Correlation Map, a goodness-of-fit test for one-dimensional X-ray scattering spectra

Daniel Franke, Cy M Jeffries & Dmitri I Svergun

Assessing similarity between data sets with the reduced χ^2 test requires the estimation of experimental errors, which, if incorrect, may render statistical comparisons invalid. We report a goodness-of-fit test, Correlation Map (CorMap), for assessing differences between one-dimensional spectra independently of explicit error estimates, using only data point correlations. Using small-angle X-ray scattering data, we demonstrate that CorMap maintains the power of the reduced χ^2 test; moreover, CorMap is also applicable to other physical experiments.

The analysis of experimental data often necessitates statistical comparisons of similarity between observations that incorporate either systematic or purely random measurement errors. Systematic errors derived from the instrument, sample or data processing steps require experiment-dependent treatment, whereas random errors are taken into account using standard procedures[1]. However, it can be difficult at a practical level to obtain accurate error estimates required for valid statistical testing. Therefore, we developed an approach to quantitatively assess data-data comparisons and to evaluate model-data correspondence without the need to explicitly estimate experimental errors. We apply a correlation-based statistical test to evaluate small-angle X-ray scattering (SAXS) data obtained from biological macromolecules in solution. This method, CorMap, should also be generally applicable for experiments producing oversampled one-dimensional data sets of discrete data points.

For SAXS, scattering intensities $I(q)$ are recorded as a function of angle, or momentum transfer $q = 4\pi \sin\theta/\lambda$, where λ is the X-ray wavelength and 2θ is the scattering angle (ref. 2). The successful interpretation of one-dimensional SAXS data is marred by a number of pitfalls that need addressing before reporting conclusions on the basis of these data[3]. In particular, when the statistical similarity between experimentally obtained intensities, $I_{exp}(q)$,

and those computed from a model, $I_{calc}(q)$, is evaluated using the reduced χ^2 statistic[4]

$$\chi^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left[\frac{I_{exp}(q_k) - I_{calc}(q_k)}{\sigma(I_{exp}(q_k))} \right]^2$$

at n experimental data points, it is necessary that the experimental errors, $\sigma(I_{exp}(q_k))$, are correctly estimated in order for the test to be statistically valid. If this condition is met and the model adequately describes the experimental data, the resulting χ^2 should be in the range $0.9 \leq \chi^2 \leq 1.1$ (**Supplementary Fig. 1**). However, the true values of $\sigma(I_{exp}(q_k))$, i.e., $\sigma(I(q_k))$, are always unknown and have to be estimated from the data assuming Poisson statistics (**Supplementary Fig. 2** and Online Methods). Accurate propagation of the recorded errors through data processing steps is also nontrivial[2]. Consequently, and especially at the modern high-throughput SAXS facilities, there is a risk of collecting thousands of data sets with poorly determined or incorrectly propagated errors that may invalidate the assessments of data-data or data-model fits. This problem is also evident in fields other than SAXS[5].

We propose a statistically valid approach that simply utilizes experimental intensities. When a photon-counting detector is used, each radially averaged $I_{exp}(q_k)$ may be considered as a sample drawn from a normally distributed random variable with expected value $I(q_k)$ and s.d. $\sigma(I(q_k))$, i.e., $I_{exp}(q_k) \sim N(I(q_k), \sigma(I(q_k)))$ (**Supplementary Fig. 2** and Online Methods). Therefore, an entire scattering profile, S, that is a collection of n normally distributed $I_{exp}(q_k)$ data points, may be conceptualized to be simultaneously drawn from an n-variate normal distribution, $S \sim N(J, \Sigma)$, where J corresponds to the vector of expected intensities of S, and Σ to its variance-covariance matrix

$$J = \begin{pmatrix} \vdots \\ I(q_k) \\ \vdots \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \ddots & & & \\ & \sigma(I(q_k))^2 & \cdots & \sigma(I(q_k),I(q_l)) \\ & \vdots & \ddots & \vdots \\ & \sigma(I(q_l),I(q_k)) & \cdots & \sigma(I(q_l))^2 \\ & & & & \ddots \end{pmatrix}$$

In synchrotron SAXS, data are usually recorded in multiple short frames to monitor for various systematic deviations, for example, for radiation damage. Inserting the observed experimental intensities $I_{exp}(q_k)$ for m frames along the diagonal of Σ,

$$\sigma(I_{exp}(q_k))^2 = \frac{1}{m-1} \sum_{i=1}^{m} (I_{exp}(q_k)_i - \bar{I}_{exp}(q_k))^2$$

