS profiles. We now go through each of these five steps sequentially.

### 2.1. Data clean-up

In this step, we remove from input SAXS profiles data values for which the expected value is not significantly different from zero. Let $\mathcal{H}_0$ be the null hypothesis of a data point being purely noise-induced. Let $\mathcal{H}_1$ be the alternative that it contains some signal. Then with a type-I error of $\alpha$, we can perform a one-sample one-sided $t$-test. Let $I(q_i)$ be a mean intensity at momentum transfer $q_i$, $s(q_i)$ the standard deviation and $N$ the number of repetitions of the experiment. Then the $t$ statistic is

$$t = \frac{I(q_i)}{s(q_i)/N^{1/2}}, \tag{1}$$

and it has a Student $t$ distribution with $\nu = N - 1$ degrees of freedom. Since we are performing a large number of tests, we apply the Bonferroni correction by defining $\alpha \equiv \tilde{\alpha}/M$ ($M$ is the total number of points in the SAXS profile) and choose $\tilde{\alpha} = 0.05$ by default. Normality of the noise is assumed, which is reasonable if no parameter varies across the $N$ replicates of an experiment.

Points with no or zero standard deviation are discarded. Optionally, points with much larger variances than average are discarded as well. This option is proposed because SAXS profiles have almost constant $s(q_i)$ values, except at extreme values for $q_i$ in which case $s(q_i)$ diverges. This behaviour is an experimental artefact, and it is reasonable to remove such points. We therefore calculate the median $\bar{s}$ and discard points which have $s(q_i) > 20\bar{s}$.

### 2.2. Profile fitting using Gaussian processes

We have a number of measurements for a SAXS profile, summarized by three sufficient statistics: intensity $I(q_i)$, standard deviation $s(q_i)$ and number of repetitions $N$ independent of $i$. The SAXS profile is modelled as the noisy measurement of an unknown smooth function $q \mapsto \mathcal{I}(q)$ at $M$ different data points. A pointwise treatment of SAXS profiles fails because of the high noise and correlation encountered in the measurements. This pointwise treatment would lead to an inconsistent classification [step (iv), data not shown]. It is crucial to account for the correlation between successive points to be able to detect outlier data points in a robust manner. Thus, we first estimate the most probable SAXS profile, which was measured with noise in a given SAXS experiment.

This functional estimation is achieved with the help of the theory of Gaussian processes. Gaussian process interpolation (GPI) is a form of non-parametric fitting which has a straightforward probabilistic interpretation and provides confidence intervals on the functional estimate. Given some data and an automatically adjusting smoothness penalty, GPI provides the most probable function that fits the data. For more information on Gaussian processes, see p. 535 of MacKay (2003), §13.43 of O'Hagan & Forster (2004), Rasmussen & Williams (2006) and http://gaussianprocess.org.

**2.2.1. Likelihood**. Define

$$\mathbf{I} = \begin{pmatrix} I(q_1) \\ \vdots \\ I(q_M) \end{pmatrix}, \quad \mathbf{S} = \mathrm{diag}\begin{pmatrix} s^2(q_1) \\ \vdots \\ s^2(q_M) \end{pmatrix}, \quad \mathcal{I} = \begin{pmatrix} \mathcal{I}(q_1) \\ \vdots \\ \mathcal{I}(q_M) \end{pmatrix}. \tag{2}$$

$\mathbf{S}$ is the sample covariance matrix, assumed to be diagonal given $\mathcal{I}$. We treat $\mathbf{I}$ as a measurement with noise of the function $\mathcal{I}$ at positions $\{q_i\}_{i=1,\ldots,M}$ so that $\mathbf{I} = \mathcal{I} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is a vector distributed as a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$. We make the assumption that $\boldsymbol{\Sigma} \equiv \sigma^2\mathbf{S}$, where $\sigma$ is a proportionality constant that will be estimated in the process. The assumption of a diagonal $\mathbf{S}$ matrix is not entirely correct, as shown by Breidt *et al.* (2012). However, correlations are expected to be non-zero only between neighbouring annuli (*i.e.* $q$ values), and the covariance model we introduce next spans much further than that. This assumption leads to the following likelihood,

$$p(\mathbf{I}|\mathcal{I}, \mathbf{S}, N) \equiv \frac{1}{(2\pi)^{M/2}\left|\sigma^2\mathbf{S}/N\right|^{1/2}} \times \exp\Big[-(1/2)(\mathbf{I}-\mathcal{I})^{\top}\left(\sigma^2\mathbf{S}/N\right)^{-1}(\mathbf{I}-\mathcal{I})\Big]. \tag{3}$$

