Journal of
Applied
Crystallography
ISSN 1600-5767

Received 22 May 2013
Accepted 17 March 2014

# Stoichiometries and affinities of interacting proteins from concentration series of solution scattering data: decomposition by least squares and quadratic optimization

**Himanshu Chandola,[a] Tim E. Williamson,[b] Bruce A. Craig,[c] Alan M. Friedman[b,d] and Chris Bailey-Kellogg[a]***

[a]Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA, [b]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA, [c]Department of Statistics, Purdue University, West Lafayette, IN 47907, USA, and [d]Markey Center for Structural Biology, Purdue Cancer Center and Bindley Bioscience Center, Purdue University, West Lafayette, IN 47907, USA. Correspondence e-mail: cbk@cs.dartmouth.edu

In studying interacting proteins, complementary insights are provided by analyzing both the association model (the stoichiometry and affinity constants of the intermediate and final complexes) and the quaternary structure of the resulting complexes. Many current methods for analyzing protein interactions either give a binary answer to the question of association and no information about quaternary structure or at best provide only part of the complete picture. Presented here is a method to extract both types of information from X-ray or neutron scattering data for a series of equilibrium mixtures containing the initial components at different concentrations. The method determines the association pathway and constants, along with the scattering curves of the individual members of the mixture, so as to best explain the scattering data for the mixtures. The derived curves then enable reconstruction of the intermediate and final complexes. Using simulated solution scattering data for four hetero-oligomeric complexes with different structures, molecular weights and association models, it is demonstrated that this method accurately determines the simulated association model and scattering profiles for the initial components and complexes. Recognizing that experimental mixtures contain static contaminants and nonspecific complexes with the lowest affinities (inter-particle interference) as well as the desired specific complex(es), a new analytical method is also employed to extend this approach to evaluating the association models and scattering curves in the presence of static contaminants, testing both a nonparticipating monomer and a large homo-oligomeric aggregate. It is demonstrated that the method is robust to both random noise and systematic noise from such contaminants, and the treatment of nonspecific complexes is discussed. Finally, it is shown that this method is applicable over a large range of weak association constants typical of specific but transient protein–protein complexes.



## 1. Introduction

Gaining a deeper understanding of the functions and mechanisms of protein–protein interactions requires extending the binary information (interaction or not) provided by high-throughput techniques and characterizing the stoichiometries, affinities and three-dimensional structures of protein complexes. However, experimental methods for the detailed study of protein complexes typically fall into two separate categories: some (*e.g.* X-ray crystallography and NMR spectroscopy) enable structure determination but do not readily reveal the association model, while others [*e.g.* H/D exchange (Codreanu *et al.*, 2005), analytical ultra-centrifugation (Lebowitz *et al.*, 2002), titration calorimetry (Velazquez-Campoy *et al.*, 2004) and composition gradient static light scattering (Attri & Minton, 2005; Kameyama & Minton, 2006)] enable characterization of the stoichiometry and strength of interaction but provide no or very limited structural information.

As we show here, small-angle scattering in solution (SAS) (Feigin & Svergun, 1987) provides an alternative experimental technique that can simultaneously provide both structural and association information for a complex. SAS has recently

gained popularity in low-resolution structural studies of protein monomers and tight complexes (Svergun & Stuhrmann, 1991; Walther *et al.*, 2000; Chacón *et al.*, 1998; Svergun, 1999; Svergun *et al.*, 2001), as it is applicable to proteins of practically any size under physiological conditions, while data can now be collected rapidly at new higher-flux X-ray or neutron sources. However, its applicability to the study of complexes has been limited owing to the requirement for a homogeneous and monodisperse sample, rendering SAS unsuitable for important, more weakly binding, transient but specific complexes (*e.g.* those associated with cellular signaling, which contain mixtures of the component monomers and intermediate and final complexes).

We recently described a method for the elucidation of weaker homo-oligomeric complexes from solution scattering data (Williamson *et al.*, 2008), and subsequent reports of similar numerical approaches applied to experimental data (Bernadó *et al.*, 2009) have demonstrated the value of such methods. However, these methods were only applied to homo-oligomers and were limited in their ability to handle systematic noise in the scattering data. Here, we extend our earlier method so as to characterize hetero-oligomeric complexes and we develop a new analytical approach to handle contaminants in the mixtures, thereby yielding a method with potential applicability to an even broader range of biological systems and experimental conditions.

An equilibrium mixture of protein components contains multiple different molecular species, including the initial components (often monomers), the desired higher-affinity complexes (both intermediate and final), nonspecific complexes of the lowest affinity (sometimes thought of as interparticle interference) and perhaps static contaminants. The method presented here focuses on the initial components and higher-affinity complexes, and includes an extension for particular static contaminants (such as a nonparticipating monomer or homo-oligomeric aggregate). The method determines the association model (stoichiometry and affinity constants) for the higher-affinity complexes from SAS data collected from a set of solutions containing the initial components in varying concentrations. In addition to the association model, this method accurately reconstructs the individual scattering curves of all the molecular species. These reconstructed curves can form the basis for low-resolution structural analysis of the intermediate and final complexes. While not addressed by the present method, the handling of interparticle interference is important, and interesting future work and possible ways to deal with the lowest-affinity nonspecific complexes are discussed.

Scattering from such equilibrium mixtures can be approximated as a fractional mass-weighted linear combination of the 'pure' scattering from the initial components and specific complexes (an approximation that is most accurate under conditions when the lowest-affinity nonspecific complexes and contaminants make only a small contribution). First, low-rank approximation is employed to remove from the observed mixture data some of the experimental noise and contributions from minor species. A search is then carried out over possible association models (which define a set of expected fractional masses for all the species), establishing a least-squares problem for each. Solution of the least-squares problem yields reconstructions of the pure scattering curves. These hypothesized reconstructions are evaluated for consistency with the data and with the postulated association model, and the best model is selected. If no model is of sufficient quality, the search can be expanded to consider association models containing a static contaminant. We have investigated the situation where the contaminant is either a nonparticipating monomer or a homo-oligomeric aggregate of one of the initial components, since these represent the most important practical situations where the contaminants are less likely to be removed by biochemical means during preparation of the initial components. In these cases, the least-squares approach is no longer applicable, so, at the cost of computational time, a convex quadratic program is employed to compute scattering curves that are consistent with the data and which satisfy the additional constraints expected of physically realistic scattering curves.

We demonstrate the effectiveness of this method on simulations of four hetero-oligomeric complexes with different association pathways, association constants, molecular weights and three-dimensional structures. Our simulation studies further demonstrate the robustness of the method to both random noise and systematic noise due to contaminants. In all cases, it is possible to infer the correct association pathway and obtain association constants that are very close to those used in the simulation, as well as scattering curves that closely approximate those of the monomers and oligomers.