Figure 1 | CorMap visualization. (**a**,**d**,**g**,**j**) One-dimensional SAXS data sets, each consisting of 20 data frames, obtained from aqueous buffer (**a**), glucose isomerase (solvent-subtracted data) (**d**), lysozyme with accumulating radiation damage (unsubtracted data) (**g**) and human serum albumin with scaling issues (**j**). (**b**,**e**,**h**,**k**) CorMaps corresponding to the data sets at left, computed from all 20 frames. (**c**,**f**,**i**,**l**) Selected pairwise comparisons of two frames each. The corresponding probabilities of similarity (*P* value) of the two-frame comparisons are 0.3400 (**c**), 0.097 (**f**), 0.0002 (**i**) and <10^{-5} (**l**).

and the off-diagonal covariance estimates between all point-to-point q_k and q_l

$$\sigma(I_{exp}(q_k), I_{exp}(q_l)) =$$

$$\frac{1}{m-1} \sum_{i=1}^{m} (I_{exp}(q_k)_i - \overline{I}_{exp}(q_k))(I_{exp}(q_l)_i - \overline{I}_{exp}(q_l))$$

where

$$\overline{I}_{exp}(q_k) = \frac{1}{m} \sum_{i=1}^{m} I_{exp}(q_k)_i$$

the corresponding correlations

$$r_{kl} = \frac{\sigma(I_{exp}(q_k), I_{exp}(q_l))}{\sigma(I_{exp}(q_k))\sigma(I_{exp}(q_l))}$$

may be computed as $-1 \le r_{kl} \le +1$. If there are no differences among the data sets, the $I_{exp}(q_k)$ values are normally distributed at all q_k and also jointly normal distributed and uncorrelated, and thus $I_{exp}(q_k)$ and $I_{exp}(q_l)$ are independent for all $1 \le k, l \le n$ and $k \ne l$ (**Supplementary Fig. 2**), i.e., the resulting correlations r_{kl} are random values. This random property allows one to evaluate the similarity between data when comparing $m \ge 2$ data frames but also modeled versus experimental scattering intensities without the need to explicitly estimate experimental errors.

The numeric values of the r_{kl} correlation matrix for successive frames may be visualized as a map, or CorMap, using gray levels ranging from −1 (black) to +1 (white), where the extent of the map corresponds to the selected q-range (**Fig. 1**). For example, the CorMap acquired from $m = 20 \times 50$-ms frames of aqueous buffer solution (**Fig. 1a,b**) displays a random pattern without any obvious features, indicating no systematic differences between the scattering intensities. A similar random pattern is noted when analyzing the correlations acquired from a protein that is stable in the X-ray beam (**Fig. 1d,e**). Conversely, if differences arise within a multiple-frame data set, the features of the map change dramatically. The effects of severe radiation damage to a lysozyme sample (**Fig. 1g,h**) show long contiguous areas of both positive and negative correlations at low q, indicating systematic differences between data frames in this region of the profile. Less obvious features can also be revealed, for example, where scaling effects have been deliberately introduced into a data set to reflect poorly recorded sample transmissions (**Fig. 1j,k**). The CorMap shows nonrandom features, even though it is difficult to assess these inconsistencies by overlaying the individual one-dimensional scattering profiles.

