IUCrJ
ISSN 2052-2525
BIOLOGY | MEDICINE

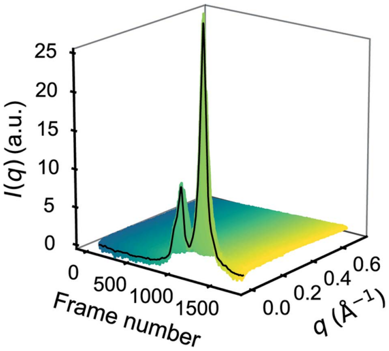
Received 14 December 2020
Accepted 14 January 2021

Edited by K. Moffat, University of Chicago, USA

‡ Equal contributors.

Keywords: deconvolution; small-angle X-ray scattering; time-resolved SAXS; AEX-SAXS; high-throughput SAXS; ligand titration; regularized alternating least squares; multivariate curve resolution; singular value decomposition; pair-distance distribution function.

Supporting information: this article has supporting information at www.iucrj.org



REGALS: a general method to deconvolve X-ray scattering data from evolving mixtures

Steve P. Meisburger,‡ Da Xu‡ and Nozomi Ando*

Department of Chemistry and Chemical Biology, Cornell University, 259 East Avenue, Ithaca, NY 14853, USA.
*Correspondence e-mail: nozomi.ando@cornell.edu

Mixtures of biological macromolecules are inherently difficult to study using structural methods, as increasing complexity presents new challenges for data analysis. Recently, there has been growing interest in studying evolving mixtures using small-angle X-ray scattering (SAXS) in conjunction with time-resolved, high-throughput or chromatography-coupled setups. Deconvolution and interpretation of the resulting datasets, however, are nontrivial when neither the scattering components nor the way in which they evolve are known *a priori*. To address this issue, the *REGALS* method (regularized alternating least squares) is introduced, which incorporates simple expectations about the data as prior knowledge, and utilizes parameterization and regularization to provide robust deconvolution solutions. The restraints used by *REGALS* are general properties such as smoothness of profiles and maximum dimensions of species, making it well suited for exploring datasets with unknown species. Here, *REGALS* is applied to the analysis of experimental data from four types of SAXS experiment: anion-exchange (AEX) coupled SAXS, ligand titration, time-resolved mixing and time-resolved temperature jump. Based on its performance with these challenging datasets, it is anticipated that *REGALS* will be a valuable addition to the SAXS analysis toolkit and enable new experiments. The software is implemented in both MATLAB and Python and is available freely as an open-source software package.

1. Introduction

Small-angle X-ray scattering (SAXS) is a widely used technique for obtaining structural information from macromolecules in solution (Putnam *et al.*, 2007). Increasingly, SAXS is applied to evolving mixtures of different molecules or conformational states (Vestergaard & Sayers, 2014; Meisburger *et al.*, 2017) during titrations (Brosey & Tainer, 2019), chromatographic separation (Pérez & Vachette, 2017) or time-resolved experiments (Kathuria *et al.*, 2011; Neutze & Moffat, 2012; Kirby & Cowieson, 2014). However, because of the fundamental limitations in the information content of the SAXS signal (Moore, 1980), multiple structures in a mixture cannot be resolved from each profile in an unambiguous manner. This inherent ambiguity can be mitigated by combining multiple measurements and carefully incorporating prior knowledge. The individual components can then be separated mathematically by analyzing the dataset as a whole using a physicochemical model for how the mixture evolves (Williamson *et al.*, 2008; Cho *et al.*, 2010; Minh & Makowski, 2013) or known scattering curves of each component (Konarev *et al.*, 2003). Often, however, both the scattering curves and physicochemical model are unknown before the experiment is performed and must be inferred from the data themselves. In such cases, the challenge is to identify appropriate mathematical tools to incorporate more general

https://doi.org/10.1107/S2052252521000555

physically motivated restraints that lead to a reliable and accurate model-free separation.

In dilute solution, SAXS intensities from non-interacting components combine linearly in proportion to their relative concentrations. A SAXS dataset from a mixture can therefore be described as the convolution of the concentration and SAXS profiles, and deconvolution can be performed using matrix factorization techniques such as singular value decomposition (SVD) (Henry & Hofrichter, 1992; Hendler & Shrager, 1994). However, to recover the scattering from each component, the basis vectors from SVD must be recombined using prior knowledge about what constitutes a physically valid solution. The field of chemometrics has developed a number of algorithms for solving this problem, known as multivariate curve resolution or MCR (de Juan & Tauler, 2003; Jaumot *et al.*, 2004). When a physicochemical model is available, the alternating least-squares (MCR-ALS) algorithm can perform deconvolution using the model as a hard restraint (Jaumot *et al.*, 2004). In the context of SAXS, deconvolution with hard restraints has been applied to time-resolved experiments (Cho *et al.*, 2010; Chen *et al.*, 1998; Segel *et al.*, 1998; Akiyama *et al.*, 2002), equilibrium titrations (Williamson *et al.*, 2008; Blobel *et al.*, 2009; Minh & Makowski, 2013; Cichocki & Zdunek, 2007), unfolding experiments (Chen *et al.*, 1996; Ayuso-Tejedor *et al.*, 2011), protein–micelle interactions (Lipfert *et al.*, 2007) and fibril formation (Herranz-Trillo *et al.*, 2017). Interestingly, MCR can be performed without assuming a hard model by imposing soft restraints such as positivity, unimodality and local rank (Jaumot *et al.*, 2004). Such model-free deconvolution is seldom applied to SAXS data because soft restraints are rarely sufficient to provide a robust and unique solution on their own (de Juan & Tauler, 2003). One exception is SAXS data collected with in-line size-exclusion chromatography (SEC-SAXS), where MCR-ALS has been combined with evolving factor analysis (EFA) (Maeder, 1987) to separate overlapping elution peaks (Meisburger *et al.*, 2016; Hopkins *et al.*, 2017).

Although the SVD and MCR algorithms are well suited to certain SAXS experiments, they are a poor fit for other more challenging datasets. A notable example is SAXS data collected with in-line anion-exchange chromatography (AEX) (Hutin *et al.*, 2016). AEX separates according to charge by applying the sample to cationic media and eluting with a salt gradient. In SAXS, the salt gradient produces a changing background scattering that must be accounted for. Because this changing background violates certain assumptions of the EFA method, model-free deconvolution of AEX-SAXS data is not possible with EFA. We previously encountered this issue when analyzing AEX-SAXS data from the large subunit of *Bacillus subtilis* ribonucleotide reductase (*Bs*RNR) (Parker *et al.*, 2018). To overcome this challenge, we incorporated a simple assumption as additional prior information, namely, that the background scattering must change gradually over time. Using the ALS algorithm with smoothness regularization applied to the concentration of background scattering components, we achieved a clean separation of multiple protein and buffer components (Parker *et al.*, 2018).

Here, we examine the generality of this strategy for the model-free deconvolution of other complex types of SAXS data where traditional 'soft' restraints are insufficient. We describe the *REGALS* (regularized ALS) toolset and introduce the *REGALS* software package, which is adaptable by design, freely available and open source. We then demonstrate the application of *REGALS* to a wide variety of SAXS experiments from evolving mixtures. Unlike most deconvolution methods that impose a physicochemical model, *REGALS* relies on very general parametric models for the SAXS profiles and concentration curves. The models include two types of restraint: smoothness and compact support. In AEX-SAXS, for example, each elution peak is assumed to be non-zero over a particular range (compact support) and the background components are assumed to be smooth. For the *Bs*RNR dataset, we find this is sufficient to deconvolve the protein scattering peaks. In other cases, such as equilibrium titration and time-resolved SAXS, where concentrations are typically non-zero in all (or nearly all) data frames, the assumption that concentrations have compact support is insufficient. However, compact support can be applied to the SAXS profiles in real space by imposing a maximum particle dimension. We show that compact support in real space, as well as boundary conditions applied to the concentration basis functions, provide sufficient information for successful deconvolution of such data.

2. Theory

2.1. Background

A dilute evolving mixture of K components scatters X-rays according to the following linear model:

$$I_{\rm calc.}(q,x) = \sum_{k=1}^{K} y_k(q)\,c_k(x), \qquad (1)$$

where $y_k(q)$ are the individual SAXS profiles and $c_k(x)$ are the relative concentrations. The SAXS profiles depend on the scattering vector magnitude $q = (4\pi/\lambda)\sin\theta$, where λ is the X-ray wavelength and 2θ is the scattering angle. The concentration profiles depend on an independent variable x (representing time, ligand concentration *etc*). Since intensities are measured at discrete values of q and x, equation (1) can be written in matrix form as follows:

$$\mathbf{I}_{\rm calc.} = \sum_k \mathbf{y}_k \otimes \mathbf{c}_k$$

$$= \begin{bmatrix} | & | & & | \\ \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_K \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & \mathbf{c}_1 & - \\ - & \mathbf{c}_2 & - \\ & \vdots & \\ - & \mathbf{c}_K & - \end{bmatrix} = \mathbf{Y}\mathbf{C}^{\rm T}, \qquad (2)$$

where $\mathbf{I}_{\rm calc.}$ contains scattering profiles arranged side by side as column vectors. Here and throughout this section, the intensity matrix has dimensions of $M \times N$ (N scattering profiles with M discrete values of q). Hence, $\mathbf{Y}$ is $M \times K$ and $\mathbf{C}$ is $N \times K$.