## 2. Methods

When several molecules are present in a solution, the observed scattering curve is the mass-fraction-weighted linear combination of the scattering intensities for the individual components. Starting with scattering intensities collected from the equilibrium mixtures of a series of different concentrations of the initial components, the goal is to infer the association model, along with the underlying scattering curves of the molecular species involved, including the initial components and intermediate and final complexes. Fig. 1 provides an overview of the present approach for an example in which the initial components $A$ and $B$ form an $AB$ complex, with an association constant $K_{AB}$ establishing the fractional amount of each of these forms at equilibrium. Each molecular species has an underlying scattering curve, but the association model and underlying scattering curves are unknown (gray shaded box in Fig. 1). At given initial concentrations of $A$ and $B$, the scattering curve for the equilibrium mixture is the weighted sum of those for pure $A$, pure $B$ and pure $AB$, weighted by the equilibrium fractional masses. The experimentally measured curve (normalized by the total mass concentration of the mixture) is then composed of this weighted sum, plus experimental error. A series of such curves is collected (or for the results presented here, simulated) over a range of initial concentrations of $A$ and $B$. A search is then carried out over

possible association models, considering alternative pathways and values for the corresponding association constants (here only $K_{AB}$). When considering a possible pathway, an associated number $p$ of molecular species present is hypothesized, and thus a corresponding reduced set of scattering curves with random experimental noise partly removed can be extracted. When considering a set of association constants under this pathway, a set of fractional masses is hypothesized and these are used to reconstruct the underlying curves. To determine the best association model and reconstructed curves, first a broad coarse-grid search is performed over possible association constants, followed by a narrow fine-grid search, and each is scored for quality of fit to the observed data and agreement between the scattering curves and proposed stoichiometries of the complexes. Finally, the best pathway and constant are returned, along with the corresponding reconstructed curves.

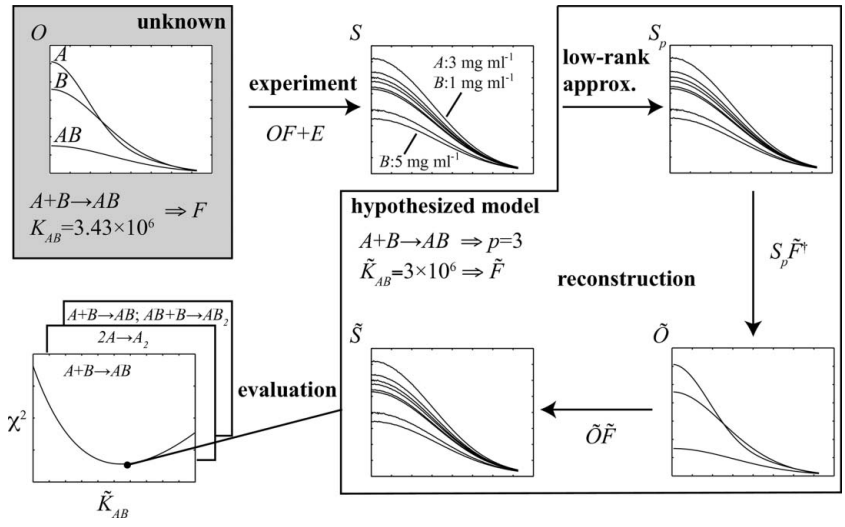


**Figure 1**
An overview of the present method for an example one-stage system. The association model and scattering curves of the various molecular species are unknown. Scattering curves are collected over a series of different initial concentrations of the components. Each observed scattering curve is a linear combination of the unknown curves of the different species, according to the association model and initial concentrations of the components, plus noise. A systematic search is carried out over possible association models; for each, a corresponding low-rank approximation is used to de-noise the data, and a least-squares formulation is employed to reconstruct the scattering curves of the different species. The agreement of the reconstructed curve of each model with the experimental data is evaluated and the best model selected. This ideal framework is then extended to account for the most problematic possible contaminants (nonparticipating monomers and homo-oligomeric aggregates have been tested) by including an additional unknown scattering curve and fractional mass, and solving a quadratic optimization problem for reconstruction.

More formally, the input scattering data are represented as an $m \times n$ matrix $S$, with $n$ columns for $n$ samples at different concentrations of the initial $A$ and $B$ components, each with $m$ rows for the scattering intensities at a fixed set of $m$ scattering angles. Each scattering curve normalized to the mass concentration (column in $S$) represents a linear combination of $p$ curves (the initial components and intermediate and product oligomers, each at the standard mass concentration), weighted according to their equilibrium fractional masses. Collecting the curves into an $m \times p$ matrix $O$ (one column per molecular species) and the fractional masses into a $p \times n$ matrix $F$ (one row per set of initial monomer concentrations), and adding experimental noise $E$ (one value per data point), we obtain

$$S = OF + E. \tag{1}$$

While $S$ is the observed data, the values in the other matrices are unknown. The goal is to infer the association model, which determines $F$ and the set of curves $O$. These in turn produce the observed matrix $S$.

We now detail each of the steps in the following subsections. The presentation is generalized from that of our previous homo-oligomeric study (Williamson *et al.*, 2008) and refocused directly on solving the underlying least-squares problem. We initially assume that only the species in the modeled association are present in the various mixtures. We subsequently show how to modify the methods to handle potential situations where the presence of a contaminant that is a nonparticipating monomer or homo-oligomeric aggregate alters the ideal situation.

### 2.1. Low-rank approximation

When considering an association pathway (recall that the search will be carried out over the possibilities), the number $p$ of molecular species that are present at equilibrium is known. Since the relationship between their mass fractions (and hence between rows of $F$) is nonlinear, and since the number of concentrations is greater than the number of molecular species, a $p$-rank approximation $S_p$ can be extracted. This low-rank approximation $S_p$ is a 'de-noised' version of $S$ (*i.e.* with $E$ partially removed) containing the appropriate number $p$ of curves with which to reconstruct the scattering curves according to the association model.

Singular value decomposition (SVD) is a popular technique for low-rank approximation and has been employed by us (Williamson *et al.*, 2008) and others (Segel *et al.*, 1998, 1999; Chen *et al.*, 1996) in the analysis of scattering data. SVD computes the low-rank approximation with the smallest distance to the input matrix, as measured by the Frobenius norm of the matrix difference,

$$\|S - S_p\|_F = \left\{ \sum_{i,j} \left[ S(i,j) - S_p(i,j) \right]^2 \right\}^{1/2}. \tag{2}$$

The SVD of the $m \times n$ matrix $S$ is given by $S = U\Sigma V^{\mathrm{T}}$, where $m \times m$ matrix $U$ and $n \times n$ matrix $V$ are orthogonal matrices whose column vectors are the left and right singular vectors,

respectively, and $m \times n$ matrix $\Sigma$ is a diagonal matrix whose elements are the singular values associated with the corresponding left/right singular vectors. The singular values are in order along the diagonal from largest to smallest, weighting the contributions from the most to the least important singular vectors. To compute the $p$th low-rank approximation, the smallest $m - p$ singular values on the diagonal of $\Sigma$ are replaced with zero to give $\Sigma_p$, and then $S_p = U\Sigma_p V^{\mathrm{T}}$ is computed.