Meaningful comparisons of two data frames are also possible. In the absence of systematic differences, the pairwise CorMaps
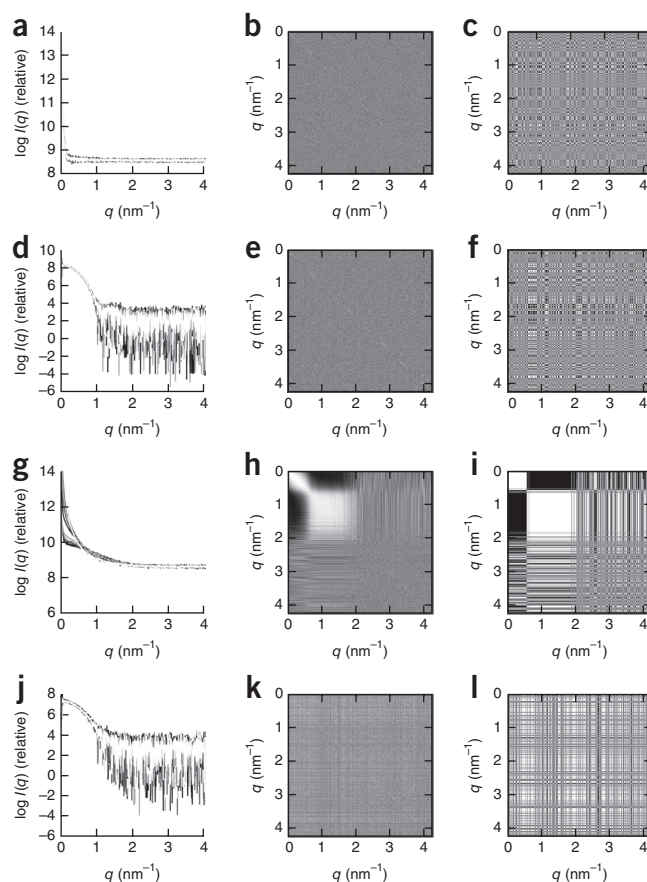
display a randomized lattice pattern (**Fig. 1c,f**). If differences exist, nonrandom and contiguous areas—or 'patches'—of positive (+1) or negative (−1) correlations emerge (**Fig. 1i,l**). Such pairwise comparisons enable the identification of subtle changes between any two data frames or data sets, for example, the onset of radiation damage during data-frame collection, interparticle interference or concentration effects (**Supplementary Figs. 3 and 4**). Furthermore, CorMap can assess the quality of fits by the scattering computed from structural models against SAXS data; nonrandom patterns clearly and reliably point to systematic deviations for incorrect model fits. Shown in **Figure 2** are examples of assessing the fits obtained during the course of *ab initio* bead modeling (**Fig. 2a–c** and **Supplementary Video 1**) and when employing rigid-body modeling (**Fig. 2d–f**).

The probability of similarity in data-data or data-model comparisons may be quantified from pairwise CorMaps by evaluating whether the largest observed patch of contiguous −1 or +1 correlations is likely to occur by chance. The maximum edge length of the patches, *C*, follows the same distribution as that of the longest head-or-tail runs in coin-toss experiments as described by Schilling[6] (**Supplementary Fig. 5**). The probability (*P* value) to obtain an edge length larger than *C* within an n-by-n correlation matrix is calculated from the Schilling distribution with parameters n and *C* (**Supplementary Fig. 6** and Online Methods). If the *P* value is less than a predetermined significance level α, preferably $\alpha = 0.01$ or less[7], the observed differences between any two SAXS patterns may be considered statistically significant.

To compare the statistical properties of CorMap with that of the reduced χ^2 test, i.e., true and false positive rates, we simulated

Figure 2 | CorMap applications. (**a**) SAXS data model fits of a starting (misfitting) and final (fitting) bead model of lysozyme obtained during *ab initio* modeling using the program DAMMIF[12]. (**b**,**c**) Corresponding CorMaps for the initial (**b**; $n = 449$, $C = 137$, $P < 10^{-6}$) and refined bead models (**c**; $n = 449$, $C = 10$, $P = 0.3532$). A spatial superposition of the final refined bead model and the lysozyme X-ray crystal structure is also displayed (black ribbons). (**d**) Comparisons of the experimental scattering from 3-phosphoglycerate kinase, with the calculated scattering from a closed (liganded) or open X-ray crystal structure (top and bottom, respectively[13]). (**e**,**f**) CorMap assessment of the rigid-body model fits (closed form (**e**): $n = 2,039$, $C = 338$, $P < 10^{-6}$; open form (**f**): $n = 2,039$, $C = 18$, $P = 0.011$). (**g**,**h**) Zero-applied-field muon spin rotation data (ZF-μSR) obtained from a crystal of MnSi at 5 K with a fitted polarization function[10] (**g**) and corresponding assessment of the fit using CorMap ($n = 200$, $C = 11$, $P = 0.0456$) (**h**).

several thousand SAXS data sets to represent a number of experimental scenarios including systematic errors, random scaling errors and radiation damage (**Supplementary Fig. 7**). The false positive rates, i.e., the proportion of cases in which differences are flagged when there are none, were found to be 0.010 ± 0.003 for the reduced χ^2 test and 0.019 ± 0.003 for CorMap, for both tests, respectively ($\alpha = 0.01$; **Supplementary Table 1**). The true positive rate of CorMap, i.e., the proportion of correctly identified differences, also known as the statistical power, we found to be similar to that of the reduced χ^2 test (**Supplementary Figs. 8–10**, **Supplementary Table 2** and Online Methods). However, CorMap is uncoupled from the requirement of explicitly defining correct experimental errors, so the test is widely applicable in situations where estimating errors is problematic.