Our aim is to determine $\mathbf{Y}$ and $\mathbf{C}$ given the measured intensity $\mathbf{I}_{\rm meas.}$, which contains noise. This is accomplished by minimizing the least-squares error between data and model:

$$\chi^2 = \sum_{ij} \sigma_{ij}^{-2} \left[(\mathbf{I}_{\rm meas.})_{ij} - (\mathbf{I}_{\rm calc.})_{ij} \right]^2, \qquad (3)$$

where σ_{ij} are the standard errors of the measured intensity. In the following, we assume that the experimental errors depend only on q, so that equation (3) can be written as a Frobenius norm of the error-weighted residual:

$$\chi^2 = \left\| \boldsymbol{\Sigma}^{-1} (\mathbf{I}_{\rm meas.} - \mathbf{I}_{\rm calc.}) \right\|_{\rm F}^2, \qquad (4)$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with $\Sigma_{ii} = N^{-1} \sum_{j=1}^{N} \sigma_{ij}$. This simplifying assumption is approximately correct for the datasets considered here.

In general, minimizing χ^2 is not sufficient to determine $\mathbf{Y}$ and $\mathbf{C}$ uniquely. The main issue is that basis vectors can be mixed (or 'rotated') without changing χ^2: for any non-singular $K \times K$ matrix Ω, replacing $\mathbf{Y} \rightarrow \mathbf{Y}\Omega$ and $\mathbf{C} \rightarrow \mathbf{C}\Omega^{-\rm T}$ leaves the product $\mathbf{Y}\mathbf{C}^{\rm T}$ unchanged. Thus, the primary challenge of deconvolution is to impose appropriate restraints that provide a unique and physically meaningful solution.

Deconvolution problems resembling equation (2) arise in many experimental contexts. A common approach is to apply SVD (Henry & Hofrichter, 1992; Hendler & Shrager, 1994), by which an error-weighted data matrix is decomposed as follows:

$$\mathbf{U}\mathbf{S}\mathbf{V}^{\rm T} = \Sigma^{-1} \mathbf{I}_{\rm meas.}, \qquad (5)$$

where $\mathbf{U}$ has the left singular vectors as columns, $\mathbf{V}$ contains the right singular vectors as columns and $\mathbf{S}$ contains the singular values along the diagonal in decreasing order. The uniqueness of the decomposition results from the fact that the singular vectors are on an orthonormal basis.

The singular values $s_j = S_{jj}$ are positive and indicate the importance, or weight, of each pair of left and right singular vectors. When the number of observations (N) is much larger than the number of independent components in the signal (which is generally the case for the examples studied here), most of the singular values will be small and represent the noise in the data, while a few large singular values correspond to the signal of interest. To detect significant singular values, it is useful to calculate a normalized singular value, as follows:

$$s_j' = \left(s_j - M^{1/2} \right) / N^{1/2}, \qquad (6)$$

where M and N are the numbers of rows and columns, respectively, in the data matrix. If no signal is present, random matrix theory shows that $s_j' < 1$ in the limit where the data matrix is large [see Vershynin (2012), and references therein]. Thus, components corresponding to signal above the noise are expected to have $s_j' > 1$.

By retaining only the K most important singular vectors ($\mathbf{U} \rightarrow \mathbf{U}_K$ *etc.*), one obtains an approximate (reduced-rank) representation of the data. Thus, a solution for $\mathbf{Y}$ and $\mathbf{C}$ can be constructed from SVD as follows:

$$\mathbf{Y}_{\rm SVD} = \boldsymbol{\Sigma}\mathbf{U}_K\mathbf{S}_K^{1/2}, \qquad \mathbf{C}_{\rm SVD} = \mathbf{V}_K\mathbf{S}_K^{1/2}. \qquad (7)$$

Here, the singular-value weights have been distributed evenly between the SAXS and concentration basis vectors, but other choices could be made depending on the normalization conditions.

Although SVD provides a unique low-rank decomposition of the data, the orthonormality of the singular vectors often produces non-physical results. For instance, the component SAXS profiles or concentrations might have negative values. It is therefore often necessary to further unmix (or 'rotate') the SVD basis vectors by applying physical restraints (Chen *et al.*, 1996; Lipfert *et al.*, 2007; Segel *et al.*, 1998; Williamson *et al.*, 2008). In traditional MCR techniques, physical restraints are imposed using 'hard' or 'soft' models, whose applicability depends on the type of experiment performed and prior knowledge. Alternatively, prior information can be imposed through Tikhonov–Miller regularization, where additional functions are minimized at the same time as χ^2 (Tikhonov & Arsenin, 1977; Miller, 1970). As described above, in an AEX-SAXS experiment, the expectation that background scattering varies gradually over time can be enforced using a regularization function that penalizes large oscillations (Parker *et al.*, 2018).

Regularization is also used in conventional SAXS data analysis to infer the pair-distance distribution function, or $P(r)$, from the measured intensity (Hansen & Pedersen, 1991). Essentially, $P(r)$ represents the probability of two electrons being a distance r apart in the sample, and it is related to the scattering intensity by a Fourier transform:

$$I(q) = 4\pi \int_0^{d_{\rm max}} P(r) \, \frac{\sin(qr)}{qr} \, {\rm d}r, \qquad (8)$$

where the integral terminates at the maximum particle dimension, $d_{\rm max}$ [since $P(r) = 0$ for $r > d_{\rm max}$]. Although equation (8) can be inverted analytically, in practice the intensity is measured over a finite q range, and thus inversion is an ill-posed problem. Since the Fourier transform is a linear operator, Tikhonov–Miller regularization can be applied. $P(r)$ is discretized as a vector $\mathbf{u}$ of length R, which samples values of $P(r)$ on a uniform grid with spacing Δr. Equation (8) can then be written as

$$\mathbf{I}_{\rm calc.} = \mathbf{A}\,\mathbf{u}, \qquad (9)$$

where $\mathbf{I}_{\rm calc}$ is a vector of length M and $\mathbf{A}$ is an $M \times R$ matrix with elements

$$A_{ij} = 4\pi\Delta r \, \frac{\sin(q_i r_j)}{q_i r_j}. \qquad (10)$$

The standard indirect Fourier transform (IFT) method for SAXS data minimizes the χ^2 between $I_{\rm calc.}$ and $I_{\rm meas.}$ plus a regularization term:

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \left[\left| \boldsymbol{\Sigma}^{-1} (\mathbf{I}_{\rm meas} - \mathbf{A}\,\mathbf{u}) \right|^2 + \lambda |\mathbf{B}\,\mathbf{u}|^2 \right]. \qquad (11)$$

Typically, the matrix $\mathbf{B}$ performs a discrete approximation of the second derivative (Hansen, 2012; Press, 2007), which enforces smoothness by penalizing wildly oscillating solutions. The regularization parameter (or Lagrange multiplier) λ controls the tradeoff between minimizing χ^2 and minimizing the regularization function. The optimization problem is solved by the method of normal equations, with the (formal) result

$$\hat{\mathbf{u}} = \left(\mathbf{A}^{\rm T}\boldsymbol{\Sigma}^{-2}\mathbf{A} + \lambda\mathbf{B}^{\rm T}\mathbf{B}\right)^{-1}\mathbf{A}^{\rm T}\boldsymbol{\Sigma}^{-2}\mathbf{I}_{\rm meas}. \qquad (12)$$

In this study, we describe a general method for deconvolving SAXS data from mixtures that applies regularization to both the concentration and SAXS profile basis vectors. We first formulate the deconvolution problem [equation (2)] using a parametric representation of the basis vectors, similar to the IFT example above. This parametric form allows the SAXS profiles to be represented in the real-space [$P(r)$] basis if desired. Then, we describe the *REGALS* algorithm for minimizing the sum of the χ^2 [equation (3)] and regularization terms.

2.2. Deconvolution by regularized least squares

In order to deconvolve SAXS data from evolving mixtures, we introduce a method to impose mathematical constraints that embody prior information (or general expectations) about a SAXS experiment. The first way that constraints are imposed is through a parameterization of the basis vectors:

$$\mathbf{y}_k = \mathbf{A}_k\mathbf{u}_k, \qquad \mathbf{c}_k = \mathbf{A}'_k\mathbf{v}_k, \qquad (13)$$

where $\mathbf{u}_k$ and $\mathbf{v}_k$ are the parameter vectors for the SAXS profile and concentration bases, respectively. Here and in the following equations, primed functions or matrices refer to the concentration basis, in order to distinguish them from the SAXS profile basis. We implemented three types of basis vector: simple, smooth and real-space [Fig. 1(*a*)]. In a simple basis vector, $\mathbf{A}_k$ is the identity matrix and the parameter vector encodes the basis vector directly. In a smooth basis vector, $\mathbf{A}_k$ performs a linear interpolation from a uniform grid of control points to the experimental grid, which need not be uniform. Finally, in a real-space basis vector (which applies exclusively to SAXS profiles), $\mathbf{u}_k$ samples $P(r)$ on a uniform grid, and $\mathbf{A}_k$ is given by equation (10). Crucially, the global model can contain a mixture of different parameterizations. This model was implemented using a flexible object hierarchy [Fig. 1(*b*)] as described in the *Methods* section.

The second way constraints are imposed is through regularization. The regularization functions $\mathcal{B}$ embody prior information (or expectations) about the data, such as smoothness in data or parameter space, and are minimized along with χ^2:

$$\{\hat{\mathbf{u}}, \hat{\mathbf{v}}\} = \underset{\mathbf{u},\mathbf{v}}{\operatorname{argmin}}\left[\chi^2(\mathbf{u},\mathbf{v}) + \mathcal{B}(\mathbf{u}) + \mathcal{B}'(\mathbf{v})\right], \qquad (14)$$

Here, $\mathbf{u}$ and $\mathbf{v}$ refer to global parameter vectors that are constructed by concatenating parameter vectors for the individual basis functions (for example, $\mathbf{u}$ is $\mathbf{u}_1, \ldots, \mathbf{u}_K$ placed end to end), and χ^2 is calculated from equations (4) and (13) as follows:

$$\chi^2(\mathbf{u},\mathbf{v}) = \left\|\boldsymbol{\Sigma}^{-1}\left[\mathbf{I}_{\rm meas.} - \sum_k \mathbf{A}_k\mathbf{u}_k\mathbf{v}_k^{\rm T}(\mathbf{A}'_k)^{\rm T}\right]\right\|_{\rm F}^2. \qquad (15)$$

The regularization functions are a sum of quadratic regularizers acting on each component's parameter vector:

$$\mathcal{B}(\mathbf{u}) = \sum_k \lambda_k\left|\mathbf{B}_k\mathbf{u}_k\right|^2, \qquad \mathcal{B}'(\mathbf{v}) = \sum_k \lambda'_k\left|\mathbf{B}'_k\mathbf{v}_k\right|^2. \qquad (16)$$

The regularization parameters λ_k and λ'_k control the tradeoff between minimizing χ^2 and each regularizing function. For smoothness regularization, $\mathbf{B}_k$ is a discrete approximation of the second derivative (Press, 2007). Zero boundary conditions are optionally imposed by removing the parameters on the boundary and deleting the corresponding rows of $\mathbf{A}_k$ and $\mathbf{B}_k$ (Press, 2007).