### 2.2. Reconstruction

When considering a set of association constants for a pathway (recall that a grid search will be conducted over possible values for the association constants), standard association equilibria can be applied to compute the resulting equilibrium fractional mass of each of the $p$ molecular species. These fractional masses are collected into a matrix $\tilde{F}$ (where the tilde indicates that it is a reconstruction of the 'true' unobserved matrix $F$). Combining this with the low-rank approximation $S_p$ in a de-noised version of equation (1), the least-squares solution is computed in order to reconstruct the scattering curves of the various species:

$$\tilde{O} = S_p \tilde{F}^{\dagger}, \tag{3}$$

where $\tilde{F}^{\dagger}$ denotes the Moore–Penrose pseudo-inverse. This formalization in terms of a $p$-rank approximation is a generalization of the approach given by Williamson *et al.* (2008), where using basis vectors from SVD was an explicit part of the equations. It clarifies the role of the decomposition and allows the use of alternative approximation approaches. It is also different from the approach of Bernadó *et al.* (2009), where a related approach employs principal component analysis to find the number of components in the solution.

If the least-squares solution $\tilde{O}$ has more than 10% negative intensity values or contains negative values in the small scattering angle range considered for Guinier analysis (Dervichian *et al.*, 1952), it is considered to be nonphysical and the reconstruction is rejected without further analysis.

The solution $\tilde{O}$ is then used to compute $\tilde{S}$, an approximation of the observed scattering curves of the equilibrium mixtures, by linearly combining the curves of the species involved at the appropriate fractional masses:

$$\tilde{S} = \tilde{O}\tilde{F} = S_p \tilde{F}^{\dagger}\tilde{F}. \tag{4}$$

Thus, the low-rank approximation is used to reconstruct the scattering data so as to be consistent with the hypothesized association model.

### 2.3. Evaluation

An association model is assessed in terms of how well the reconstructed scattering curves $\tilde{S}$ match the experimental ones $S$. The two scoring approaches of our homo-oligomeric work (Williamson *et al.*, 2008) are employed, customized for hetero-oligomers.

First, a $\chi^2$ score quantifies the differences over the entire set of scattering curves, weighted by the estimated error $\sigma(i, j)$ for each experimental data point:

$$\chi^2 = \frac{1}{m(n-p)} \sum_{j=1}^{n} \sum_{i=1}^{m} \left[ \frac{S(i,j) - \tilde{S}(i,j)}{\sigma(i,j)} \right]^2. \tag{5}$$

The sum of the squared differences between points on the reconstructed and original curves is normalized by $m(n - p)$ degrees of freedom to yield a $\chi^2$ score. While there are $mn$ data points, $p$ of the $n$ degrees of freedom are fixed by the low-rank approximation. In practice, this score is observed to be approximately 1 for the best fit to the data with Gaussian simulated noise.

Second, the mean-squared mass ratio difference (MSMRD) score calculates whether the zero-angle intensities match the stoichiometry of the hetero-oligomeric forms. The scattering intensity at zero angle, estimated by Guinier analysis (Dervichian *et al.*, 1952), is proportional to the molecular weight. Thus, for example, one would expect $I(0)$ for species $AB$, denoted $I_{AB}(0)$, to equal $I_A(0) + I_B(0)$, and thus $I_{AB}(0)/[I_A(0) + I_B(0)] = 1$. Thus, the MSMRD score computes the average, over the various hetero-oligomeric forms, of the deviations of such ratios from the ideal value of 1. Its expected value is thus zero. For a hetero-oligomer formed by $A$ and $B$ monomers, we compute the MSMRD as

$$\mathrm{MSMRD} = \frac{1}{p-2} \sum_{(a,b)\in C} \left[ 1 - \frac{I_{A_a B_b}(0)}{a\, I_A(0) + b\, I_B(0)} \right]^2, \tag{6}$$

where $C$ is a set of $(a, b)$ pairs indicating the various $A_aB_b$ hetero-oligomeric forms and $I_{A_aB_b}(0)$ represents their zero-angle intensity. For example, if the association model is $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$, then

$$\mathrm{MSMRD} = \frac{1}{2} \left\{ \left[ 1 - \frac{I_{AB}(0)}{I_A(0) + I_B(0)} \right]^2 + \left[ 1 - \frac{I_{AB_2}(0)}{I_A(0) + 2I_B(0)} \right]^2 \right\}. \tag{7}$$

These two scores are complementary. The $\chi^2$ value is global, assessing the overall agreement between the reconstruction and the data. However, two related association pathways (with an appropriate choice of association constants) can generate similar solutions and similar $\chi^2$ values. For example, this can happen with a one-stage association pathway $A + B \rightarrow AB$ and the extended two-stage association pathway $A + B \rightarrow AB$, $A + B \rightarrow AB_2$, with similar association constants $K_{AB}$ for both cases and a very weak $K_{AB_2}$ for the second (see *Results*, §3). This is because equation (4) can give similar solutions for two different matrices $F$, as long as the column space spanned by the fractional matrix is the same. On the other hand, the MSMRD is very local, ignoring the agreement over most of the curve and focusing on the zero-angle intensity in order to assess the agreement between the independent (and not directly optimized) expected molecular weights and the stoichiometry. We have found that considering both $\chi^2$ and the MSMRD improves the determination of the correct association model (see *Results*, §3).

### 2.4. Association model search

We have discussed how to reconstruct and evaluate scattering curves for a given association model defined by a pathway and a corresponding set of association constants. In order to determine the best model, models for a set of plausible pathways are separately reconstructed and evaluated over a grid of possible association constants.

The pathways to be considered are chosen from the set of oligomers that could possibly be present in the equilibrium mixture. Although that set is potentially infinite, a most likely set of oligomers can be selected, for example, from an analysis of the zero-angle scattering or by the radii of gyration of the experimental scattering curves. Then all pathways that could form complexes with the allowed sets of subunits are considered. For example, if it is known that there are two monomers, $A$ and $B$, and it has been determined that the final oligomer has at most three subunits, then the one-stage associations $2A \rightarrow A_2$, $2B \rightarrow B_2$ and $A + B \rightarrow AB$ would be evaluated, along with the two-stage associations that extend these to yield $A_2B$ and $AB_2$. Like other approaches, for example analytical ultracentrifugation, where postulated association models are fitted to the data, assumptions have to be made for the most likely models to be assessed.

Coarse- and fine-grid searches are performed over possible values for the association constants. Each association constant is an independent dimension in the grid. The results presented below use grids covering the range of plausible constants: $10^{-6}$ to $10^{25}$ for a one-stage association and $10^{1}$ to $10^{15}$ for a two-stage association. An initial coarse grid is searched at integer multiples of the powers of 10 (*e.g.* $1 \times 10^3, 2 \times 10^3, 3 \times 10^3, \ldots, 9 \times 10^3, 1 \times 10^4, 2 \times 10^4, \ldots$). For each point (representing one or a pair of association constants), the curves are reconstructed and evaluated by $\chi^2$ and MSMRD, as described above. The constants with the best scores establish a region for a fine-grid search, plus or minus one unit in each dimension, with a spacing of 1% of that of the coarse grid. Fine-grid searches are only performed for the models with the best $\chi^2$ and MSMRD values from the coarse-grid search and for which the best coarse-grid association constants from the $\chi^2$ and MSMRD scores are in sufficient agreement. Finally, the model with the best fine-grid $\chi^2$ and MSMRD scores is selected, determining the corresponding pathway, association constants and reconstructed curves. In cases where the fine-grid search fails to yield an acceptable model, owing to either a high $\chi^2$ for the best fine-grid point or a large disagreement between the best $\chi^2$ and MSMRD fine-grid points, the methods in the next section can be employed to account for contaminants.