**2.2.2. Prior**. The likelihood alone does not constrain the vector $\mathcal{I}$, which is still free to vary. However, we believe that the function $\mathcal{I}$ is smooth. This belief is modelled by assuming that the vector $\mathcal{I}$ follows a multivariate normal distribution with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{W}$ which have been carefully chosen (see below),

$$p(\mathcal{I}|\mathbf{m}, \mathbf{W}) \equiv \frac{1}{(2\pi)^{M/2}|\mathbf{W}|^{1/2}} \times \exp\big[-(1/2)(\mathcal{I}-\mathbf{m})^{\top}\mathbf{W}^{-1}(\mathcal{I}-\mathbf{m})\big]. \tag{4}$$

Equivalently, one can say that the function $\mathcal{I}$ has a Gaussian process prior distribution with prior covariance function $w$ and prior mean function $m$.

**2.2.3. Choice of *w***. The covariance function determines the smoothness of the Gaussian process. We choose the commonly used squared exponential form, which yields continuous and infinitely differentiable functions. Therefore, this approach is in principle only usable on smooth SAXS profiles,

$$w(q, q') \equiv \tau^2 \exp\left[-\frac{(q-q')^2}{2\lambda^2}\right]. \tag{5}$$

The covariance function has two parameters: $\tau^2$ is the variance that the Gaussian process assigns in regions far from any data point; $\lambda$ is the persistence length of the profile, in units of $q$. With this covariance function, we define

$$\mathbf{w}(q) \equiv \begin{pmatrix} w(q_1, q) \\ \vdots \\ w(q_M, q) \end{pmatrix},$$
$$\mathbf{W} \equiv \begin{pmatrix} w(q_1, q_1) & \cdots & w(q_1, q_M) \\ \vdots & \ddots & \vdots \\ w(q_M, q_1) & \cdots & w(q_M, q_M) \end{pmatrix}. \tag{6}$$

**2.2.4. Choice of *m***. Gaussian process interpolation is a non-parametric approach. However, it is possible to incorporate some prior knowledge in the form of a parametric mean function, making the approach semi-parametric. In our case, this way of proceeding has the advantage of providing an estimate of the radius of gyration and other constants. At the same time, the Gaussian process fits the signal in the data that is unexplained by the parametric mean function so that even high deviations from the prior mean function will be followed by the Gaussian process.

At very low angle, the Guinier plot allows for an estimation of the radius of gyration,

$$I(q) \propto \exp\left(-\frac{R_G^2}{3} q^2\right). \tag{7}$$

For the higher-angle portion of the profile, Porod's law is

$$I(q) \propto q^{-4}. \tag{8}$$

Hammouda (2010) constructed a smooth function encompassing both behaviours, which we use as a starting point for $m$,

$$m(q) \equiv A + \begin{cases} (G/q^s)\exp\left[-q^2 R_G^2/(3-s)\right] & \text{if } q \le q_1, \\ D/q^d & \text{if } q > q_1, \end{cases} \tag{9}$$

$$q_1 \equiv (1/R_G)[(d-s)(3-s)/2]^{1/2}, \tag{10}$$

$$D \equiv G q_1^{d-s} \exp\left[-q_1^2 R_G^2/(3-s)\right]. \tag{11}$$

This function has five parameters: $A$, $G$, $R_G$, $d$ and $s$. Some of them can be fixed to certain values, generating a nested family of parametric functions. For example, setting $G = 0$ reduces $m$ to a constant function. Setting $d$ such that $q_1$ is larger than any input $q$-value reduces $m$ to the Guinier model with a constant offset. Finally, setting $s = 0$ reduces $m$ to the simpler Guinier–Porod model described in the first section of Hammouda (2010) (up to a constant offset). Define