The specification of experimental errors has a major effect on data analysis and model fitting in many physical experiments. For SAXS, there are no agreed-upon standards with respect to correct error estimation or propagation, thus prompting suggestions to find replacements for the reduced χ^2 test. The recently proposed χ^2_{free} (ref. 8), a resampling-based adaptation of the reduced χ^2 test, has the same limitation as the reduced χ^2 test in that χ^2_{free} is valid only when correct error estimates are available. Moreover, under these valid-only circumstances, χ^2_{free} and the reduced χ^2 test are equivalent (**Supplementary Figs. 11** and **12**); if the errors are randomly chosen, correspondingly randomized results may be observed (**Supplementary Fig. 13**). An alternative approach using the paired *t*-test[9] cannot be applied to SAXS data analysis, as the *t*-test's requirement of identically distributed data is not met (**Supplementary Fig. 2b**).

CorMap offers a valid approach to evaluate discrepancies between SAXS data sets or data-model fits that overcomes the issue of correctly estimating experimental errors and identifies the *q*-range where the largest dissimilarity occurs. Beyond SAXS, we anticipate that CorMap may be applied to assess differences between discrete oversampled one-dimensional data from various scattering, reflectometry and other spectroscopic techniques as well as from other fields of physics. We demonstrate one example: a polarization function used to model experimental zero-applied-field muon spin rotation (ZF-μSR) data[10] (**Fig. 2g,h**); here, the patch sizes appear larger in the ZF-μSR CorMap compared to in the SAXS examples because the ZF-μSR spectra have a lower number of experimental points. CorMap is implemented in the ATSAS software package[11] as a command-line module
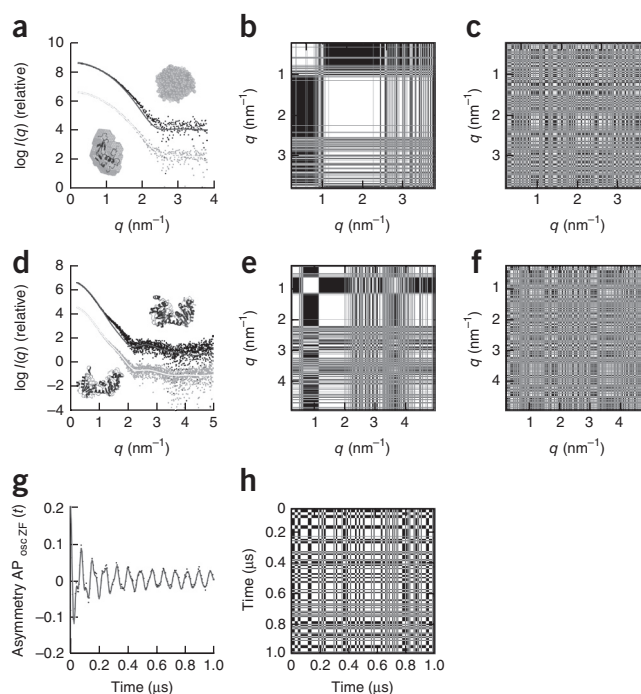
and graphical user interface, and is freely available for academic use (http://www.embl-hamburg.de/biosaxs/download.html). In addition, the source code of the calculation of the *P* value is available to academic users upon request. The results of CorMap should be reported as: "The hypothesis of similarity of experimental data and model could [not] be rejected (CorMap test, *n* points, $C = $ XXX, $P = $ x.xxx)."

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

AUTHOR CONTRIBUTIONS
The initial idea was conceived of and simulation studies were done by D.F. Experimental data were collected by C.M.J. D.F., C.M.J. and D.I.S. participated in critical discussion and wrote the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Bevington, P.R. & Robinson, K.D. in *Data Reduction and Error Analysis for the Physical Sciences* 3rd edn. 36–51 (McGraw-Hill, 2002).
2. Svergun, D.I., Koch, M.H.J., Timmins, P.A. & May, R.P. *Small Angle X-Ray and Neutron Scattering from Solutions of Biological Macromolecules* (Oxford Univ. Press, 2013).
3. Jacques, D.A., Gus, J.M., Svergun, D.I. & Trewhella, J. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 620–626 (2012).
4. Pearson, K. *Philos. Mag.* **50**, 157–175 (1900).
5. Andrae, R., Schulze-Hartung, T. & Melchior, P. Preprint at http://arxiv.org/abs/1012.3754 (2010).
6. Schilling, M.F. *Coll. Math. J.* **21**, 196–207 (1990).
7. Johnson, V.E. *Proc. Natl. Acad. Sci. USA* **110**, 19313–19317 (2013).
8. Rambo, R.P. & Tainer, J.A. *Nature* **496**, 477–481 (2013).
9. Trewhella, J. *et al. Structure* **21**, 875–881 (2013).
10. Amato, A. *et al. Phys. Rev. B Condens. Matter Mater. Phys.* **89**, 184425 (2014).
11. Petoukhov, M.V. *et al. J. Appl. Crystallogr.* **45**, 342–350 (2012).
12. Franke, D. & Svergun, D.I. *J. Appl. Crystallogr.* **42**, 342–346 (2009).
13. Varga, A. *et al. FEBS Lett.* **580**, 2698–2706 (2006).