2.3. *REGALS* algorithm

The optimization problem described in the previous section [equations (14), (15) and (16)] is nonlinear and therefore does not afford a straightforward solution. We chose to adapt the alternating least-squares (ALS) algorithm, which is often used in classic MCR (Jaumot *et al.*, 2015, 2004; de Juan & Tauler, 2003). ALS replaces the single nonlinear optimization problem with two linear problems that are solved in an alternating fashion over many iterations. Beginning with an

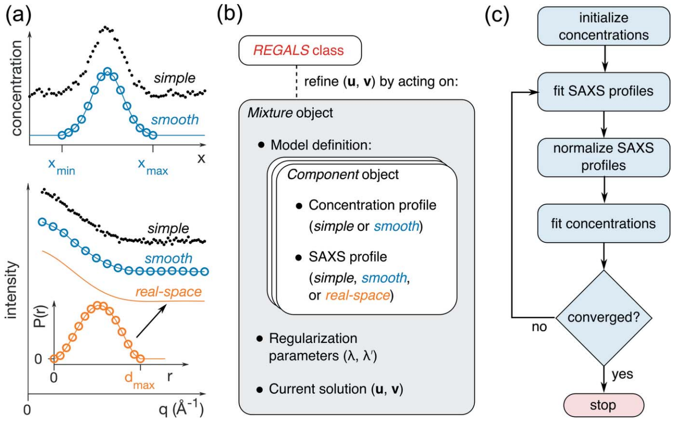


Figure 1
Overview of the *REGALS* method. (*a*) Parametric basis vectors representing concentrations (top panel) and SAXS profiles (bottom panel). In simple vectors, each sample (q or x) is given an independent parameter (black dots). A smooth vector represents the data by linear interpolation between control points (blue circles) over the region of support ($x_{\rm min}$ and $x_{\rm max}$, top panel). A real-space vector samples the $P(r)$ function (orange circles) up to the maximum particle dimension ($d_{\rm max}$) and the corresponding SAXS intensities (orange curve) are calculated by Fourier transform [equation (8)]. (*b*) Experimental restraints are expressed in the software by mixing and matching basis vector types using a flexible object hierarchy. The basis vectors representing SAXS profiles ($\mathbf{u}$) and concentrations ($\mathbf{v}$) are refined by methods in the high-level *REGALS* class. (*c*) The refinement algorithm based on regularized ALS. At each iteration, regularized linear least-squares fits are performed on the SAXS profiles [equation (17)] and concentrations [equation (18)] in an alternating fashion until a user-specified convergence test is satisfied.

initial guess, one set of basis functions is optimized (*e.g.* the concentrations) with the other held fixed, and then the other basis functions are optimized. This is repeated until the change in basis vectors from one iteration to the next is smaller than a certain tolerance, or the maximum number of iterations has been reached.

The *REGALS* algorithm solves equation (14) iteratively using ALS with regularization [Fig. 1(*c*)]. First, an initial guess is made for the concentration basis parameters ($\hat{\mathbf{v}}$). This can be supplied by the user or generated automatically based on the parameterization type and boundary conditions. In the first least-squares step, the SAXS basis functions are optimized while the concentrations are held fixed:

$$\hat{\mathbf{u}} := \underset{\mathbf{u}}{\operatorname{argmin}}\left[\chi^2(\mathbf{u}, \hat{\mathbf{v}}) + \mathcal{B}(\mathbf{u})\right]. \qquad (17)$$

The profiles are then normalized according to a their parameterization type; for simple and smooth types, the parameters are divided by the root-mean-squared value, while for the real-space type, the parameters are normalized by the scattering intensity at $q = 0$ calculated from the area under the $P(r)$ curve [see equation (8)]. In the second least-squares step, the concentration basis functions are optimized while the SAXS profiles are held fixed:

$$\hat{\mathbf{v}} := \underset{\mathbf{v}}{\operatorname{argmin}}\left[\chi^2(\hat{\mathbf{u}}, \mathbf{v}) + \mathcal{B}'(\mathbf{v})\right]. \qquad (18)$$

Statistics are calculated at this stage, including the change in the basis vector from the previous iteration (the sum of the absolute value of the difference) and the χ^2 for the current model [equation (3)]. Finally, the cycle is repeated until reaching convergence according to user-specified termination conditions. Further details about parameter estimation, error analysis and implementation can be found in the *Methods* section.

3. Methods

3.1. Computational details

3.1.1. Least-squares optimization in each *REGALS* iteration. The two regularized linear least-squares problems within each *REGALS* iteration [equations (17) and (18)] are solved using the method of normal equations. For the SAXS profile basis [equation (17)], the best-fit parameters satisfy K sets of linear equations (with $k = 1, 2, \ldots, K$):

$$\lambda_k \mathbf{B}_k^{\mathrm{T}} \mathbf{B}_k \mathbf{u}_k + \sum_{l=1}^{K} \mathbf{M}_{kl} \mathbf{u}_l = \mathbf{b}_k, \qquad (19)$$

where

$$\mathbf{M}_{kl} = (\mathbf{c}_k \cdot \mathbf{c}_l)\left(\mathbf{A}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-2} \mathbf{A}_l\right), \qquad (20)$$

$$\mathbf{b}_k = \mathbf{A}_k^{\mathrm{T}} \boldsymbol{\Sigma}^{-2} \mathbf{I}_{\text{meas.}} \mathbf{c}_k. \qquad (21)$$

Note that these equations can be combined and written in the form $(\mathbf{M} + \mathbf{H})\mathbf{u} = \mathbf{b}$, making them straightforward to solve using standard numerical methods. Similarly, the parameters for the concentration basis [equation (18)] are found by solving the K sets of linear equations:

$$\lambda'_k (\mathbf{B}'_k)^{\mathrm{T}} \mathbf{B}'_k \mathbf{v}_k + \sum_{l=1}^{K} \mathbf{M}'_{kl} \mathbf{v}_l = \mathbf{b}'_k, \qquad (22)$$

where

$$\mathbf{M}'_{kl} = \left(\mathbf{y}_k \cdot \boldsymbol{\Sigma}^{-2} \mathbf{y}_l\right)\left[(\mathbf{A}'_k)^{\mathrm{T}} \mathbf{A}'_l\right], \qquad (23)$$

$$\mathbf{b}'_k = (\mathbf{I}_{\text{meas.}} \mathbf{A}'_k)^{\mathrm{T}} \boldsymbol{\Sigma}^{-2} \mathbf{y}_k. \qquad (24)$$

3.1.2. Extracting scattering curves and error estimates. After fitting a dataset with a *REGALS* model for $\mathbf{Y}$ and $\mathbf{C}$, the results are typically smooth versions of the concentrations and SAXS profiles. However, for further analysis (such as fitting atomistic models to the SAXS data), it is desirable to extract curves resembling experimental data with properly estimated errors. Previously, we applied a projection algorithm which uses the pseudo-inverse of the concentration matrix to generate SAXS profiles and associated error bars (Meisburger *et al.*, 2016). For the datasets examined here, we found that the pseudo-inverse method amplifies noise in certain cases. Therefore, we developed an alternative method which makes use of the regularized basis vectors to overcome this issue. In order to extract a particular component, a residual data matrix is reconstructed by subtracting the model with component k excluded:

$$\mathbf{D}^{(k)}_{\text{resid.}} = \mathbf{I}_{\text{meas.}} - \sum_{j \neq k} \mathbf{y}_j \otimes \mathbf{c}_j. \qquad (25)$$

The unregularized basis functions $\mathbf{y}$ and $\mathbf{c}$ are extracted by minimizing

$$\left\|\boldsymbol{\Sigma}^{-1}\left[\mathbf{D}^{(k)}_{\text{resid.}} - \mathbf{y} \otimes \mathbf{c}\right]\right\|_{\mathrm{F}}^{2}, \qquad (26)$$

with either the scattering profile or the concentration held fixed. The solutions can be written as weighted averages of the residual data matrix, as follows:

$$\mathbf{y}^{(k)}_{\text{extract}} = \mathbf{D}^{(k)}_{\text{resid}} \mathbf{m}_k, \qquad \mathbf{c}^{(k)}_{\text{extract}} = \left[\mathbf{D}^{(k)}_{\text{resid}}\right]^{\mathrm{T}} \mathbf{m}'_k, \qquad (27)$$

with coefficients

$$\mathbf{m}_k = \frac{\mathbf{c}_k}{\mathbf{c}_k \cdot \mathbf{c}_k}, \qquad \mathbf{m}'_k = \frac{\boldsymbol{\Sigma}^{-2} \mathbf{y}_k}{\mathbf{y}_k \cdot \boldsymbol{\Sigma}^{-2} \mathbf{y}_k}. \qquad (28)$$