$$\mathbf{m} \equiv \left[\, m(q_1) \cdots m(q_M) \,\right]^\top. \tag{12}$$

**2.2.5. Hyperprior**. The parameters arising in the prior mean or covariance functions as well as $\sigma$ are collectively called *hyperparameters*. In this hierarchical approach we can in turn assign a prior to these hyperparameters. Since our knowledge of their plausible values is rather vague, we give a Jeffreys prior to $\sigma^2$ and a uniform prior to the other parameters. However, for the sake of model comparison, parameters are bounded within a finite interval to allow for a normalized prior,

$$p(\sigma^2) = \frac{1}{\log(\sigma^{\max}/\sigma^{\min})}\,\frac{1}{\sigma^2},$$
$$p(P_i) = \frac{1}{P_i^{\max} - P_i^{\min}} \qquad P_i \in \{G, R_G, d, s, A, \tau, \lambda\}. \tag{13}$$

**2.2.6. Fitting the SAXS profile**. In order to find the best fit of the SAXS profile, it is required to optimize the hyperparameters. Defining $\mathbf{\Theta} \equiv (G, R_G, d, s, A, \tau, \lambda, \sigma)^\top$ and $D \equiv (\mathbf{I}, \mathbf{S}, N)$, this optimization can be achieved by maximizing $p(\mathbf{\Theta}|D)$ with respect to $\mathbf{\Theta}$. With the help of Bayes' rule, we obtain

$$p(\mathbf{\Theta}|D) \propto p(\mathbf{I}|\mathbf{S}, N, \mathbf{\Theta})\, p(\mathbf{\Theta}), \tag{14}$$

where $p(\mathbf{\Theta})$ is given in equation (13) and $p(\mathbf{I}|\mathbf{\Theta}, \mathbf{S}, N)$ is called the marginal likelihood,

$$p(\mathbf{I}|\mathbf{S}, N, \mathbf{\Theta}) \equiv \int p(\mathbf{I}|\mathcal{I}, \mathbf{S}, N) p(\mathcal{I}|\mathbf{\Theta})\, \mathrm{d}\mathcal{I}. \tag{15}$$

Since both the likelihood [equation (3)] and the prior [equation (4)] appearing in this integral are multivariate Gaussian distributions, it is possible to give an analytical expression of the marginal likelihood,

$$p(\mathbf{I}|\mathbf{S}, N, \mathbf{\Theta}) = \frac{1}{(2\pi)^{M/2}|\mathbf{\Omega}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}^\top \mathbf{\Omega}^{-1} \boldsymbol{\varepsilon}\right), \tag{16}$$

with $\boldsymbol{\varepsilon} \equiv \mathbf{I} - \mathbf{m}$ and $\mathbf{\Omega} \equiv \sigma^2 \mathbf{S}/N + \mathbf{W}$.

**2.2.7. Obtaining functional deviates**. To make predictions of $\mathcal{I}$ at a new point $q$ we average over all possible values for $\mathbf{\Theta}$, weighted by their posterior probability,

$$p[\mathcal{I}(q)|D] = \int p[\mathcal{I}(q)|\mathbf{\Theta}, D]\, p(\mathbf{\Theta}|D)\, \mathrm{d}\mathbf{\Theta}. \tag{17}$$

Let us examine the two terms appearing in this last integral. The posterior probability density of the hyperparameters $p(\mathbf{\Theta}|D)$ was already encountered in equation (14).

The remaining term, $p(\mathcal{I}(q)|\mathbf{\Theta}, D)$, is the posterior predictive probability density of a new noise-free observation given the hyperparameters. It is called posterior predictive because it allows new values of the SAXS profile given the noisy observations to be predicted. Since the function $\mathcal{I}$ has a Gaussian process prior and a multivariate normal likelihood,

the posterior distribution for $\mathcal{I}$ is also a Gaussian process, with mean function $\hat{\mathcal{I}}$ and covariance function $\hat{\sigma}^2_{\mathcal{I}}$ given by

$$\forall q \qquad \hat{\mathcal{I}}(q) = m(q) + \mathbf{w}(q)^{\top}\mathbf{\Omega}^{-1}(\mathbf{I} - \mathbf{m}), \tag{18}$$

$$\forall q, q' \qquad \hat{\sigma}^2_{\mathcal{I}}(q, q') = w(q, q') - \mathbf{w}(q)^{\top}\mathbf{\Omega}^{-1}\mathbf{w}(q'). \tag{19}$$