ONLINE METHODS

Experimental SAXS and sample details. Continuous-flow sample injection was performed at 20 °C using a temperature-controlled EMBL/ESRF automated sample changer equipped with a 1.8-mm quartz capillary sample cell held under vacuum[14]. SAXS data (20×50-ms frames) were collected from several protein samples and their associated matched solvent blanks: (i) glucose isomerase (Hampton Research): 5.8 mg/ml dialyzed against 200 mM Na_2SO_4, 50 mM K_2SO_4, 1 mM $MgCl_2$, 50% (v/v) ^2H_2O, 20 mM HEPES, pH 7.0; (ii) chicken egg white lysozyme (USB Corporation): 2.6 mg/ml dialyzed against 20 mM CH_3COONa, 20 mM HEPES, pH 6.8; (iii) bovine pancreatic RNase A, RNAse; (Sigma): 3.7 mg/ml, 7.5 mg/ml and 15 mg/ml in phosphate buffered saline, pH 7.0; (iv) human serum albumin, HSA, (Sigma): 5 mg/ml, 10 mg/ml and 20 mg/ml in phosphate buffered saline, pH 7.0; (v) HSA (Sigma): 3.65 mg/ml dialyzed against 50 mM HEPES, pH 7.5.

The solvent scattering contributions were subtracted to obtain the scattering from each protein in solution. To test the onset of radiation damage (**Supplementary Fig. 3**), we collected lysozyme SAXS data as described above, without solvent subtraction, from protein samples prepared at 7.9 mg/ml in 40 mM NaCl, 20 mM CH_3COONa, 20 mM HEPES, pH 4.5. Severe radiation damage was tested on lysozyme samples at 8.4 mg/ml in 150 mM NaCl, 40 mM CH_3COONa, pH 3.8 (**Fig. 1j**). To simulate poorly recorded sample transmissions, SAXS data from HSA in HEPES buffer (v) were recorded and deliberate scaling errors applied to randomly selected data frames via multiplication of the sample scattering data before solvent subtraction (**Fig. 1d**). The multiplication factors introduced were, in order: 1.00, 1.01, 1.02, 0.99, 0.96, 1.04, 1.00, 1.00, 1.00, 0.96, 1.00, 1.00, 1.02, 0.96, 1.00, 0.94, 1.04, 0.96, 1.00 and 1.00. In all instances, protein concentrations were estimated using the extinction coefficient calculated from the amino acid sequence (ProtParam[15]).

Experimental setup and data collection. SAXS intensities, $I(q)$ vs. q, where $q = 4\pi \sin\theta/\lambda$, $\lambda = 0.124$ nm is the X-ray wavelength and 2θ is the scattering angle, were collected at the EMBL BioSAXS P12 beam line (PETRA-III storage ring, DESY, Hamburg, Germany) equipped with a DECTRIS Pilatus 2M photon-counting detector. Radial averaging to produce 1D scattering profiles from the recorded 2D data was performed using RADAVER that is integrated into the automated P12 data acquisition and analysis pipeline[16]. All experimental data were recorded on a relative scale. The radial averaging process, which assumes error estimates based on Poisson counting statistics, where the red square donates the beam center (**Supplementary Fig. 14**), is as follows

1. To obtain the intensity for a single value of q, find all N pixels with constant distance (black) from the beam-center (red) (**Supplementary Fig. 14**), without applying anti-aliasing or 'pixel splitting', as this produces correlations in neighboring intensities.

2. Compute the average intensity I_{poi} at distance

$$I_{poi} = \frac{1}{N} \sum_{k=1}^{N} p_k$$

where p_k the kth pixel intensity.