The uncertainties are estimated by standard propagation of experimental errors:

$$\begin{aligned} \left[\Delta \mathbf{y}^{(k)}_{\text{extract}}\right]_i &= \left[\sum_j \sigma^2_{ij} (\mathbf{m}_k)^2_j\right]^{1/2}, \\ \left[\Delta \mathbf{c}^{(k)}_{\text{extract}}\right]_j &= \left[\sum_i \sigma^2_{ij} (\mathbf{m}'_k)^2_i\right]^{1/2}. \end{aligned} \qquad (29)$$

3.1.3. Regularization parameter estimation. The regularization parameters λ_k and λ'_k reflect prior information about the smoothness of the parameters. They are not known in advance, so initial values must be chosen by the user and further adjusted if *REGALS* fails to converge. However, the regularization parameter is not an intuitive quantity and it depends in a complicated fashion on the noise level in the data and the particular regularizer chosen. To assist the user in selecting initial values, we provide the option of specifying a more intuitive parameter, the 'number of good parameters,' or n_k. This parameter comes from the Bayesian interpretation of regularized linear regression (MacKay, 1992, 1996) and it estimates how many parameters are effectively determined by the data (as opposed to the regularizer). The number of good parameters determined for $\mathbf{u}_k$ (SAXS basis) is as follows:

$$n_k = \sum_j \frac{(\mathbf{d}_k)_j}{(\mathbf{d}_k)_j + \lambda_k}, \qquad (30)$$

where $\mathbf{d}_k$ is the vector of generalized eigenvalues of the matrices $\mathbf{M}_{kk}$ [which depends on $|\mathbf{c}_k|$, see equation (20)] and $\mathbf{H}_k = \mathbf{B}_k^{\mathrm{T}}\mathbf{B}_k$. To determine λ_k given n_k, equation (30) is solved numerically using the initial guess for $\mathbf{c}_k$. Strictly speaking, n_k should be determined using the final value of $\mathbf{c}_k$ (after *REGALS* has converged), but we have found that initial estimates of n_k are usually close to the final values. Similarly, regularization parameters for the concentration basis are found by solving equation (30) where $\mathbf{d}_k$ are the generalized eigenvalues of $\mathbf{M}'_{kk}$ [equation (23)] and $\mathbf{H}'_k = (\mathbf{B}'_k)^{\mathrm{T}}\mathbf{B}'_k$.

3.1.4. Software implementation. The *REGALS* method was developed in MATLAB and subsequently translated into Python. The two implementations have similar organization and produce equivalent results. Both versions are available for the convenience of future users and developers.

The code is organized using a hierarchy of classes to facilitate mixing and matching of basis vector types [Fig. 1(*b*)]. At the lowest level are Concentration and Profile classes for each type (simple, smooth and real-space), which share a common interface and are responsible for calculating the $\mathbf{A}_k$ and $\mathbf{B}_k$ matrices [equations (13) and (16)] given parameters such as boundary conditions, number of samples and extent. At an intermediate level is the Component class, which represents a single component in the mixture and contains one Concentration object and one Profile object. At the top level is the Mixture class, which contains an array of Component objects as well as the parameter vectors and regularization parameters. Methods are included to compute the terms appearing in the normal equations [equations (19) and (22)], estimate regularization parameters [inversion of equation (30)] and extract basis vectors [equations (27) and (29)]. Finally, the *REGALS* class implements alternating least squares and it includes a high-level method (`REGALS.run`) that controls flow through the algorithm with user-specified termination conditions.

The process of setting up, running and analyzing a *REGALS* calculation is performed by writing scripts to interact with the objects. We have included example scripts in the form of live notebooks (*Jupyter* notebooks in Python) for each of the application examples presented here. Source code, documentation and examples are available at https://github.com/ando-lab/regals. The release associated with this publication has been tagged as Version 1.0.

3.2. Example data

3.2.1. AEX-SAXS of *Bs*RNR large subunit. The collection and preprocessing of AEX-SAXS from the large subunit of *B. subtilis* ribonucleotide reductase (*Bs*RNR) were described in the original publication (Parker *et al.*, 2018). Briefly, the as-isolated protein was eluted from a MonoQ column using a linear gradient of 100 to 500 m*M* NaCl directly into a SAXS flow cell. Scattering images were recorded continously during elution (q range of 0.008 to 0.700 Å^{-1}). After integration, each profile was normalized by the transmitted beam intensity, and buffer-only curves collected before the start of the gradient were averaged and subtracted from the remaining curves. A set of 1737 frames was retained for further analysis, beginning just after the start of the linear gradient and ending before the gradient completed, when the NaCl concentration had reached approximately 400 m*M*. These preprocessed data are available in `NrdE_mix_AEX.mat` (a MATLAB-formatted HDF5 file).

3.2.2. Equilibrium titration of PheH. A SAXS titration of phenylalanine hydroxylase (PheH) with phenylalanine (L-phe) was performed previously (Meisburger *et al.*, 2016). In the original publication, SAXS curves from PheH at 25 µ*M* (monomer concentration) were processed to produce 16 background-subtracted scattering curves, each with a different amount of L-phe (0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 1, 3, 6, 10, 20, 40 and 80 m*M*). The same amount of L-phe was present in the buffer-only samples used for subtraction. The q range was 0.01 to 0.96 Å^{-1} and the scattering was normalized by the transmitted beam intensity. These preprocessed data are available in `PheH_titration.mat` (a MATLAB-formatted HDF5 file).

3.2.3. Time-resolved mixing of MsbA NBD with ATP. As a first example of time-resolved SAXS data, we chose a recently published stopped-flow mixing dataset (Josts *et al.*, 2020). In the experiment, a soluble nucleotide binding domain (NBD) construct [residues 330–581 of the adenosine triphosphate (ATP)-binding cassette transporter MsbA] was mixed with Mg^{2+}-ATP in a 1:1 (*v*:*v*) ratio (final concentrations 500 µ*M* NBD and 450 µ*M* ATP). One X-ray exposure of 35 ms was acquired per shot after a variable time delay of 20 ms to 120 s.

The time-resolved MsbA NBD dataset consisting of 23 buffer-subtracted scattering curves ($0.01 < q < 0.5$ Å^{-1}) was downloaded from a public database (the Small-Angle Scattering Biological Data Bank, https://www.sasbdb.org/data/SASDGV5/), minimally reformatted and saved as `MsbA_time_resolved.mat` (a MATLAB-formatted HDF5 file). Minor preprocessing was performed before running *REGALS*. Upon inspection, we noted a strong negative-going feature at low q, suggesting a background subtraction error. We therefore truncated the low q at 0.015 Å^{-1}. In addition, we found that the average intensity displayed a slight random jitter shot to shot. We corrected for this by applying a scale factor to each curve, which was found by fitting a fifth-order polynomial to the mean intensity versus $\log_{10}$(time). The resulting scale factors were close to 1 (standard deviation of 0.013).

3.2.4. Time-resolved temperature jump of CypA. As an example of a pump–probe time-resolved experiment, we chose recently reported temerature-jump (T-jump) SAXS/wide-angle X-ray scattering (WAXS) data collected on the *cis*-proline isomerase CypA (Thompson *et al.*, 2019). Here we analyzed one particular set of experiments corresponding to the wild-type CypA protein and buffer blanks following a T-jump to $29.9 \pm 0.1^\circ$C. After downloading the raw T-jump data (Fraser *et al.*, 2019), we repeated the published data-reduction protocol (Thompson *et al.*, 2019) using a custom MATLAB script (available upon request). Briefly, difference scattering curves ($\Delta I = I_{\rm on} - I_{\rm off}$) were calculated for both the protein and buffer blanks, and a series of scaling operations was performed to correct for shot-to-shot variations, most crucially the scaling of buffer difference profiles in order to minimize $\Delta I_{\rm protein} - \Delta I_{\rm buffer}$ in the WAXS regime, where solvent scattering predominates. This produced a set of 28 difference profiles: 27 logarithmically distributed time points after the T-jump, and one control where the laser was off prior to X-ray exposure. The control profile was close to zero, indicating that the data processing had not introduced errors and the remaining profiles resembled those reported in the original publication. After initial data reduction, the WAXS data were discarded and the SAXS portion of the curves ($0.025 \le q < 1$ Å^{-1}) was saved as `CypA_Tjump.mat` (a MATLAB-formatted HDF5 file).

4. Results and discussion

4.1. *REGALS* deconvolution of AEX-SAXS data

During an AEX separation, sample bound to the column is eluted by flowing buffer with increasing salt concentrations. The main challenge in deconvolving AEX-SAXS data is to account for the changing background scattering from the buffer. As described in the *Methods* section, we analyzed a dataset previously reported for the large subunit of *Bs*RNR (Parker *et al.*, 2018), which eluted from the column in two main peaks during a linear gradient of 100 to 400 m*M* NaCl [Fig. 2(*a*)]. The salt gradient produced a rising background intensity during elution, seen clearly in a plot of the total intensity per frame [Fig. 2(*b*), top panel].