### 2.5. Accounting for contaminants

An extension to the current methodology has been developed to deal with the case when the scattering data contain a substantial contaminant. Since contaminants that are unrelated to the initial components are generally readily purified out by current protein-separation methods, we seek to solve the biochemical situations that arise most frequently. Thus, our focus is on cases in which the contaminant is either a nonparticipating monomer or a large homo-oligomeric aggregate of one of the components.

Let us assume that the contaminant is a nonparticipating monomer or homo-oligomeric aggregate of $A$ (the methodology works the same for any component and could be generalized to multiple such contaminants). Note that, in our approach, the contributions from all species in a polydisperse homo-oligomeric aggregate can be accounted for by one combined scattering curve and one total contaminant fraction. Let $c$ be the unknown mass fraction of $A$ that forms the contaminant. As part of the grid search, possible values for $c$ will be considered along with those for the association constant(s). Given hypothesized values for $c$ and the association constant(s), a fractional mass matrix $\tilde{F}$ must be built for each, now containing $p + 1$ rows, with the extra row for the contaminant. In constructing this matrix, let $a_i$ be the initial amount of $A$ in sample $i$. Then the amount of $a_i$ still participating in the hypothesized association (rather than in the contaminant) is $a_i(1 - c)$. The equilibrium concentrations, and thereby the masses of the other forms, are determined from the reduced $A$ concentration and the initial concentrations of the other initial component(s).

Unfortunately, the extended matrix $\tilde{F}$ is no longer of full rank in the presence of contaminant, as the fractional mass vector for the contaminant is linearly dependent on $A$. This in turn implies that there is an infinite set of widely varying least-squares solutions $\tilde{O}$ satisfying $\tilde{O}\tilde{F} = S_p$. One of these, denoted $\tilde{O}_0$, is the solution from equation (3), $\tilde{O}_0 = S_p\tilde{F}^{\dagger}$. Using this $\tilde{O}_0$ to reconstruct $\tilde{S}$, as in equation (4), gives $S_p\tilde{F}^{\dagger}\tilde{F}$, denoted $S_{p,\tilde{F}}$. Each least-squares solution $\tilde{O}$ produces this same $S_{p,\tilde{F}}$ and thus cannot be distinguished by comparison with the data $S$ or the de-noised data $S_p$. This equivalence of solutions $\tilde{O}$ is due to the fact that the set of least-squares solutions is composed of the sum of $\tilde{O}_0$ with an infinite set of matrices of row vectors (that is, adjustments to the scattering curves) from the null space of $\tilde{F}^{\mathrm{T}}$. Post-multiplication by $\tilde{F}$ then reduces the second matrix in this sum to zero, resulting in no change in $S_{p,\tilde{F}}$.

In summary, there are an infinite number of reconstructions of the pure curves $\tilde{O}$, but each produces the same reconstructed data $S_{p,\tilde{F}}$. Since the reconstructed data are used to compute $\chi^2$ [equation (5)], the best association model (best $\tilde{F}$) can be found *via* coarse- and fine-grid searches as before, with an additional dimension of the contaminant fraction in addition to the association constant(s). However, this approach does not produce correct reconstructed pure scattering curves and thus also does not give MSMRD values. Therefore, after identifying the best $\chi^2$ point (or a set of feasible points for consideration), one must search over the space of satisfying $\tilde{O}$ to reconstruct and evaluate pure scattering curves and identify the best one.

A quadratic optimization framework has been developed that seeks a solution $\tilde{O}$ that not only explains the data (which all $\tilde{O}$ do equally) but also has properties desirable of physically realistic scattering curves. In particular, smoothness is established as the objective function and constraints are incorporated limiting the sub-optimality of $\chi^2$, while the

expected decaying exponential trend in the Guinier region of the scattering curves is also enforced, as well as the expected ratios of $I(0)$ values (as also employed in the MSMRD score). Note that, if the contaminant only involves form $A$, for example, then the row for $B$ in the fractional mass matrix is linearly independent of the contaminant and yields a unique least-squares solution (the same in $\tilde{O}$ for any $\tilde{O}_0$). Thus, after computing $\tilde{O}_0$, the row for initial component $B$ is removed from $\tilde{F}$ and from $S_{p,\tilde{F}}$ (*via* its row in $\tilde{F}$ and its column in $\tilde{O}_0$). For simplicity, we continue to refer to $\tilde{O}$ and $\tilde{F}$ without distinguishing the reduced-parameter versions.

We now outline the components of the quadratic program: the objective to optimize and the constraints to limit the considered solutions.

**2.5.1. Objective: smoothness**. With the available freedom in $\tilde{O}$, there are curves that use wildly fluctuating values to obtain good $\chi^2$ scores upon post-multiplication by $\tilde{F}$. Since physical curves are expected to be relatively smooth, a discrete evaluation of smoothness is established as the objective function. A finite difference matrix $D$ is constructed that, when multiplied by $\tilde{O}$, approximates the second-order derivative at each point on the curve. The quadratic program then seeks to minimize the total of the squared differences, *i.e.* the square of the Frobenius norm of $D\tilde{O}$:

$$\min_{\tilde{O}} ||D\tilde{O}||_F^2. \tag{8}$$

**2.5.2. Constraint: $\chi^2$ deviation**. A reconstruction is sought with the optimum $\chi^2$ (as with all the $\tilde{O}$, satisfying $\tilde{O}\tilde{F} = S_{p,\tilde{F}}$), but since the data are noisy, one may sacrifice a little in the $\chi^2$ score in order to ensure a feasible optimization problem and do better in terms of smoothness and other characteristics. Thus, a constraint is imposed that the reconstructed curves are no more than a tolerance $\varepsilon_{\rm fit}$ away from the one that gives the lowest $\chi^2$. This tolerance should be set fairly low to keep the identified curves near the optimum one; for the present results, a value of $10^{-3}$ is used. The constraint then requires

$$(1 - \varepsilon_{\rm fit})S_{p,\tilde{F}} \leq \tilde{O}\tilde{F} \leq (1 + \varepsilon_{\rm fit})S_{p,\tilde{F}}. \tag{9}$$