3. Compute the standard error for intensity I_{poi}

$$SE_{poi} = \frac{\sqrt{I_{poi}}}{\sqrt{N}}$$

with $\sqrt{I_{poi}}$ the s.d. of the Poisson variable I_{poi} and $\sqrt{N}$ the reduction factor of the s.e.m.

4. Normalize by exposure time t, intensity of the transmitted beam, d and to unit time T

$$I_{norm} = \frac{I_{poi}}{t \times d} \times T$$

$$SE_{norm} = \frac{SE_{poi}}{t \times d} \times T$$

Scaling intensity and the error term by the same value seems counterintuitive; however, the standard error is an estimate of accuracy of the point estimate, not its variation. On scaling of the intensity, the accuracy does not change.

Scattering patterns as random variables. For assessment of the statistical properties of experimentally recorded scattering intensities at each point in q, several thousand SAXS data sets from water were recorded spanning a momentum transfer of $0.03 < q < 4.4$ nm^{-1}: (i) 10,000 consecutive frames of 0.1 s, (ii) 2,000 consecutive frames of 1 s and (iii) 500 consecutive frames of 10 s. The analysis of the statistical properties of SAXS intensities recorded from water (**Supplementary Fig. 2**) indicate that, for a data set consisting of K q points ($k = 1, …, K$), the experimental $I_{exp}(q_k)$ at each q_k follow Poisson counting statistics[1]. With a sufficiently large number of counts, the distribution of $I_{exp}(q_k)$ limits to a Gaussian distribution with mean N_k and s.d. $\sigma(I_{exp}(q_k)) = \sqrt{N_k}$ (cf. radial averaging above), i.e., the $I_{exp}(q_k)$ values follow a normal distribution in accordance with the central limit theorem (CLT), and the variances decrease with extended exposure time in accordance with the s.e.m. (**Supplementary Fig. 2a,b**).

An analysis of the pairwise joint distribution for any q_k and q_l where $k \neq l$ shows a good agreement with a two-dimensional normal distribution (**Supplementary Fig. 2c**), whereas the correlation matrix (**Supplementary Fig. 2d**) shows that no two points are correlated with each other. Therefore, the scattering intensities recorded at each value of q are statistically independent, i.e., (pairwise) jointly normal distributed and uncorrelated.

Correlation Map: theoretical distribution, *P* value and approximation. When analyzing 5,000 independent pairwise correlation matrices derived from 10,000 experimental water frames, we found that the distribution of the edge length C of the largest contiguous area of similar correlation, i.e., contiguous patches of −1 or +1, may be described by the same distribution that models the longest head-or-tail runs in coin-toss experiments, for example, as in Schilling[6] (**Supplementary Fig. 5**). Consequently, the probability to obtain a run R_n of more than C consecutive data points with the same direction of correlation is given by

$$P(R_n > C) = 1 - \frac{A_n(C)}{2^n}$$

The longest run of heads defined by Schilling's equation (1) defines $A_n(C)$ as

$$A_n(C) = \begin{cases} \sum_{j=0}^{C} A_{n-1-j}(C) & \text{for} \quad n > C \\ 2^n & \text{for} \quad n \leq C \end{cases}$$

where "the distribution of the longest run of heads or tails for a fair coin is simply the distribution of the longest run of heads alone for a sequence containing one fewer coin toss, shifted to the right by one"[6].

The toss-a-coin principle directly applies to a sequence of one-dimensional discrete data. Indeed, if two data sets are identical up to noise, the difference between them is that of two random numbers. The difference of two normally distributed random variables is a normally distributed random variable. The mean is the difference of the two means, i.e., 0 in our case. Hence, owing to symmetry of the normal distribution, the probability of the variable being positive or negative is exactly as for tossing a fair coin, i.e., 0.5. And the two subsequent data points are independent, exactly as for toss-a-coin.

The probability of how likely it is that two SAXS data sets are similar may thus be obtained from a pairwise CorMap consisting of $n \times n$ data points by calculating the exact Schilling distribution for n points and computing the probability of obtaining an edge length larger than the edge length of the observed patch size. If this probability is less than the predefined significance level α, the hypothesis of similarity between the frames has to be rejected. Examples of the Schilling distribution for various n are shown in **Supplementary Figure 6** as well the effects of SAXS data rebinning in q, for example, averaging intensities over a selected Δq intervals, that is often employed as a data processing step.