First, we performed SVD to estimate the number of scattering components associated with the protein and background signals. SVD of the entire dataset yields four significant singular values [Fig. 2(*b*), bottom panel, black circles]. To determine which of these four correspond to buffer versus protein, we repeated SVD on a truncated dataset consisting of the first 700 frames, collected before the protein elution [Fig. 2(*b*), top panel, blue region]. Interestingly, this region alone produces two significant singular values [Fig. 2(*b*), bottom panel, blue circles], suggesting that two components are needed to describe the background and that the remaining two correspond to protein. Inspection of the basis vectors obtained from SVD of the full dataset [equation (7)] further confirms this assignment. A rising background signal is present in two of the concentration profiles [Fig. 2(*c*), left panel, orange and blue curves]. However, it is also evident in the concentration profiles that protein peaks appear in all four components, mixing with the background in two cases, and that the corresponding SAXS profiles [Fig. 2(*c*), right panel] are similarly non-physical, containing negative intensities. The

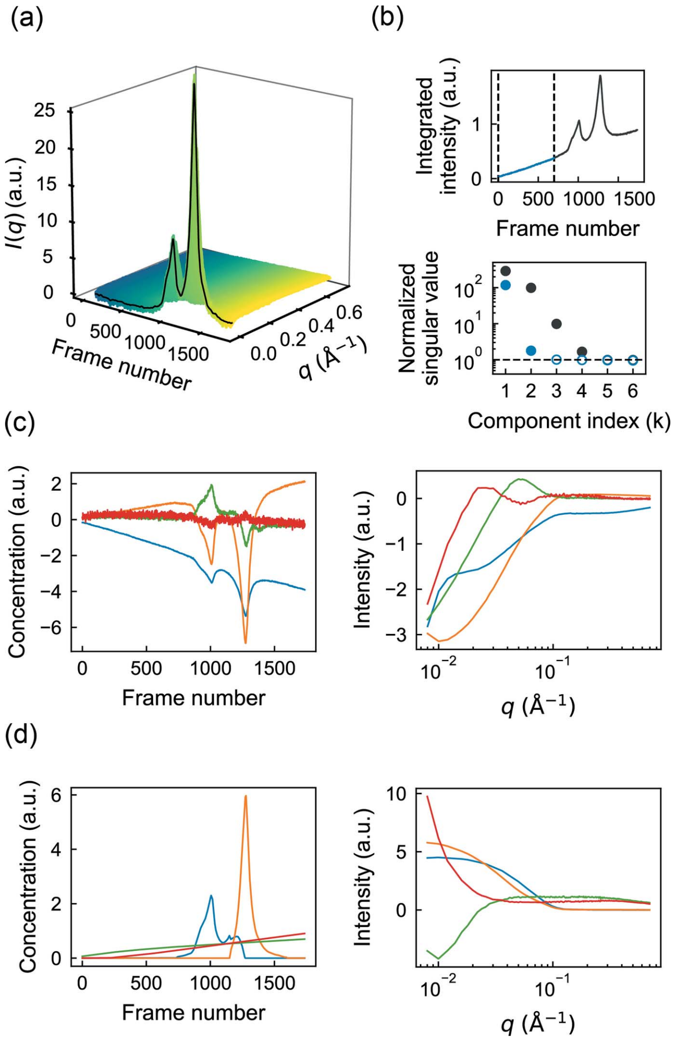


Figure 2
REGALS deconvolution of an AEX-SAXS dataset with a changing background. (*a*) The scattering intensities obtained in a previously reported AEX-SAXS experiment on the large subunit of *Bs*RNR (Parker *et al.*, 2018) plotted as a function of frame number. (*b*) In the top panel, the integrated intensities across the elution display two prominent peaks over a rising background. SVD of the full dataset shows four significant singular values above the noise level [gray circles above the dashed line in the bottom panel, see equation (7)]. SVD of only those scattering profiles prior to the protein peaks (blue region between dashed lines in the top panel) shows two significant singular values (blue filled circles in the bottom panel), indicating the presence of two background components. (*c*) The deconvolution derived from SVD of the full dataset [equation (7)] is non-physical. On the left are the concentration profiles (right singular vectors) and on the right the corresponding scattering profiles (left singular vectors). (*d*) *REGALS* gives physical concentration profiles (left panel) and scattering profiles (right panel).

fact that many of the concentration and SAXS profiles oscillate around zero is expected given the orthonormality restraint imposed on the SVD basis vectors. Thus, although SVD provides useful information on the number and types of scattering components, different restraints are needed in the deconvolution process to obtain a physically meaningful interpretation.

With initial insight from SVD, we next constructed a Mixture model [Fig. 1(*b*)] that takes into account basic expectations about the data. The simplest assumption is that each peak in the chromatogram corresponds to a different protein component and that the concentrations of the background components should evolve smoothly over the course of elution. Each protein component (C1 and C2) was thus parameterized using a smooth concentration basis vector with a region of support encompassing each peak. In order to arrive at a unique deconvolution, the two background components must be differentiated in some way within the model. Because SVD revealed that one of the background-containing components is close to zero for the first ∼200 frames [Fig. 2(*c*), left panel, orange curve], we modeled one of the background components (B1) to span the full range of frames, while the other (B2) had a region of support beginning at frame 200 with a zero boundary condition there. We implemented this model again using smooth concentration basis vectors for B1 and B2.

Finally, we refined the model using regularization to enforce smoothness of the background components. The SAXS profiles were not parameterized (simple basis vectors were used). To ensure that each protein concentration model fully encompassed the peak for each component but was not larger than necessary, we performed several trial refinements with *REGALS* while varying the region of support and inspecting a plot of residual χ^2 versus frame number (not shown). The model parameters are summarized in Table S1 in the supporting information. Finally, *REGALS* was run for 50 iterations, at which point it was well converged. The overall reduced χ^2 was 1.011, suggesting that the refined model accounted for most of the signal.

The results obtained by *REGALS* are shown in Fig. 2(*d*). The concentrations of the background components (B1 and B2) rise in an approximately linear fashion during elution [Fig. 2(*d*), left panel, green and red], reflecting the influence of the smoothness regularizer on these components. The corresponding SAXS profiles show that B1 is associated with an increase in scattering at high q, while B2 is primarily a low-q feature [Fig. 2(*d*), right panel, green versus red]. The protein components (C1 and C2) have compact peaks with positive concentrations and corresponding SAXS profiles that appear well subtracted [Fig. 2(*d*), blue and orange]. We previously showed that these protein components were in excellent agreement with models of the monomeric and dimeric forms derived from crystal structures (Parker *et al.*, 2018).

Although SVD had suggested that two background components were needed to describe the data, we wondered whether two components were strictly necessary for deconvolving the protein peaks. To test this, we removed the minor component from the model (B2) and performed the deconvolution using *REGALS*. As expected, the quality of fit was noticeably worse when the background was modeled with one component compared with two [Fig. 3(*a*), bottom versus top]. Interestingly, the fit of the one-background model is worse in the buffer-only region of the data, but it achieves a near-perfect fit ($\chi^2 \simeq 1$) in the region where the proteins elute. This observation suggests that the protein components have absorbed the background subtraction error. Indeed, a comparison of the extracted SAXS profiles for C1 shows a significant deviation from expected shape in the low-q region if only one background is used [Fig. 3(*b*)]. These results indicate that the buffer scattering in AEX-SAXS can be complex and must be modeled well to achieve well subtracted SAXS profiles. Furthermore, they underscore the importance of collecting the full buffer scattering before and after the peak in AEX-SAXS experiments, as this information is effectively used to extrapolate the complex behavior underneath the elution peaks.

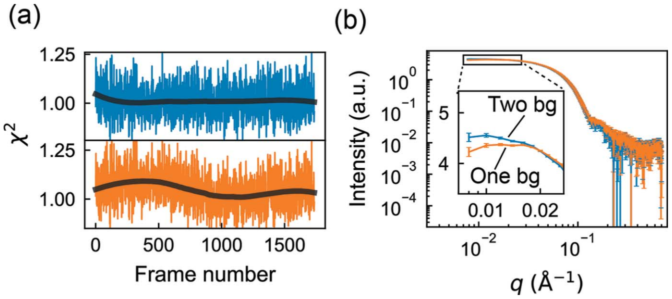


Figure 3
The changing background in AEX-SAXS can be complex. (*a*) Comparison of χ^2 values from the *REGALS* deconvolution of the AEX-SAXS dataset in Fig. 2 with two background components (top panel, blue) versus one (bottom panel, orange). The former is relatively uniform around the expected value of 1, whereas the latter shows unevenness throughout the elution (black curves are the smoothed χ^2 values shown as trend lines), indicating that dataset is better described with two background components. (*b*) Because the background scattering includes significant low-q features, failing to take proper account of the changes in the background can lead to artifacts in the extracted protein scattering profiles. Here, the use of only one background component in the analysis leads to a downturn in the low-q region of the scattering from component 1, which will lead to an underestimation of the protein size.

4.2. *REGALS* deconvolution with real-space SAXS restraints

In SAXS datasets from time-resolved or ligand titration experiments, it is common for components to have non-zero concentrations in most or all of the measurements, and a compact support cannot be assumed in the concentration basis as in the AEX-SAXS example above. To deconvolve such datasets robustly, it is necessary to incorporate additional prior information. Within *REGALS*, this can be done in two ways: (i) by imposing boundary conditions on the concentration basis vectors and (ii) by limiting the maximum dimension of certain components through real-space parameterization of the SAXS basis vectors.