These equations arise from the fact that the vector $[\mathcal{I}(q), I(q_1), \ldots, I(q_M)]^{\top}$ has a multivariate normal distribution with mean vector $[m(q), m(q_1), \ldots, m(q_M)]^{\top}$ and covariance matrix

$$\begin{pmatrix} w(q, q) & \mathbf{w}(q)^{\top} \\ \mathbf{w}(q) & \mathbf{\Omega} \end{pmatrix}. \tag{20}$$

The distribution for $\mathcal{I}(q)$ then results from the conditioning of the multivariate normal distribution on the observed values,

$$p[\mathcal{I}(q)|\mathbf{\Theta}, D] = \frac{1}{(2\pi)^{1/2}\,\hat{\sigma}_{\mathcal{I}}(q, q)} \exp\left\{-\frac{\left[\mathcal{I}(q) - \hat{\mathcal{I}}(q)\right]^2}{\hat{\sigma}^2_{\mathcal{I}}(q, q)}\right\}. \tag{21}$$

Note that it is also possible to generate random functional deviates from the posterior predictive distribution. If $k$ points are wanted for each functional estimate, one can draw them from the multivariate normal distribution with mean vector and covariance matrix built, respectively, from the posterior mean function $\hat{\mathcal{I}}$ and the posterior covariance function $\hat{\sigma}^2_{\mathcal{I}}$ at the values $q_1, \ldots, q_k$.

Although we could in principle perform the interpolation by numerically integrating equation (17) for every value of $q$ needed, this approach would be costly in terms of computation power. In fact, two integrals would need to be computed numerically, equation (17) and also the normalization constant of equation (14),

$$p(\mathbf{I}|\mathbf{S}, N) = \int p(\mathbf{I}|\mathbf{\Theta}, \mathbf{S}, N)\, p(\mathbf{\Theta})\, \mathrm{d}\mathbf{\Theta}. \tag{22}$$

Luckily, as Gibbs & MacKay (1997) have pointed out, a Laplace approximation of this last integral is a very good approximation because hyperparameters are usually quite peaked around their most probable value. This approach is known as a type-II maximum likelihood (ML-II),

$$p(\mathbf{I}|\mathbf{S}, N) \simeq p(\mathbf{I}|\hat{\mathbf{\Theta}}, \mathbf{S}, N)\, p(\hat{\mathbf{\Theta}})\Delta\hat{\mathbf{\Theta}}, \tag{23}$$

$$\Delta\hat{\mathbf{\Theta}} = (2\pi)^{N_p}\left|\frac{\partial^2 E}{\partial\mathbf{\Theta}^2}(\mathbf{I}, \hat{\mathbf{\Theta}})\right|^{-1/2}, \tag{24}$$

$$E(\mathbf{I}, \mathbf{\Theta}) = -\log p(\mathbf{I}|\mathbf{\Theta}, \mathbf{S}, N) - \log p(\mathbf{\Theta}). \tag{25}$$

$N_p \equiv \dim(\mathbf{\Theta})$ is the number of parameters. $\Delta\hat{\mathbf{\Theta}}$ is the phase space volume in which values of $\mathbf{\Theta}$ are acceptable given $D$, and is usually small (Rasmussen & Williams, 2006). This procedure has a considerable practical advantage, since optimization of the hyperparameters then does not need to be performed for each new $\mathcal{I}(q)$ but only once for this dataset. The optimization itself has been described in §2.2.6.

Once the most probable $\mathbf{\Theta}$ has been found, the Laplace approximation gives the normalization constant of $p(\mathbf{\Theta}|D)$,

$$p(\mathbf{\Theta}|D) \simeq \frac{p(\mathbf{I}|\mathbf{\Theta}, \mathbf{S}, N)\, p(\mathbf{\Theta})}{p(\mathbf{I}|\hat{\mathbf{\Theta}}, \mathbf{S}, N)\, p(\hat{\mathbf{\Theta}})\Delta\hat{\mathbf{\Theta}}} \tag{26}$$