The P value has a clear statistical meaning and, as such, is a correct measure of similarity. The situation is similar to the case of reduced χ^2 test, where, given the χ^2 value, the goodness of fit should be judged on the basis of the associated P value computed from the χ^2 distribution. In practice, χ^2 itself is most often reported (cf. **Supplementary Fig. 1**) and not its P value. To have a similar shorthand measure for the CorMap, one may consider, as a rule of thumb, a z-score approximation directly related to the longest edge length

$$z = \frac{C - \log_2(n)}{2.5} \approx N(0,1)$$

In practical terms, z values exceeding 3 indicate statistically significant differences between the data sets. However, use of the exact P value is strongly recommended.

Correlation Map: Bonferroni correction for multiple testing. In instances when it is necessary to evaluate multiple pairwise tests derived from several comparisons (for example, **Supplementary Figs. 3** and **4**), the P values for the CorMaps are adjusted by the Bonferroni method

$$P_{\text{adj}} = m \times p$$

where $p = P(R_n > C)$ and m is the number of tests. The adjusted P value is then compared to the predefined significance level α.

Of note, other adjustments for multiple testing are possible and possibly more powerful (Bonferroni-Holm, Benjamini-Hochberg); however, in the context of this manuscript, the classical approach of Bonferroni was chosen for its simplicity.

Correlation Map: visualization. In Octave (http://www.octave.org/) or Matlab (http://www.mathworks.com/):

```
% Load two or three column data (s,I,err) as ascii
% without headers or footers
d1 = dlmread('file1.dat');
d2 = dlmread('file2.dat');
% Build data matrix, one set of intensities per column
data = [d1(:,2), d2(:,2)]';
% Plot the correlation map
imagesc(corr(data, data))
```

Comparing statistical tests. To assess the ability of the CorMap to hold the type I error (false positive rate) and evaluate its power (true positive rate) in comparison to the reduced χ^2 test, we generated a large pool of data consisting of several thousand simulated scattering profiles of monomeric bovine serum albumin (BSA, PDB: 3V03) and tested several hypotheses. Simulated SAXS data are very useful for comparative testing in that they provide a standard frame of reference that consists of a sufficiently large sample set for significance testing without being influenced by unknowable experimental uncertainties in intensity and error estimates. Notably, this approach ensures that the CorMap is always compared to a reduced χ^2 test with known errors. The simulated data were computed by taking the expected solution scattering intensities, $I(q_k)$, from the atomic structure of BSA using Crysol[17] and introducing statistical variations based on the variation information obtained from real data (**Supplementary Fig. 2**). All calculations were done using Octave (http://www.octave.org/) and Matlab (http://www.mathworks.com/). Each of the simulated data sets is composed of $m = 10$ data frames to more realistically model frame-by-frame variations.

The hypotheses used to test the reduced χ^2 test and the CorMap were formulated to reflect 'real-world' scenarios and were as follows:

H0: no systematic differences among SAXS data frames (10,000 times ten frames of simulated data to evaluate type 1 error, i.e., the probability that a difference is detected although there are none).

H1: random shifts in $I(q)$ simulating systematic errors, for example, incorrectly matched solvents or sample fluorescence (2,000 times ten simulated frames).

H2: random scaling to simulate systematic error due to incorrect sample transmissions (2,000 times ten simulated frames).

H3,H4,H5: examples of radiation damage (2,000 times ten simulated frames). Note, as it is very difficult to specifically characterize radiation damage at the molecular level, the H3–H5 scenarios were modeled using simple additive contributions to the scattering patterns, based on empirical estimates of radiation damage observed during experiments at P12 (ref. 18) (**Supplementary Fig. 7**).

Both the CorMap and reduced χ^2 were applied to all pairs of frames within a data set, and the P value for each test was adjusted for multiple testing by the Bonferroni correction. The smallest

P value after this adjustment was selected as the global result of all pairwise tests, and the number of statistically significant results across the simulated data sets was counted, i.e., where the adjusted *P* value was less than the significance level $\alpha = 0.01$, to determine empirical values for type I error and power. Finally the 99% Clopper-Pearson confidence intervals[19] were computed to facilitate the comparison of results; overlapping confidence intervals indicate equivalent tests at that effect size; fully separated intervals indicate significant differences between the tests. Both the reduced χ^2 and the CorMap tests approximately hold the type I error level at $\alpha = 0.01$ (0.010 ± 0.003 and 0.019 ± 0.003, respectively; **Supplementary Table 1**). The CorMap test is more powerful, according to the Clopper-Pearson confidence intervals, than the reduced χ^2 test in four out of five alternatives (**Supplementary Fig. 8b–e** and **Supplementary Table 2**). Only when detecting random shifts (H1) does χ^2 display slightly more power than the CorMap (**Supplementary Fig. 8a**).