4.2.1. Equilibrium titrations. As a first test of real-space restraints, we examined a challenging ligand titration dataset

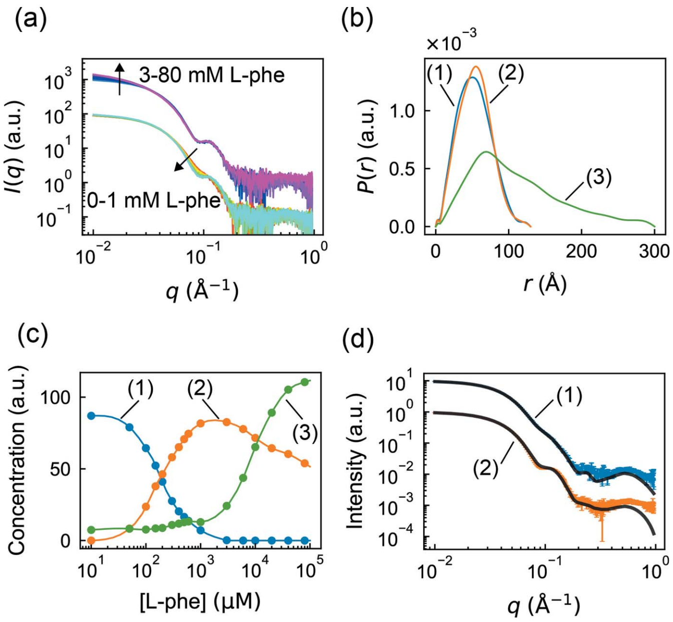


Figure 4

Separation of aggregation from ligand-induced conformational changes in a titration dataset with real-space regularization in *REGALS*. (*a*) Scattering profiles from a previously reported phenylalanine (L-phe) titration experiment on PheH (Meisburger *et al.*, 2016). Up to 1 m*M* L-phe (red to cyan), the change in scattering occurs mainly at mid q, corresponding to internal conformational changes. At [L-phe] greater than 1 m*M* (blue to magenta), an increase at low q can be observed, indicative of aggregation. The two sets of profiles are offset for clarity. (*b*) Regularized $P(r)$ functions from *REGALS* deconvolution. Different cutoffs for $P(r)$ functions differentiate aggregation (green) from normal conformations (blue and orange). (*c*) Concentration profiles from *REGALS* deconvolution (continuous curves) are consistent with observations from panel (*a*), with conformational switching occurring below 1 m*M* L-phe and aggregation gradually becoming dominant above 1 m*M* L-phe. Circles show unregularized concentrations [equation (27)]. (*d*) Extracted profiles for components (1) and (2) agree with the scattering profiles of inactive and activated PheH, respectively. Here, the black curves are the $P(r)$ regularized scattering profiles from SEC-SAXS (Meisburger *et al.*, 2016).

of phenylalanine hydroxylase (PheH) (Meisburger *et al.*, 2016). The tetrameric enzyme undergoes a conformational change upon binding its allosterically activating ligand, L-phenyalanine (L-phe). In SAXS, the signature of this conformational change is an oscillating mid-q feature that appears at physiological concentrations of L-phe [Fig. 4(*a*), 0–1 m*M* L-phe]. At higher concentrations of ligand, the mid-q feature does not change further, but an increase in scattering at low q is observed [Fig. 4(*a*), 3–80 m*M* L-phe], indicating an increase in the average molecular weight. This larger oligomer or aggregate is likely to be non-physiological, and therefore previous analysis focused on the 0–1 m*M* concentration range. However, SEC-SAXS experiments at 0 and 1 m*M* L-phe revealed the presence of a small amount of aggregate, indicating that all of the SAXS curves in the titration were corrupted by aggregation to some extent, inflating estimates of size and molecular weight even at low L-phe concentrations. The presence of a small population of aggregates is extremely common in SAXS, and thus a direct method to deconvolve it from other components is of particular interest.

To deconvolve the PheH titration dataset, we constructed a *REGALS* model with three components: resting tetramer, activated tetramer and aggregate, numbered (1)–(3), respectively. The SAXS profiles for each component were modeled using the real-space parameterization [$P(r)$]. The resting and activated tetramers were estimated to have a maximum dimension of 130 Å based on previous studies (Meisburger *et al.*, 2016), and the aggregate was assigned a maximum dimension of 300 Å, the largest dimension that could be measured based on the Shannon limit for this dataset ($d_{\max} < \pi/q_{\min}$). Boundary conditions of $P(r) = 0$ were imposed at both $r = 0$ and $r = d_{\max}$. We also imposed prior information on the concentration basis vectors using a smooth parameterization. According to the equilibrium model for this system (Meisburger *et al.*, 2016), the concentration of activated tetramer is negligible at 0 m*M* L-phe, so a zero boundary condition was imposed. For the resting tetramer, we limited the range of the basis vector to 0–3 m*M* and imposed a zero boundary condition at 3 m*M* based on the observation that the mid-q feature saturates above this concentration. No limits or boundary conditions were imposed on the aggregate concentration. The independent variable x was calculated as the logarithm of [L-phe], reflecting the higher density of samples at low [L-phe] and the standard practice of visualizing titration data on a logarithmic scale. Regularization was used to enforce smoothness of the concentration profiles and the $P(r)$ functions. The model parameters are summarized in Table S2. The basis vectors were optimized using the *REGALS* algorithm, which converged after 50 iterations with an overall reduced χ^2 of 1.41.

One advantage of using the real-space parameterization is that $P(r)$ functions are obtained directly from the deconvolution and provide immediate insight into particle shape. For the resting and activated PheH tetramers, we find that the $P(r)$ functions decay to zero smoothly at $d_{\max}$ [Fig. 4(*b*), components (1) and (2)], as expected for compact particles. The peak in $P(r)$ shifts toward larger dimensions in the activated tetramer, indicating that it has a less compact conformation. The $P(r)$ for the aggregated species decays toward zero at $d_{\max}$ in an approximately linear fashion, which is characteristic of elongated or rod-like shapes [Fig. 4(*b*), component (3)].

In ligand titration datasets like this one, the concentrations of different components are often of great interest, since they give insight into the equilibrium behavior of the system, including cooperativity and binding affinities. The *REGALS* deconvolution of the PheH titration produced concentration profiles that appear physically reasonable [Fig. 4(*c*), continuous curves]. To verify that smoothness regularization had not overly biased the result, we also extracted concentration estimates at each point without regularization [equation (27)] and found that they agree with the regularized curve [Fig. 4(*c*), circles]. We find that the aggregate is present under all conditions, staying at a low level between 0 and 1 m*M* L-phe, before rising sharply at high concentrations, in agreement with prior SEC-SAXS experiments (Meisburger *et al.*, 2016). The resting tetramer converts into the activated tetramer in a manner characteristic of cooperative two-state transition, as shown previously (Meisburger *et al.*, 2016).

Further analysis of this equilibrium is beyond the scope of this study. However, we note that the arbitrary concentration scale of Fig. 4(*c*) can be transformed readily into the fraction of resting and activated species, which can be fitted using an equilibrium model. The *REGALS* results are normalized by the area under $P(r)$ [equal to $I(q \rightarrow 0)$] and this quantity is expected to be the same for components with the same molecular weight. Thus, in this case the tetramer concentrations in Fig. 4(*c*) differ from the true concentrations (*e.g.* in mg ml^{-1}) by the same scale factor.

One assumption in the *REGALS* model was that the aggregate did not change in size or shape as a function of [L-phe], which may not be the case, particularly since its shape appears to be rod-like and therefore its growth might be nonterminating. To check whether this assumption was supported by the data, we examined the reduced χ^2 of the model at each L-phe concentration. Interestingly, χ^2 at the highest [L-phe] is 3.4, which is significantly larger than at other concentrations (1.3 on average). Thus, it seems likely that the scattering profile of the aggregate does change, at least at very high L-phe concentrations. If this is the case and the aggregate is improperly modeled by *REGALS*, another technique such as SEC-SAXS might be necessary to obtain reliable scattering curves for the tetramers. Nonetheless, the tetramer SAXS curves extracted from the *REGALS* deconvolution are in excellent agreement with those obtained by SEC-SAXS [Fig. 4(*d*)], suggesting that inaccuracy of the aggregation model had a minimal effect on deconvolution.

4.2.2. Time-resolved SAXS. Based on the successful application of real-space *REGALS* to the challenging PheH titration dataset, we considered whether similar models might be applied to time-resolved SAXS.

Time-resolved experiments can be performed with two different techniques: mixing and pump–probe. In mixing experiments, a rapid change in solution conditions (such as by rapid dilution, or addition of denaturant, allosteric ligand or reactant) is followed by SAXS measurements after some time has elapsed. This technique is well suited to irreversible reactions or those that cannot be initiated except by mixing. In contrast, pump–probe experiments are usually initiated by a laser pulse and followed, after a time delay, by the X-ray measurement. Compared with mixing, pump–probe measurements can access very short time scales if fast lasers and pulsed X-ray sources are used. Pump–probe datasets are also special in that very small changes can be measured by examining difference profiles (laser on minus laser off), which removes systematic error. Given these differences, we chose to evaluate *REGALS* with both mixing and pump–probe datasets, as described below.

First, we chose to analyze a stopped-flow mixing dataset from the soluble nucleotide binding domains (NBDs) of the membrane transport protein MsbA, which was recently published and deposited in a public database (Josts *et al.*, 2020). In the experiment, a solution with nucleotide-free NBD monomers was rapidly mixed with ATP, resulting in ligand binding and dimerization, followed by ATP hydrolysis and dissociation back to the monomeric state. This transient increase in average size can be observed in a Kratky plot [$q^2I(q)$ versus q], where the main peak shifts to the left (lower q) and then to the right [Fig. 5(*a*)]. In the original publication, the relative concentrations of NBD monomer and dimer at each time point were fitted using calculated scattering profiles from known crystal structures. However, in time-resolved experiments generally, it is often the case that atomically detailed structures are not available, either because they have not been characterized at high resolution or because they are dynamic. Therefore, we considered whether *REGALS* could deconvolve the MsbA dataset using only general properties of the molecules.