With the additional hypothesis that $p[\mathcal{I}(q), \mathbf{\Theta}, D]\, p(\mathbf{\Theta}|D)$ has the same maximum for $\mathbf{\Theta}$ as $p(\mathbf{\Theta}|D)$ alone, equation (17) becomes

$$p[\mathcal{I}(q)|D] \simeq p[\mathcal{I}(q)|\hat{\mathbf{\Theta}}, D]\left|I_n + AB^{-1}\right|^{-1/2} \tag{27}$$

$$A = \left\{-\frac{\partial^2 \log p[\mathcal{I}(q)|\mathbf{\Theta}, D]}{\partial\mathbf{\Theta}^2}\right\}_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}} \tag{28}$$

$$B = \frac{\partial^2 E(\mathbf{I}, \hat{\mathbf{\Theta}})}{\partial\mathbf{\Theta}^2}. \tag{29}$$

It is also possible to compute the posterior mean and covariance functions averaged over all values of $\mathbf{\Theta}$,

$$\mathring{\mathcal{I}}(q) \simeq \hat{\mathcal{I}}(q)\left|I_n + A'B^{-1}\right|^{-1/2} \tag{30}$$

$$\mathring{\sigma}^2_{\mathcal{I}}(q, q) \simeq \hat{\sigma}^2_{\mathcal{I}}(q, q)\left|I_n + A''B^{-1}\right|^{-1/2} \tag{31}$$

$$A' = \left[-\frac{\partial^2}{\partial\mathbf{\Theta}^2}\log\hat{\mathcal{I}}(q)\right]_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}} \tag{32}$$

$$A'' = \left[-\frac{\partial^2}{\partial\mathbf{\Theta}^2}\log\hat{\sigma}^2_{\mathcal{I}}(q, q)\right]_{\mathbf{\Theta}=\hat{\mathbf{\Theta}}} \tag{33}$$

**2.2.8. Choice between different mean functions *via* model comparison**. Sometimes, the information content of a SAXS profile is so low that the number of parameters in the mean function exceed the number of identifiable parameters of the SAXS profile. In that case, overfitting occurs, and it is preferrable to try a simpler mean function.

The previously presented mean function has five parameters. It has been noted that it generates a nested family of parametric functions when some parameters are held fixed. For globular proteins, $s$ can be set to zero, reducing the number of parameters to four. It is also possible to use simpler functions. For example,

$$m_G(q) = A + G\exp\left(-\frac{q^2R_{\mathrm{G}}^2}{3}\right) \tag{34}$$

assumes the SAXS profile only contains the Guinier region; it has three parameters. The flat function has one parameter: $m_F(q) = A$.

Fitting can be performed using a number of different mean functions. The one that is the most plausible is then selected by model comparison. Suppose $M_1$ ($M_2$) represents the model in which the mean and covariance functions total $N_p^1$ ($N_p^2$) parameters, summarized in the parameter vector $\mathbf{\Theta}_1$ ($\mathbf{\Theta}_2$). The best mean function is the one which has the highest Bayes factor,

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)}\frac{p(D|M_1)}{p(D|M_2)} \tag{35}$$

$$= 1 \cdot \frac{\int p(D|\mathbf{\Theta}_1, M_1)\, p(\mathbf{\Theta}_1|M_1)\, \mathrm{d}\mathbf{\Theta}_1}{\int p(D|\mathbf{\Theta}_2, M_2)\, p(\mathbf{\Theta}_2|M_2)\, \mathrm{d}\mathbf{\Theta}_2}. \tag{36}$$

The Bayes factor is the ratio of the evidences if both models have an equal *a priori* probability. As was just discussed, we simplify this assumption further by performing a Laplace approximation of the integral around the maximum posterior set of parameters. This expansion yields

$$\frac{p(M_1|D)}{p(M_2|D)} \simeq \frac{p(\mathbf{I}|\hat{\mathbf{\Theta}}_1, \mathbf{S}, N, M_1)\, p(\hat{\mathbf{\Theta}}_1|M_1)}{p(\mathbf{I}|\hat{\mathbf{\Theta}}_2, \mathbf{S}, N, M_2)\, p(\hat{\mathbf{\Theta}}_1|M_2)} \left[(2\pi)^{1/2}\right]^{N_p^1 - N_p^2} \times \left|\frac{\partial^2 E_1}{\partial \mathbf{\Theta}_1^2}(\mathbf{I}, \hat{\mathbf{\Theta}}_1)\right|^{-1/2} \Big/ \left|\frac{\partial^2 E_2}{\partial \mathbf{\Theta}_2^2}(\mathbf{I}, \hat{\mathbf{\Theta}}_2)\right|^{-1/2}. \tag{37}$$