For comparing the performance of the CorMap with the reduced χ^2 test for assessing SAXS data model fits, an additional 10,000 simulated data sets of BSA were generated and compared to the scattering calculated for a set of BSA models (**Supplementary Fig. 9**). Fits against the data were determined for the native BSA monomer (**Supplementary Fig. 10a,c**) to estimate the type I error, as well as a set of 23 hypothetical structures where the Tyr496-to-Val497 bond angle was rotated in consecutive steps of 1° to introduce ever-increasing systematic differences relative to the native structure (**Supplementary Fig. 10b,d**). Both the CorMap and χ^2 test hold the type I error level (0.012 ± 0.003 and 0.013 ± 0.003, respectively), and the CorMap is about as powerful to detect systematic differences in data fitting as the reduced χ^2 test (**Supplementary Fig. 10e**).

Evaluation of χ^2_{free}. After performing comparisons of the CorMap with the reduced χ^2, we also intended to rigorously compare its statistical power for discerning data model fits with the recently developed χ^2_{free} test[8]. However, when 23,000 pairwise model-fit test comparisons of χ^2 and χ^2_{free} are plotted, the corresponding values of χ^2 and χ^2_{free} are, up to the sampling variation of χ^2_{free}, identical. A one-to-one correspondence between χ^2 and χ^2_{free} is observed for all data model fits when the errors on the data have been correctly specified (**Supplementary Fig. 11**). To verify that this is not an artifact, we tested another example with correct, but different, error structure. We simulated 1,000 frames of BSA with the commonly assumed constant 3% relative error across the whole *q*-range instead of an empirical error structure

(**Supplementary Fig. 1b**). Again, we found the values of χ^2 and χ^2_{free} are, up to the sampling variation of χ^2_{free}, in essence identical (**Supplementary Fig. 12**). Therefore, the result of one-to-one correspondence, if the correct errors are available, is independent of the applied error structure. We note that the observed upwards shift from the diagonal, i.e., from a perfect 1:1 correlation between χ^2 and χ^2_{free} (**Supplementary Figs. 11** and **12b**), may be attributed to the maximum particle dimension, D_{max}, parameter of χ^2_{free}; changing the estimated D_{max} modulates the outcome of the computation.

Four additional cases comparing χ^2_{free} and the reduced χ^2 test were considered (**Supplementary Fig. 13**): (i) the general proportions of the errors are correct, but exactly half the magnitude; (ii) likewise, but twice the magnitude; (iii) random permutation of the previously correct error estimates to random positions; and (iv) assumption of a constant 75% relative error. A total of 23,000 model fit test comparisons of χ^2 and χ^2_{free} were calculated and compared to the 23,000 cases with correct error structure (identical to **Supplementary Fig. 11**, marked with a circle in each panel of **Supplementary Fig. 13** as a reference). It is notable that with incorrect errors, both tests report values that indicate that differences are present (i.e., outside the interval 0.9–1.1) even in the cases where there are no statistical differences. This corresponds to a false positive error rate of 100%; hence, both tests are invalid. Any previously reported improvements of stability of χ^2_{free} over χ^2 (ref. 8) have thus to be attributed to coincidence. Consequently, χ^2_{free} affords no advantage over the reduced χ^2 test when data-model fits are evaluated because both are equivalent under valid test conditions. Therefore, the statistical power of a valid χ^2_{free} test in comparison to the CorMap is the same as that of a valid reduced χ^2 test, and thus χ^2_{free} was not considered in more detail. All results shown here were computed using the reference implementation of χ^2_{free}, which was provided on request by R.P. Rambo.

14. Round, A. *et al. Acta Crystallogr. D Biol. Crystallogr.* **71**, 67–75 (2015).
15. Gasteiger, E. *et al.* in *The Proteomics Protocols Handbook* (ed. Walker, J.M.) 571–607 (Humana Press, 2005).
16. Franke, D., Kikhney, A.G. & Svergun, D.I. *Nucl. Inst. Methods Phys. Res. A* **689**, 52–59 (2012).
17. Svergun, D., Barberato, C. & Koch, M.H.J. *J. Appl. Crystallogr.* **28**, 768–773 (1995).
18. Jeffries, C.M., Graewert, M.A., Svergun, D.I. & Blanchet, C.E. *J. Synchrotron Radiat.* **22**, 273–279 (2015).
19. Clopper, C.J. & Pearson, E.S. *Biometrika* **26**, 404–413 (1934).