The *REGALS* model consisted of two components representing NBD monomer and dimer. The parameterization was similar to the PheH titration example above: the smooth parameterization was used for concentrations and the real-space one for SAXS profiles. Based on the full-length structure of dimeric MsbA [Protein Data Bank (PDB) ID 3b60], we estimated the $d_{\rm max}$ of the dimeric and monomeric forms of the NBD portions to be 70 and 62 Å, respectively. Reflecting the prior observation that NBDs relax to a fully dimeric state after ATP hydrolysis, we applied a zero boundary condition to the

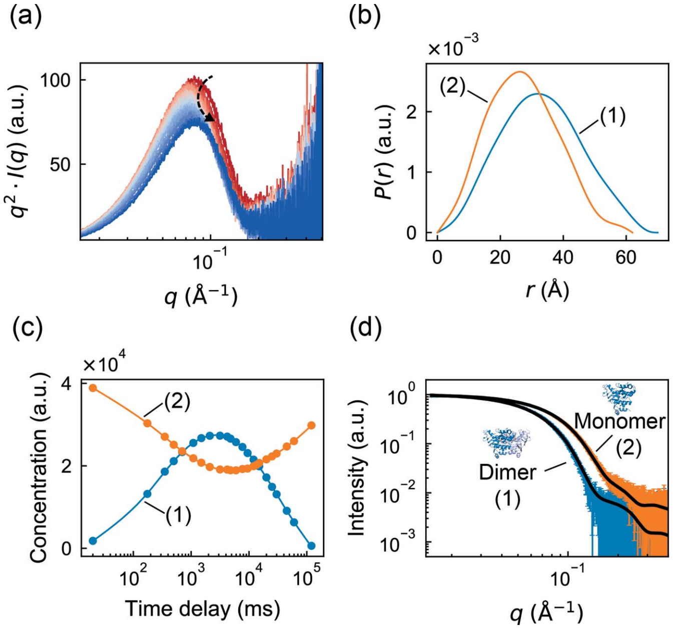


Figure 5
Model-free deconvolution of a time-resolved mixing dataset in *REGALS*. (*a*) Scattering profiles from a previously reported time-resolved mixing experiment with MsbA NBD and ATP (Josts *et al.*, 2020) shown as Kratky plots (red to blue). The peak position shifts to $q \sim 0.07$ Å^{-1} before returning to $q \sim 0.08$ Å^{-1} (denoted by the curved arrow), indicating a transient increase in size. (*b*) Regularized $P(r)$ functions of dimer (blue) and monomer (orange) components have well defined shapes with $d_{\rm max}$ estimates based on the crystal structure of full-length dimeric MsbA (PDB ID 3b60). (*c*) Concentration profiles from *REGALS* deconvolution (continuous curves) and unregularized concentrations (circles), showing transient formation of the NBD dimer. (*d*) The extracted scattering profiles of components (1) and (2) agree with predictions using the NBD dimer and monomer from the full-length crystal structure [black curves are *CRYSOL* (Svergun *et al.*, 1995) fits].

dimer concentration at the final time point (approximately 2 min after mixing). The model parameters are summarized in Table S3. *REGALS* was run for 100 iterations, resulting in an overall reduced χ^2 of 0.335. The fact that $\chi^2 < 1$ here suggests that the reported experimental errors were overestimated, so χ^2 is not a reliable statistic for quality of fit. However, the quality of fit was confirmed by examining the residual (not shown).

The deconvolved concentration profiles show a rise and fall of the dimer component, with a concomitant dip in the monomer, which resembles the profiles obtained by fitting scattering from crystal structure models in the original publication (Josts *et al.*, 2020). The $P(r)$ functions are also physically reasonable, with single peaks that decay smoothly to zero as r approaches the maximum dimension. Using the *REGALS* deconvolution, we extracted the SAXS profiles [equation (27) in *Methods*] for the monomer and dimer and compared them with models derived from crystallography [Fig. 5(*d*)]. The excellent agreement suggests that the atomistic models accurately reflect the structures of the NBDs in solution. Although the analysis presented here used estimates for the maximum dimension based on a crystal structure, no assumptions were made about the shape of the individual components.

For a pump–probe dataset, we chose a temperature-jump (T-jump) SAXS/WAXS experiment which was performed on the protein CypA (Thompson *et al.*, 2019). These experiments involved rapidly heating the sample by approximately 10°C with an infrared laser pulse of several nanoseconds duration, followed by a synchrotron X-ray pulse of approximately 500 ns duration after a delay of 562 ns to 1 ms. Following the methods in the original publication (Thompson *et al.*, 2019), difference profiles were constructed (laser on minus laser off) for both the protein and buffer blanks, and these were scaled together in the WAXS regime and subtracted. The remaining signal, attributed to the effect of the rapid temperature change, is most significant in the SAXS regime [Fig. 6(*a*)], and it evolves non-trivially as a function of the time delay [Fig. 6(*a*), inset]. Note that the difference profiles are negative and this is thought to result from the differential thermal expansion coefficients of protein and water, which would reduce the scattering contrast at high temperature (Thompson *et al.*, 2019).

Previously, the biphasic appearance of the mean intensity [Fig. 6(*a*), inset] was interpreted as a fast transition to excited states of the molecule, followed by a slow relaxation toward equilibrium (Thompson *et al.*, 2019). Although SVD analysis revealed three significant components, no kinetic or structural interpretation of the basis vectors was reported. We wondered whether a real-space *REGALS* deconvolution might offer additional insight. Based on the SVD result, we chose to model three components (C1, C2 and C3). For all three, a smooth parameterization was used for the concentration basis and a real-space one for the SAXS profile basis. The first component (C1) was assigned to represent the transient process following the T-jump, with a concentration of zero at both end points. No constraints were applied to the concentrations of the other two components. In real space, C2 was assigned a maximum dimension of 46 Å estimated from a crystal structure of CypA (PDB ID 3k0n). Lacking further information with which to restrain the model, the maximum dimensions for C1 and C3 were adjusted by trial and error based on the quality of fit and subjective appearance of the $P(r)$ functions. The final model parameters are summarized in Table S4. Note that, since difference intensities are fitted, this parameterization represents the difference $P(r)$ function, $\Delta P = P_{\rm on} - P_{\rm off}$, and $d_{\rm max}$ represents the maximum dimension over which changes to $P(r)$ occur after heating.

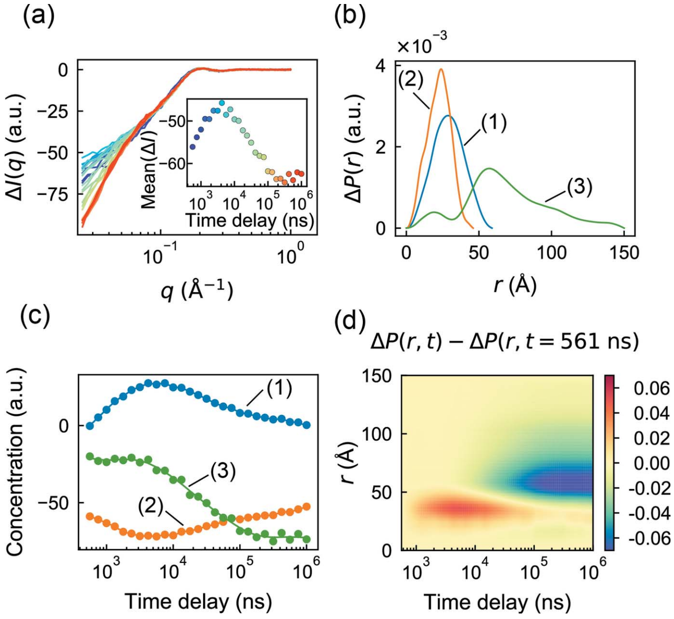


Figure 6
Separation of changes at different length scales in a time-resolved T-jump dataset in *REGALS*. (*a*) Difference scattering from a previously reported time-resolved T-jump experiment on CypA (Thompson *et al.*, 2019) as a function of time delay (violet to red). (Inset) The mean intensity ΔI over q = 0.03–0.05 Å^{-1} increases before decreasing. (*b*) Regularized $\Delta P(r)$ functions from *REGALS* deconvolution. Three cut-offs were chosen to separate changes at different length scales: the equilibrium CypA structure (49 Å, orange), the thermally excited intermediate (59 Å, blue) and the large length-scale change (150 Å, green). (*c*) Concentration profiles from *REGALS* deconvolution (continuous curves) and unregularized concentrations (circles), showing different kinetics for conformational changes at different length scales. (*d*) Reconstructed $\Delta P(r, t)$ from deconvolved components, displaying two distinct processes occurring at small and large length scales.

The *REGALS* algorithm was run for 400 iterations, converging to an overall reduced χ^2 of 1.667. Although the difference intensities are negative [Fig. 6(*a*)], the deconvolved $\Delta P(r)$ functions are all positive [Fig. 6(*b*)] because the *REGALS* algorithm normalizes SAXS basis functions by the integral of $P(r)$. Consequently, some of the concentrations are negative [Fig. 6(*c*)]. Negative concentrations (or SAXS curves) are a necessary feature when analyzing difference intensities, and they can be a challenge to conceptualize. However, two immediate observations can be made. First, the concentration of C3 is approximately constant for the first ~4 µs after the T-jump and the change on those timescales is captured by C1 and C2. According to the $\Delta P(r)$ functions for C1 and C2, we conclude that the fast processes occur on length scales up to ~60 Å, somewhat larger than the size of the CypA monomer. Changes on longer timescales additionally involve C3, which has a much longer range of 150 Å, and probably involve interparticle interactions because the experiments were done at a relatively high protein concentration of 50 mg ml^{-1}.

To gain a more intuitive picture of the changes following the T-jump, we used the regularized basis functions to reconstruct the time evolution of $\Delta P(r)$. This removes, to some extent, the influence of choices made during the *REGALS* parameterization and resolves the sign ambiguity. Since the signal is dominated by the contrast decrease (not shown), we subtracted the first time point to obtain $\Delta\Delta P(r, t) \equiv \Delta P(r, t) - \Delta P(r, t = 561\ \text{ns})$, which tracks the change in signal after the T-jump [Fig. 6(*d*)]. This reconstructed signal reveals a clear positive feature with a peak at $r \simeq 35$ Å that appears at fast time scales, followed by a negative feature with a peak at $r \simeq 60$ Å on slower time scales. The physical explanation is not entirely clear, but one hypothesis might be transient partial unfolding followed by an increase in inter-particle repulsion (or a decrease in attraction). As experiments which rely on difference intensities are often performed at high protein concentrations, further investigations of inter-particle interactions are of great interest.