**2.5.3. Constraint: non-negativity**. This requires that the scattering curves are non-negative,

$$\tilde{O} \geq 0. \tag{10}$$

**2.5.4. Constraint: Guinier**. Scattering curves decay exponentially in the Guinier region (Dervichian *et al.*, 1952). Therefore, a constraint is imposed that the curves are non-increasing (within a tolerance) in the initial Guinier region. To approximate the Guinier region in the scattering curves in $\tilde{O}$ without iterating on $R_{\rm g}$ (radius of gyration) values, $q_{\max} = 1.33/R_{\rm g}$ (Guinier & Fournet, 1955) and a fixed $R_{\rm g} = 40$ Å are used. To allow for noise, this property is enforced only to within a tolerance $\varepsilon_{\rm Guinier}$: within the Guinier region, a given intensity is no more than $(1 + \varepsilon_{\rm Guinier})$ times the intensity at the next lower scattering angle. A reasonable value for $\varepsilon_{\rm Guinier}$ can be estimated by examining some pure intensity curves that have been reconstructed from uncontaminated simulations with standard noise; a value of $2 \times 10^{-2}$ is used here. Note that this value is dependent on the extent of the noise and the spacing of the scattering angles. This constraint is formulated with a matrix $G$ which, when multiplied by $\tilde{O}$, gives the differences between $(1 + \varepsilon_{\rm Guinier})$ times a particular point and the next point, for points in the scattering curves in $\tilde{O}$ at $q < q_{\max}$:

$$G\tilde{O} \geq 0. \tag{11}$$

**2.5.5. Constraint: molecular weights**. When considering a contaminant $X$ that is a nonparticipating form of $A$ (either monomer or aggregate), its native mass must be at least that of $A$, *i.e.* $M_X > M_A$. Thus, the zero-angle intensity of its scattering curve should be at least equal to that of $I_A(0)$. Since the extrapolation to obtain $I(0)$ requires an exponential fit (which would render the system nonlinear), the intensity at the smallest angle measured, $I(q_{\min})$, is used instead:

$$I_X(q_{\min}) - I_A(q_{\min}) \geq 0, \tag{12}$$

where the scattering curves $I_A$ and $I_X$ (for $A$ and the contaminant $X$) are particular vectors of $\tilde{O}$.

Imposing this constraint on $I(q_{\min})$ instead of $I(0)$ results in negligible error, since, from the Guinier relationship,

$$\frac{I_X(q_{\min})}{I_A(q_{\min})} = \frac{M_X}{M_A}\exp\left\{-\frac{1}{3}\,q_{\min}^2\left[R_{\rm g}(X)^2 - R_{\rm g}(A)^2\right]\right\}. \tag{13}$$

Given that $q_{\min}^2$ is generally quite small (of the order of $10^{-6}$ in experimental data), the difference in radii of gyration is not large enough to impact the results substantially.

Furthermore, since a unique scattering curve for $B$ has been found, its intensity at $q_{\min}$ can be used to constrain the intensity at $q_{\min}$ of the scattering curve for $A$ and other forms (excluding the contaminant). For example, it is expected that

$$\frac{I_A(q_{\min})}{I_B(q_{\min})} \simeq \frac{M_A}{M_B}. \tag{14}$$

Essentially, this is encoding MSMRD (relative to the independent form $B$) as a constraint, but for intensities at $q_{\min}$ instead of at zero angle. As with most other constraints, a tolerance is used to allow for some noise. Thus, the approximate equality of the intensity ratio and the mass ratio is encoded as a constraint on the ratio between these two ratios – it must be within a tolerance $\varepsilon_{\rm MSMRD}$ of the desired value of 1. A value of $\varepsilon_{\rm MSMRD} = 0.1$ has been found to work well for the present tests, but for other data this tolerance could potentially be tightened further, as long as feasible solutions still result. For the scattering from $A$ ($I_A$) and every other molecular species $A_kB_l$ ($I_{A_kB_l}$), constraints are added of the form

$$I_A(q_{\min}) \geq (1 - \varepsilon_{\rm MSMRD})I_B(q_{\min})\frac{M_A}{M_B}, \tag{15}$$

$$I_A(q_{\min}) \leq (1 + \varepsilon_{\rm MSMRD})I_B(q_{\min})\frac{M_A}{M_B}, \tag{16}$$

$$I_{A_kB_l}(q_{\min}) \geq (1-\varepsilon_{\mathrm{MSMRD}})I_B(q_{\min})\frac{kM_A+lM_B}{M_B}, \qquad (17)$$

$$I_{A_kB_l}(q_{\min}) \leq (1+\varepsilon_{\mathrm{MSMRD}})I_B(q_{\min})\frac{kM_A+lM_B}{M_B}, \qquad (18)$$

where again the scattering curves $I$ are particular vectors in $\tilde{O}$.

**2.5.6. Solving the system**. While the objective and constraints have been written in terms of $\tilde{O}$ and other matrices, these matrices can be reshaped into long vectors (*i.e.* by stacking columns). The combination of the objective function and constraints yields a convex quadratic optimization problem that can be solved by numerous solvers. If the quadratic optimization program is not feasible for a hypothesized association model, that model is discarded. If more than one feasible model were to remain, MSMRD values could be computed and the best selected, but that did not happen in the simulation studies presented below.

### 2.6. Implementation

The methods have been implemented in a platform-independent Python package that is available from the authors upon request. The package calls the IBM *ILOG CPLEX* optimizer to solve the system of equations. The program allows a user to search over possible association models based on specifications provided *via* the command line or in an input file. The package contains implementations for both a contaminant-free search and an extension to handle non-participating monomers and homo-oligomeric contaminants. In addition to the methods in this paper, it also contains an implementation for homo-oligomeric association models from our previous work (Williamson *et al.*, 2008).

To obtain the results presented below, coarse- and fine-grid searches for a one-stage model took less than a minute, while searches for a two-stage model took a few minutes on a single-core Intel Xeon 2.50 GHz processor. The three-stage searches took a few hours. Grid searches with contaminant for one-stage association took a few minutes, while contaminant searches for two-stage association took a few hours. The quadratic program solver usually took less than a minute.

## 3. Results

In order to evaluate the effectiveness of the present method in a range of scenarios, an extensive set of simulation studies were performed with different association pathways and association constants, and varying levels of random noise, data resolution and monomer size. Fig. 2 summarizes the complexes used in these studies, and illustrates their crystal structures and the simulated scattering curves of the monomers and intermediate and final oligomers at a constant mass concentration. The complex structures were taken from the Protein Data Bank (PDB; Berman *et al.*, 2000) (PDB codes indicated), and monomer and intermediate complex structures were extracted. The association models for simulation were not taken from experimental data; instead, they were chosen to challenge the ability of the method to determine the correct