Details of the calculation, along with gradient and hessian of Hammouda's Generalized Guinier Porod function (Hammouda, 2010), are given in the supporting information.[1]

### 2.3. Rescaling

Suppose $\mathcal{I}_0$ is given, and we want to find the scaling factor $\gamma$ between $\mathcal{I}_0$ and $\mathcal{I}_1$, such that the distance between $\mathcal{I}_0$ and $\gamma\mathcal{I}_1$ is minimized under some metric. We propose three similar models to rescale the SAXS profiles: normal model, normal model with constant offset, and lognormal model (using the logs of the intensities rather than the intensities themselves). In this section we assume $\mathcal{I}_i$ are evaluated at $M$ points, and treat $\mathcal{I}_i$ as a vector with $M$ entries.

In the normal model we use the squared error loss

$$\mathcal{L} \equiv (\mathcal{I}_0 - \gamma\mathcal{I}_1)^\top \mathbf{A} (\mathcal{I}_0 - \gamma\mathcal{I}_1), \tag{38}$$

where $\mathbf{A}$ is a symmetric positive definite matrix. The risk is

$$\mathcal{R} \equiv \mathbb{E}_{\mathcal{I}_1}\left\{\mathbb{E}_{\mathcal{I}_0}\left[(\mathcal{I}_0 - \gamma\mathcal{I}_1)^\top \mathbf{A} (\mathcal{I}_0 - \gamma\mathcal{I}_1)\right]\right\}. \tag{39}$$

It can be put in the form

$$\mathcal{R} = \left(\hat{\mathcal{I}}_0 - \gamma\hat{\mathcal{I}}_1\right)^\top \mathbf{A}\left(\hat{\mathcal{I}}_0 - \gamma\hat{\mathcal{I}}_1\right) + \mathrm{tr}\left[\mathbf{A}\left(\mathbf{\Sigma}_0 + \gamma^2\mathbf{\Sigma}_1\right)\right], \tag{40}$$

where $\mathbf{\Sigma}_i$ is the covariance matrix of $\mathcal{I}_i$. We would like to choose $\gamma$ and $\mathbf{A}$ so that the risk is minimal,

$$\frac{\partial\mathcal{R}}{\partial\gamma} = -2\left(\hat{\mathcal{I}}_0 - \gamma\hat{\mathcal{I}}_1\right)^\top \mathbf{A}\,\hat{\mathcal{I}}_1 + 2\gamma\,\mathrm{tr}(\mathbf{A}\mathbf{\Sigma}_1), \tag{41}$$

$$\frac{\partial\mathcal{R}}{\partial\mathbf{A}} = \left(\hat{\mathcal{I}}_0 - \gamma\hat{\mathcal{I}}_1\right)\left(\hat{\mathcal{I}}_0 - \gamma\hat{\mathcal{I}}_1\right)^\top + \mathbf{\Sigma}_0 + \gamma^2\mathbf{\Sigma}_1. \tag{42}$$

The second equation is a sum of positive matrices, and cannot be zero. Therefore there is no choice of $\mathbf{A}$ that minimizes the risk. We choose $\mathbf{A} \equiv \mathbf{\Sigma}_1^{-1}$. Minimizing the first equation gives the target value for $\gamma$,

$$\hat{\gamma} \equiv \frac{\hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1} \hat{\mathcal{I}}_0}{\hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1} \hat{\mathcal{I}}_1 + M}. \tag{43}$$

[1] Supporting information for this paper is available from the IUCr electronic archives (Reference: CO5036).

The mean vectors are computed from equation (18) or (30); the covariance matrices from equation (19) or (31).