5. Conclusions

In this work, we have introduced *REGALS* as a robust and generally applicable technique to deconvolve challenging SAXS datasets from evolving mixtures. The strategy implemented in *REGALS* has several key advantages. Most notably, prior knowledge is taken into account without having to impose a physicochemical ‘hard’ model or known scattering curves. Having flexible restraints is important in cases where such models are not available, or when SAXS is to be used for cross validation. Second, the method is readily adapted to a range of experiments. As we have demonstrated, AEX-SAXS, ligand titrations, time-resolved mixing and time-resolved pump–probe datasets can all be analyzed successfully by *REGALS*. Finally, *REGALS* is not a black box; the model assumptions are physically motivated, easily explained and completely specified by the user. Because deconvolution can be ambiguous and strongly influenced by model assumptions, this transparency is essential when communicating scientific results.

The flexibility of the *REGALS* method is reflected in our software implementation (see *Methods*). The model is specified using object-oriented code, which facilitates mixing and matching parameterizations to suit the experiment. In order to provide feedback to the user and support customization, the code is run using a live notebook that performs data import, model definition, optimization and visualization. Example notebooks are provided for each of the datasets described here. Since SAXS is a rapidly developing technique, we designed *REGALS* with future changes in mind. Its hierarchical object structure allows for new linear parameterizations and quadratic regularizers to be added with minimal changes to the existing code. Finally, to facilitate future development and adoption by the community, we have provided two functionally equivalent implementations of *REGALS* in MATLAB and Python. The code is version-controlled, open source and free to use.

Future work will focus on augmenting the *REGALS* toolkit to expand the range of applications further. Here, we found that two simple restraints, smoothness and compact support, proved powerful for expressing prior knowledge. However, many other types of restraint are possible within the *REGALS* framework. Examples of particular interest to SAXS include sparseness and non-negativity (Cichocki & Zdunek, 2007), hard restraints on certain components with known scattering curves, and fixed non-zero or derivative boundary conditions. In addition, *REGALS* could be applied to datasets with more than one independent variable using methods from MCR of multi-way data (de Juan & Tauler, 2003). For example, the CypA time series analyzed here was one among several conducted at different initial temperatures (Thompson *et al.*, 2019), and thus the entire dataset might be analyzed using a multi-way *REGALS* decomposition. Furthermore, the assumption of a dilute solution can also be relaxed by adding extra components to represent terms in the Taylor expansion of the structure factor (Lipfert *et al.*, 2007), which may be of particular interest for time-resolved experiments that require high protein concentrations. Finally, certain parameter choices in *REGALS* may be automated by leveraging the Bayesian interpretation of regularized linear regression (MacKay, 1992, 1996), much as the regularization parameter and $d_{\max}$ are determined automatically in Bayesian IFT (Hansen, 2012). We anticipate that the *REGALS* method described here, and future developments, will be a valuable addition to the SAXS data analysis toolset and enable new applications.

Acknowledgements

We thank M. B. Watkins and W. C. Thomas for critical reading of the paper. The authors declare no conflicts of interest.

Funding information

Funding for this research was provided by: National Science Foundation (grant No. MCB-1942668 to Nozomi Ando); National Institutes of Health, National Institute of General Medical Sciences (grant No. GM124847 to Nozomi Ando); Cornell University (start-up funds awarded to Nozomi Ando).

References

Akiyama, S., Takahashi, S., Kimura, T., Ishimori, K., Morishima, I., Nishikawa, Y. & Fujisawa, T. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 1329–1334.

Ayuso-Tejedor, S., García-Fandiño, R., Orozco, M., Sancho, J. & Bernadó, P. (2011). *J. Mol. Biol.* **406**, 604–619.

Blobel, J., Bernadó, P., Svergun, D. I., Tauler, R. & Pons, M. (2009). *J. Am. Chem. Soc.* **131**, 4378–4386.

Brosey, C. A. & Tainer, J. A. (2019). *Curr. Opin. Struct. Biol.* **58**, 197–213.

Chen, L., Hodgson, K. O. & Doniach, S. (1996). *J. Mol. Biol.* **261**, 658–671.

Chen, L., Wildegger, G., Kiefhaber, T., Hodgson, K. O. & Doniach, S. (1998). *J. Mol. Biol.* **276**, 225–237.

Cho, H. S., Dashdorj, N., Schotte, F., Graber, T., Henning, R. & Anfinrud, P. (2010). *Proc. Natl Acad. Sci. USA*, **107**, 7281–7286.

Cichocki, A. & Zdunek, R. (2007). *Advances in Neural Networks – ISNN 2007. Proceedings of the 4th International Symposium on Neural Networks*, 3–7 June 2007, Nanjing, China, edited by D. Liu, S. Fei, Z. Hou, H. Zhang & C. Sun, Part III, pp. 793–802. Heidelberg: Springer.

Fraser, J., Anfinrud, P. & Thompson, M. (2019). *X-ray Scattering Curves (SAXS/WAXS) Used for the Analysis Described in 'Temperature-Jump Solution X-ray Scattering Reveals Distinct Motions in a Dynamic Enzyme'*. https://doi.org/10.35092/yhjc.9177143.v1.

Hansen, S. (2012). *Bayesian Methods in Structural Bioinformatics*, edited by T. Hamelryck, K. Mardia & J. Ferkinghoff-Borg, pp. 313–342. Heidelberg: Springer.

Hansen, S. & Pedersen, J. S. (1991). *J. Appl. Cryst.* **24**, 541–548.

Hendler, R. W. & Shrager, R. I. (1994). *J. Biochem. Biophys. Methods*, **28**, 1–33.

Henry, E. & Hofrichter, J. (1992). *Numerical Computer Methods. Methods in Enzymology*, Vol. 210, pp. 129–192. New York: Academic Press.

Herranz-Trillo, F., Groenning, M., van Maarschalkerweerd, A., Tauler, R., Vestergaard, B. & Bernadó, P. (2017). *Structure*, **25**, 5–15.

Hopkins, J. B., Gillilan, R. E. & Skou, S. (2017). *J. Appl. Cryst.* **50**, 1545–1553.

Hutin, S., Brennich, M., Maillot, B. & Round, A. (2016). *Acta Cryst.* D**72**, 1090–1099.

Jaumot, J., de Juan, A. & Tauler, R. (2015). *Chemom. Intell. Lab. Syst.* **140**, 1–12.

Jaumot, J., Vives, M. & Gargallo, R. (2004). *Anal. Biochem.* **327**, 1–13.

Josts, I., Gao, Y., Monteiro, D. C. F., Niebling, S., Nitsche, J., Veith, K., Gräwert, T. W., Blanchet, C. E., Schroer, M. A., Huse, N., Pearson, A. R., Svergun, D. I. & Tidow, H. (2020). *Structure*, **28**, 348–354.e3.

Juan, A. de & Tauler, R. (2003). *Anal. Chim. Acta*, **500**, 195–210.

Kathuria, S. V., Guo, L., Graceffa, R., Barrea, R., Nobrega, R. P., Matthews, C. R., Irving, T. C. & Bilsel, O. (2011). *Biopolymers*, **95**, 550–558.

Kirby, N. M. & Cowieson, N. P. (2014). *Curr. Opin. Struct. Biol.* **28**, 41–46.

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 1277–1282.

Lipfert, J., Columbus, L., Chu, V. B. & Doniach, S. (2007). *J. Appl. Cryst.* **40**, s235–s239.

MacKay, D. J. C. (1992). *Neural Comput.* **4**, 415–447.

MacKay, D. J. C. (1996). *Models of Neural Networks III*, pp. 211–254. Heidelberg: Springer.

Maeder, M. (1987). *Anal. Chem.* **59**, 527–530.

Meisburger, S. P., Taylor, A. B., Khan, C. A., Zhang, S., Fitzpatrick, P. F. & Ando, N. (2016). *J. Am. Chem. Soc.* **138**, 6506–6516.

Meisburger, S. P., Thomas, W. C., Watkins, M. B. & Ando, N. (2017). *Chem. Rev.* **117**, 7615–7672.

Miller, K. (1970). *SIAM J. Math. Anal.* **1**, 52–74.

Minh, D. L. & Makowski, L. (2013). *Biophys. J.* **104**, 873–883.

Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.

Neutze, R. & Moffat, K. (2012). *Curr. Opin. Struct. Biol.* **22**, 651–659.

Parker, M. J., Maggiolo, A. O., Thomas, W. C., Kim, A., Meisburger, S. P., Ando, N., Boal, A. K. & Stubbe, J. (2018). *Proc. Natl Acad. Sci. USA*, **115**, E4594–E4603.

Pérez, J. & Vachette, P. (2017). *Biological Small Angle Scattering: Techniques, Strategies and Tips*, edited by B. Chaudhuri, I. G. Muñoz, S. Qian & V. S. Urban, pp. 183–199. Singapore: Springer.

Press, W. H. (2007). *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. Cambridge University Press.

Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). *Q. Rev. Biophys.* **40**, 191–285.

Segel, D. J., Fink, A. L., Hodgson, K. O. & Doniach, S. (1998). *Biochemistry*, **37**, 12443–12451.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Thompson, M. C., Barad, B. A., Wolff, A. M., Sun Cho, H., Schotte, F., Schwarz, D. M. C., Anfinrud, P. & Fraser, J. S. (2019). *Nat. Chem.* **11**, 1058–1066.

Tikhonov, A. N. & Arsenin, V. Y. (1977). *Solutions of Ill-posed Problems.* Philadelphia: Society for Industrial and Applied Mathematics.

Vershynin, R. (2012). *Compressed Sensing: Theory and Applications*, edited by G. Kutyniok & Y. C. Eldar, pp. 210–268. Cambridge University Press.

Vestergaard, B. & Sayers, Z. (2014). *IUCrJ*, **1**, 523–529.

Williamson, T. E., Craig, B. A., Kondrashkina, E., Bailey-Kellogg, C. & Friedman, A. M. (2008). *Biophys. J.* **94**, 4906–4923.