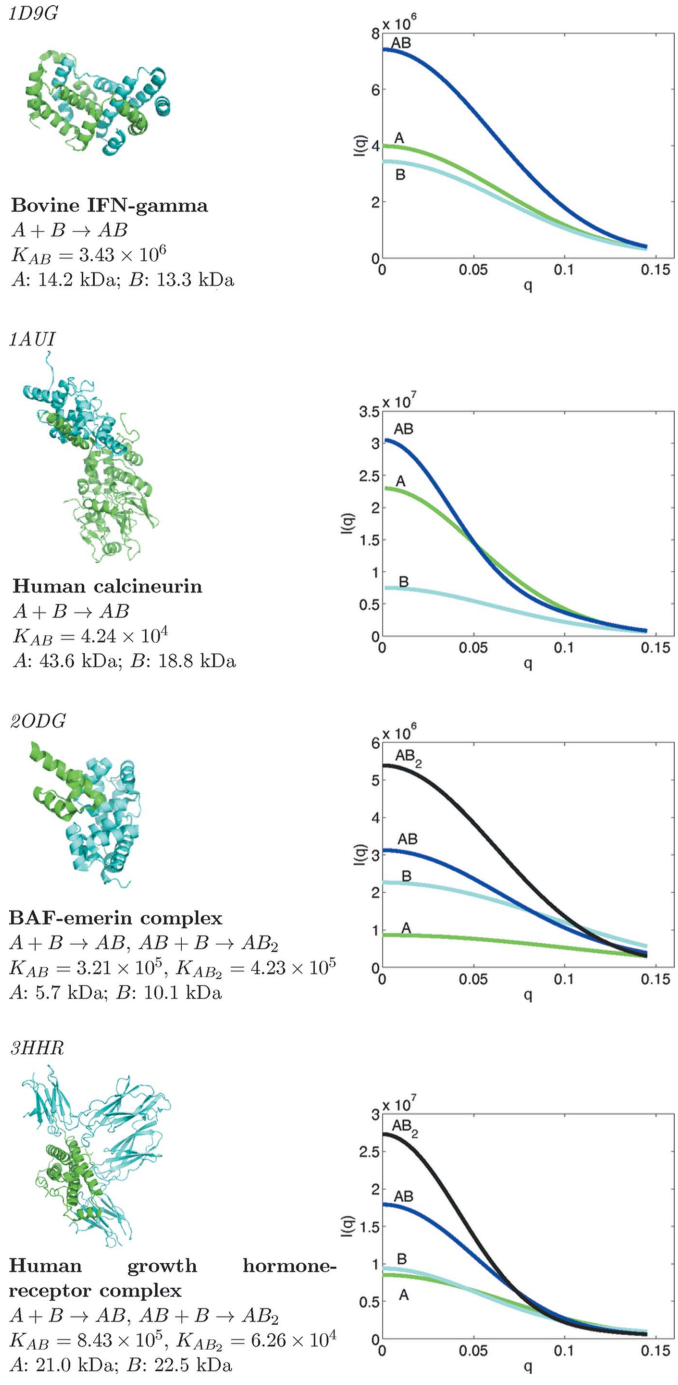


**Figure 2**
The four case studies discussed in this article.

model even in the presence of alternatives that have intermediate and final complexes of similar mass (note the similarity of the initial component masses in the bovine IFN-$\gamma$ and human growth hormone-receptor cases). Association constants were chosen in the middle of a feasible range. However, the impact of the constants was explicitly assessed in one set of simulations.

It was found that as few as eight different initial concentrations provides a sufficient set of different scattering curves for subsequent reconstruction, and the results shown are

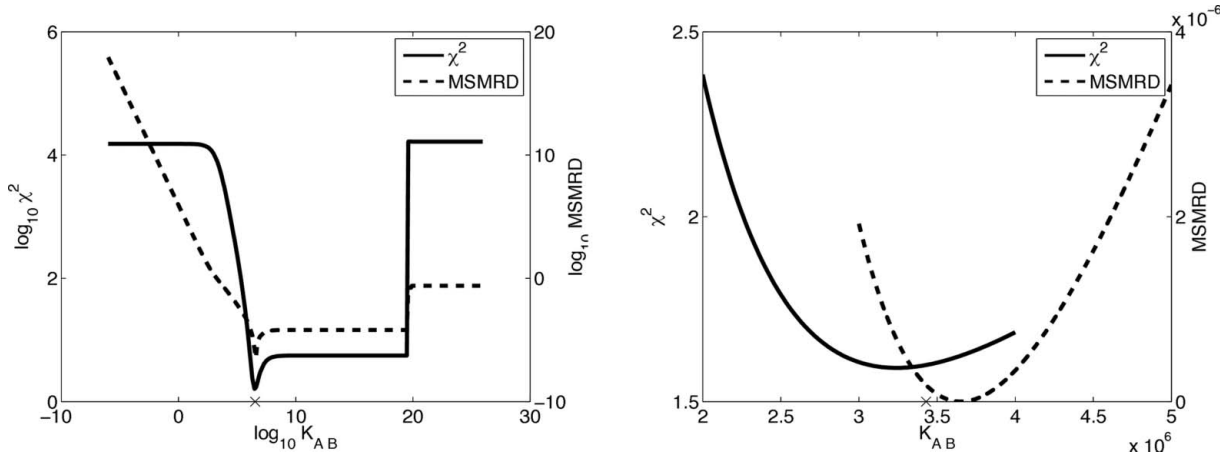


**Figure 3**
Association constant searches for one bovine IFN-$\gamma$ data set, for the correct $A + B \rightarrow AB$ pathway. The '$\times$' mark on the $x$ axis indicates the simulated association constant ($3.43 \times 10^6$).

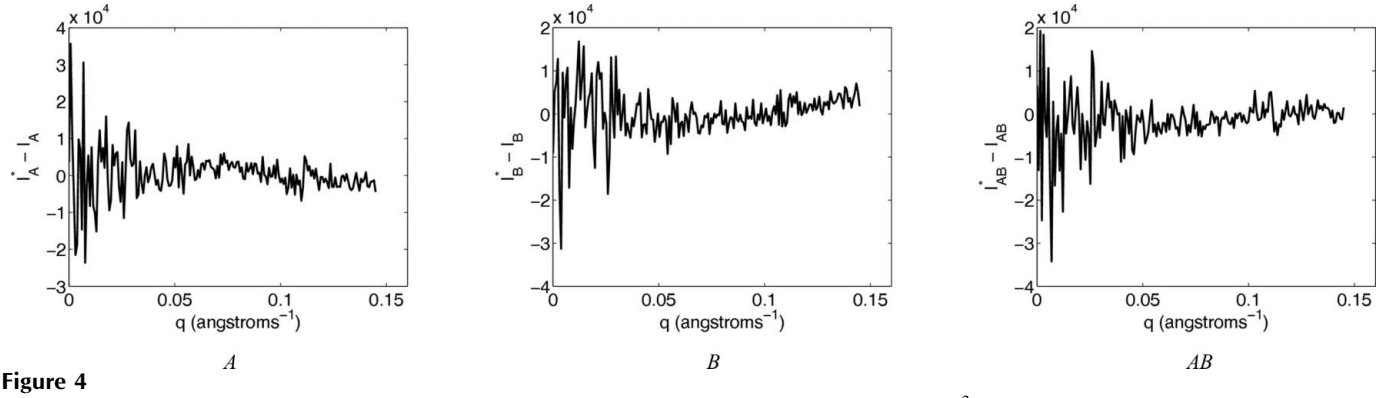


**Figure 4**
Residuals between pure simulated scattering intensities and the reconstructed ones for bovine IFN-$\gamma$ $\chi^2$ optimum association models.

based on eight for all test cases. The initial concentrations used (Supplementary Tables 1 and 2[1]) are all in the 0.5–5.0 mg ml$^{-1}$ range, where SAS data are easily collected. They were chosen so as to yield a diverse set of row vectors (fractional masses) in the fractional mass matrix $F$, adequately sampling the space and ensuring that important vectors (scattering from intermediate and final complexes) are included in the low-rank approximation. Even so, the equilibrium mixtures are rarely more than 70% of one form. In practice, of course, $F$ cannot be assessed initially, but it is still recommended that the user ensures that there is a diverse set of initial concentrations, with different combinations of low and high monomer concentrations. In the absence of approximate knowledge of the association constants that determine $F$, a first-round analysis can be used to identify a definitive set of initial concentrations from which to collect data. Pure monomer solutions (only $A$, only $B$) are included as initial components so as to characterize them better and account for their contributions to the mixtures. Of course, pure monomers may not be biochemically available, but the method is not dependent on this and any available components could be used.