The normal model with offset has loss function

$$\mathcal{L} \equiv \left[\mathcal{I}_0 - \gamma(\mathcal{I}_1 + c\mathbf{J})\right]^\top \mathbf{\Sigma}_1^{-1} \left[\mathcal{I}_0 - \gamma(\mathcal{I}_1 + c\mathbf{J})\right], \tag{44}$$

where $\mathbf{J}$ is a vector of ones. This model leads to the estimates

$$\hat{c} \equiv \frac{M\hat{\mathcal{I}}_0^\top \mathbf{\Sigma}_1^{-1}\mathbf{J} + \hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1}\left(\hat{\mathcal{I}}_1\hat{\mathcal{I}}_0^\top - \hat{\mathcal{I}}_0\hat{\mathcal{I}}_1^\top\right)\mathbf{\Sigma}_1^{-1}\mathbf{J}}{\hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1}\left(\hat{\mathcal{I}}_0\mathbf{J}^\top - \mathbf{J}\hat{\mathcal{I}}_0^\top\right)\mathbf{\Sigma}_1^{-1}\mathbf{J}}, \tag{45}$$

$$\hat{\gamma} \equiv \frac{\hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1} \hat{\mathcal{I}}_0}{\hat{\mathcal{I}}_1^\top \mathbf{\Sigma}_1^{-1}(\hat{\mathcal{I}}_1 + \hat{c}\,\mathbf{J}) + M}. \tag{46}$$

Finally the lognormal model has loss function

$$\mathcal{L} \equiv \left[\mathbf{J}\log\gamma - \log\left(\frac{\mathcal{I}_0}{\mathcal{I}_1}\right)\right]^\top \mathbf{\Sigma}_1^{-1} \left[\mathbf{J}\log\gamma - \log\left(\frac{\mathcal{I}_0}{\mathcal{I}_1}\right)\right], \tag{47}$$

which is defined because the intensities are expected to be positive. The estimate for $\gamma$ is then

$$\log\hat{\gamma} \equiv \frac{\left[\log\left(\hat{\mathcal{I}}_0/\hat{\mathcal{I}}_1\right)\right]^\top \mathbf{\Sigma}_1^{-1}\mathbf{J}}{\mathbf{J}^\top \mathbf{\Sigma}_1^{-1}\mathbf{J}}. \tag{48}$$

By default, all profiles are rescaled to the last profile, which has usually the widest range.

### 2.4. Classification

To classify the SAXS profiles, it is necessary to rank them. SAXS profiles are ranked as follows. For each profile $i$, we compute $I_i(0)$ by fitting the Guinier region, and the median $\bar{s}_i$ of the errors. We use the median instead of the mean because it is more robust to outliers. The profiles are then ranked by ascending $I_i(0)/\bar{s}_i$. This quantity is expected to increase with either concentration or X-ray dose.

The first profile has the reference status on all intervals that have not been discarded by the first step (*i.e.* as long as its signal-to-noise ratio is sufficiently high). Let $\mathcal{I}$ be the candidate profile, and $\mathcal{I}_{\rm ref}$ the reference profile, for which we have just derived a distribution in the fitting step. Because correlation has been accounted for in the profile fitting step (§2.2.6), pointwise statistical treatment is sufficient. The SAXS profiles are then compared by using a two-sample two-sided $t$ test and regions of disagreement are determined.

We would like to know which measurements of $\mathcal{I}(q)$ and $\mathcal{I}_{\rm ref}(q)$ are compatible. We simply assume that each new observation at scattering angle $q$ is drawn from a normal distribution with mean $\mu(q)$ and standard deviation $\sigma_{\rm exp}(q)$, where

$$\mu(q) = \hat{\gamma}\mathring{\mathcal{I}}(q), \tag{49}$$

$$\sigma_{\rm exp}^2(q) = \hat{\gamma}^2\mathring{\sigma}_{\mathcal{I}}^2(q). \tag{50}$$

$\mathring{\mathcal{I}}(q)$ and $\mathring{\sigma}_{\mathcal{I}}(q)$ are given by equations (30) and (31) and $\hat{\gamma}$ by equations (48), (46) or (43). If no parameter averaging was

performed, one can use $\mathcal{I}$ and $\sigma_{\mathcal{I}}$ instead of $\mathring{\mathcal{I}}$ and $\mathring{\sigma}_{\mathcal{I}}$ given by equations (18) and (19), respectively.