The program *CRYSOL* (Svergun *et al.*, 1995) was used at the default settings to simulate noiseless scattering intensities $O$ from the three-dimensional structures of each initial component and complex. The noiseless equilibrium mixture intensities were then simply calculated as $OF$. Note that these curves include only the scattering within the initial components and complex members and do not capture contributions from any weak interparticle interactions. Noise $E$ was then added, following the method employed by Williamson *et al.* (2008), to simulate realistic angle-dependent Gaussian noise based on noise levels observed in experimental samples. Ten data sets were generated for each example, with different random noise added for each data set.

While two one-stage associations and two two-stage associations were studied, detailed results are presented for only one of each and the second is summarized, since the results were similar in each category. We first show that the method yields the correct association model on the initial simulated data, for both one-stage and two-stage examples. We then demonstrate the robustness of the method to noise and investigate the range of association constants for which the

[1] Supporting information discussed in this paper is available from the IUCr electronic archives (Reference: KK5143).

method is applicable. Finally, we consider test cases with simulated contamination and present results from the expanded method that accounts for the contaminant.

### 3.1. Baseline simulations

**3.1.1. Bovine IFN-$\gamma$ (one stage)**. We first examine the results for one of the ten simulated data sets (*i.e.* one Gaussian noise matrix $E$), with the correct pathway $A + B \rightarrow AB$ and varying the association constants on a coarse grid (Fig. 3, left) and fine grid (Fig. 3, right). Both plots show a steep decline in $\chi^2$ and MSMRD scores around the simulated association constant value ($3.43 \times 10^6$), with a minimum $\chi^2$ of 1.59 at $3.34 \times 10^6$ and a minimum MSMRD of $1.67 \times 10^{-11}$ at $3.65 \times 10^6$. The close agreement of these association constants and the high quality of the scores under these complementary metrics gives confidence in this solution.

Whereas in an experimental setting one would not have access to the 'true' scattering curves of the various molecular species ($O$), here one does (from the *CRYSOL* calculation on the model components and complexes), and one can evaluate how well the reconstructed curves agree with them [$\tilde{O}$, computed by equation (3)]. Fig. 4 shows the approximately random residuals between the reconstructed and simulated curves, at the association constant $K_{AB} = 3.34 \times 10^6$ which yields the best $\chi^2$ score. [The apparent deviation from random residuals seen at higher resolution for component $B$ (Fig. 4, middle) is not explained by deviation between simulated and best $\chi^2$ association constants.] To quantify the extent of agreement, the median of the absolute relative deviation (MARD) is computed as a percentage deviation of the reconstructed curve from the simulated one; a MARD value close to zero indicates that the reconstructed curve is very close to the original noiseless *CRYSOL* curve. MARD scores confirm the agreement illustrated in the figure: $A$ has a MARD of 0.24%, $B$ 0.16% and $AB$ 0.22%, averaged across the ten data sets with different simulated noise.

Table 1 summarizes the results of the best-scoring pathway (which is the correct one) over all ten simulated noisy data sets; Supplementary Table 3 includes results for alternatives. The $A + B \rightarrow AB$ pathway was always chosen and the average association constant was close to the simulated one, with only a small variation between data sets. Only the related two-stage pathways $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ and $A + B \rightarrow AB$, $AB + A \rightarrow A_2B$ obtained coarse-grid $\chi^2$ scores (averaging 1.62 and 1.55, respectively) competitive with that of the correct model (1.53); the rest were much worse. Both alternative models extend the correct model with an additional association of weak affinity, keeping the $A + B \rightarrow AB$ association as the primary one. Any additional association has an adverse effect on the MSMRD scores ($1.19 \times 10^{-3}$ and $8.21 \times 10^{-4}$, *versus* $3.74 \times 10^{-7}$ for the correct model), as the low-angle data do not support an oligomer with a molecular weight corresponding to $AB_2$ or $A_2B$. In addition, while the optimum association constants for $\chi^2$ and MSMRD are very similar for the correct model, the best association constants by these two metrics are quite different for the alternative models. Furthermore, there is no choice of constants that scores moderately well under both metrics, and the association constants giving the best $\chi^2$ score yield a poor MSMRD score and *vice versa*. For pathway $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$, the MSMRD for the association constant with the best $\chi^2$ score averages $9.53 \times 10^{-2}$ across the ten data sets, *versus* an average best MSMRD of $1.19 \times 10^{-3}$. On the other hand, the $\chi^2$ score for the association constants with the best MSMRD score is 33.45 on average. These values are more than an order of magnitude worse than the best $\chi^2$ and MSMRD scores for the correct pathway. Similar results are found for the second alternative pathway. Even though the $\chi^2$ scores are not good discriminators, the substantial deterioration in the MSMRD and the disagreement between MSMRD and $\chi^2$ metrics for the alternative models point to the correct $A + B \rightarrow AB$ pathway.

**Table 1**
Coarse- and fine-grid search results for the best-scoring (and correct) models for uncontaminated simulations, over ten sets of simulated noise.

Simulated association constants: bovine IFN-$\gamma$, $K_1 = 3.43 \times 10^6$; BAF–emerin complex, $K_1 = 3.21 \times 10^5$, $K_2 = 4.23 \times 10^5$.

| Search | $K_1$ | $K_2$ | Score |
|---|---|---|---|
| Bovine IFN-$\gamma$, $A + B \rightarrow AB$ | | | |
| Coarse $\chi^2$ | $3.30 \times 10^6 \pm 4.8 \times 10^5$ | – | $1.53 \pm 0.11$ |
| Coarse MSMRD | $3.70 \times 10^6 \pm 4.8 \times 10^5$ | – | $3.74 \times 10^{-7} \pm 3.0 \times 10^{-7}$ |
| Fine $\chi^2$ | $3.40 \times 10^6 \pm 7.1 \times 10^4$ | – | $1.49 \pm 0.12$ |
| Fine MSMRD | $3.64 \times 10^6 \pm 5.2 \times 10^5$ | – | $1.82 \times 10^{-11} \pm 1.8 \times 10^{-11}$ |
| BAF–emerin complex, $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ | | | |
| Coarse $\chi^2$ | $3.00 \times 10^5 \pm 0.0$ | $4.00 \times 10^5 \pm 0.0$ | $1.17 \pm 0.24$ |
| Coarse MSMRD | $3.00 \times 10^5 \pm 0.0$ | $4.00 \times 10^5 \pm 0.0$ | $1.33 \times 10^{-6} \pm 6.8 \times 10^{-7}$ |
| Fine $\chi^2$ | $3.21 \times 10^5 \pm 4.5 \times 10^3$ | $4.24 \times 10^5 \pm 6.9 \times 10^3$ | $1.12 \pm 0.24$ |
| Fine MSMRD | $3.24 \times 10^5 \pm 1.3 \times 10^4$ | $4.29 \times 10^5 \pm 1.9 \times 10^4$ | $1.03 \times 10^{-9} \pm 6.6 \times 10^{-10}$ |