We then perform Welch's two-sample two-sided $t$-test at confidence level $\alpha$ (Welch, 1947). Similar to §2.1, we compute the $t$ statistic

$$t = \frac{|\mu(q) - \mu_{\rm ref}(q)|}{\left[\sigma^2_{\rm exp}(q)/N + \sigma^2_{\rm exp,ref}(q)/N_{\rm ref}\right]^{1/2}} \qquad (51)$$

with $N$ and $N_{\rm ref}$ the number of repetitions of each experiment. The degrees of freedom are given by the Satterthwaite approximation,

$$\nu = \frac{\left[\sigma^2_{\rm exp}(q)/N + \sigma^2_{\rm exp,ref}(q)/N_{\rm ref}\right]^2}{\left[\sigma^2_{\rm exp}(q)/N\right]^2/(N-1) + \left[\sigma^2_{\rm exp,ref}(q_i)/N_{\rm ref}\right]^2/(N_{\rm ref}-1)}. \qquad (52)$$

If the $p$-value of this test is smaller than $\alpha$ then the functions are locally different and $\mathcal{I}(q_i)$ is discarded.

Usually, the second profile spans a wider $q$ range, so that comparison with the reference profile cannot be carried out at higher angles. In such a case the remaining portion of the second profile is marked as valid, and becomes the reference. Next, the third profile is compared with the low-angle part of the first profile and with the high-angle part of the second profile. If the third profile spans a wider $q$ range than the second profile, its tail becomes the reference for the remaining $q$ values, and so on until all SAXS profiles have been compared.

### 2.5. Merging

The merging step simply consists of pooling all compatible data points, keeping track of their origins. Gaussian process interpolation is then performed on this merged dataset. It can then happen that some datasets overlap, leading to multiple intensities for the same $q$ values. In that case we discard the points which have the largest standard deviations. This behaviour can be disabled.

## 3. Conclusion

In this article we have developed *SAXS Merge*, a fully automated method for merging SAXS profiles in a robust and statistically principled way. It has been released as both a software package and a webserver, as described by Spill *et al.* (2014). The required input consists only of the buffer-subtracted profile files in a three-column format ($q$, intensity, standard deviation).

YGS thanks Riccardo Pellarin for discussion about Bayesian scoring. MN acknowledges funding from the European Union (FP7-IDEAS-ERC 294809).

## References

Breidt, F. J., Erciulescu, A. & van der Woerd, M. (2012). *J. Time Ser. Anal.* **33**, 704–717.
Gibbs, M. & MacKay, D. J. (1997). *Efficient Implementation of Gaussian Processes.* Technical Report. Cavendish Laboratory, Department of Physics, Cambridge University, UK.
Glatter, O. & Kratky, O. (1982). *Small Angle X-ray Scattering.* New York: Academic Press.
Hammouda, B. (2010). *J. Appl. Cryst.* **43**, 716–719.
Hura, G., Menon, A., Hammel, M., Rambo, R., Poole II, F., Tsutakawa, S., Jenney Jr, F., Classen, S., Frankel, K., Hopkins, R., Yang, S., Scott, J., Dillard, B., Adams, M. & Tainer, J. (2009). *Nat. Methods*, **6**, 606–612.
Jacques, D. A. & Trewhella, J. (2010). *Protein Sci.* **19**, 642–657.
Kuwamoto, S., Akiyama, S. & Fujisawa, T. (2004). *J. Synchrotron Rad.* **11**, 462–468.
Lipfert, J., Herschlag, D. & Doniach, S. (2009). *Riboswitches*, edited by A. Serganov, *Methods in Molecular Biology*, Vol. 540, pp. 141–159. Totowa: Humana Press.
MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*, 1st ed. Cambridge University Press.
O'Hagan, A. & Forster, J. (2004). *Bayesian Inference.* London: Arnold.
Rambo, R. P. & Tainer, J. A. (2010). *Curr. Opin. Struct. Biol.* **20**, 128–137.
Rasmussen, C. & Williams, C. (2006). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* Cambridge: MIT Press.
Schneidman-Duhovny, D., Kim, S. & Sali, A. (2012). *BMC Struct. Biol.* **12**, 17.
Spill, Y. G., Kim, S. J., Schneidman-Duhovny, D., Russel, D., Webb, B., Sali, A. & Nilges, M. (2014). Submitted.
Svergun, D. I. (2010). *Biol. Chem.* **391**, 737–743.
Welch, B. L. (1947). *Biometrika*, **34**, 28–35.