**3.1.2. BAF–emerin complex (two stage)**. Fig. 5 shows both $\chi^2$ and MSMRD scores on the coarse and fine grids for the correct $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ pathway, for one example noisy data set. As in the one-stage case, there are well defined minima, with the best association constants yielding much better $\chi^2$ and MSMRD scores than nearby alternatives, at both coarse and fine resolutions. Again there is good agreement as to the best association constants under the two scores: $\chi^2$ gives $K_{AB} = 3.16 \times 10^5$, $K_{AB_2} = 4.16 \times 10^5$, and MSMRD gives $K_{AB} = 3.27 \times 10^5$, $K_{AB_2} = 4.35 \times 10^5$, with the simulated constants being $K_{AB} = 3.21 \times 10^5$, $K_{AB_2} = 4.23 \times 10^5$. Interestingly, under both metrics, the best association constants lie on a diagonal line in which $K_{AB}$ and $K_{AB_2}$ increase at a similar rate, ensuring that if more $AB$ is produced than the data dictate it is also converted to $AB_2$. While this keeps the fraction of $AB$ relatively constant, the resulting excessive depletion of $A$ and excessive formation of $AB_2$ yield worse scores at points along the diagonal line other than the minimum. The reconstructed intensities at the best association

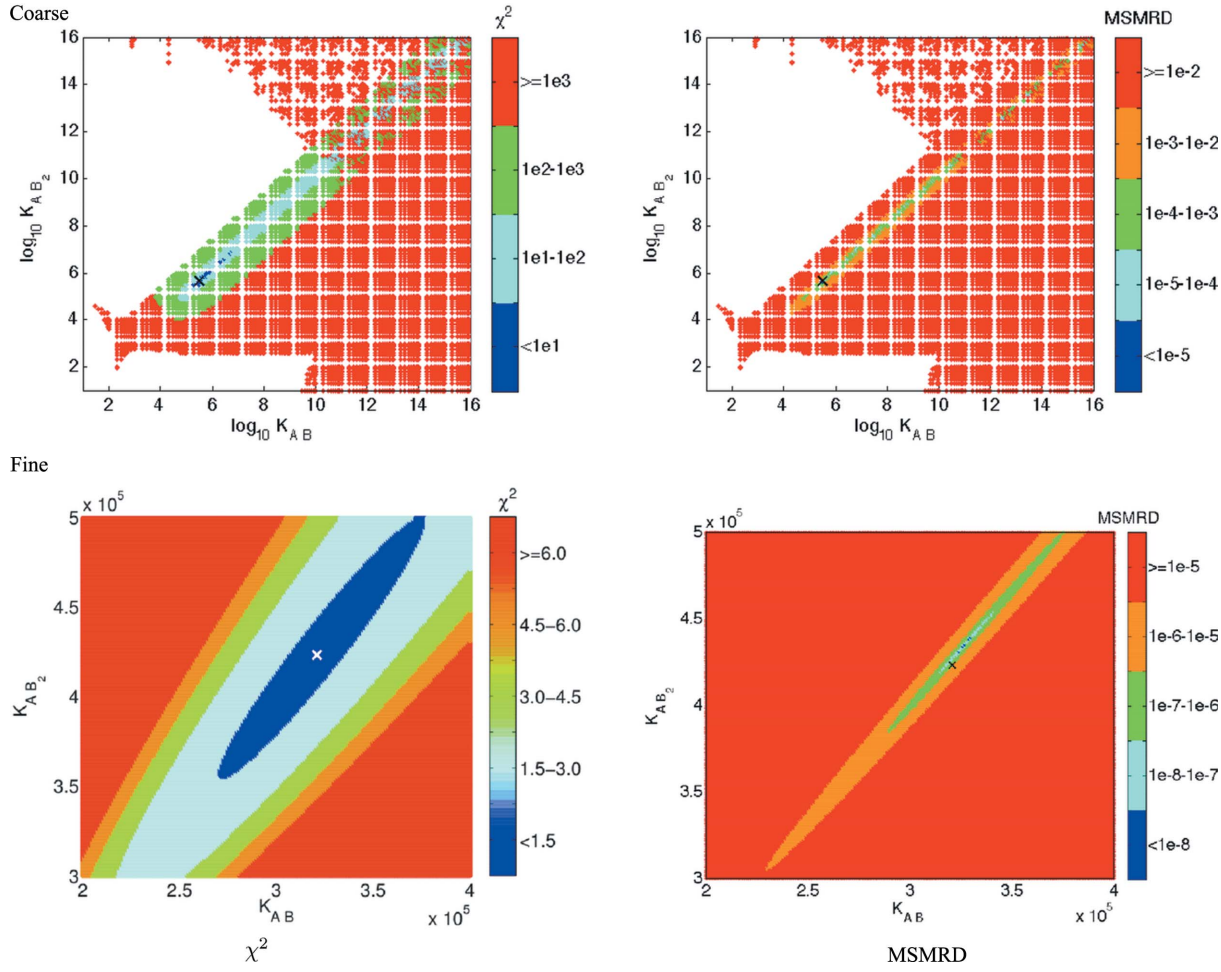


**Figure 5**
Association constant searches for one BAF–emerin complex data set, for the correct $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ pathway. The '×' marks indicate the simulated association constants ($K_{AB} = 3.21 \times 10^5$, $K_{AB_2} = 4.23 \times 10^5$). The white regions in the coarse-grid plots indicate constants yielding nonphysical scattering curves (those with substantial negative intensities).

constants are quite similar to the original simulated noiseless ones, as illustrated in the residuals (not shown) and quantified by average MARD values of 0.08% for $A$, 0.08% for $B$, 0.15% for $AB$ and 0.06% for $AB_2$.

Table 1 summarizes the results from ten simulations for the correct, best-scoring model; Supplementary Table 4 includes those for alternatives. The best coarse-grid $\chi^2$ score, averaging 1.17, is obtained by the correct pathway ($A + B \rightarrow AB$, $AB + B \rightarrow AB_2$). The next best $\chi^2$ scores, averaging 1.28 and 4.49, are obtained by alternative three-stage pathways that add weak association reactions $AB_2 + B \rightarrow AB_3$ or $AB_2 + A \rightarrow A_2B_2$ to the correct pathway. As before, larger changes in the MSMRD scores are seen. The first alternative (adding $AB_3$) has an MSMRD score that is more than 40 times higher than the best MSMRD score ($6.09 \times 10^{-5}$, compared with $1.33 \times 10^{-6}$ for the correct pathway). The second alternative (adding $A_2B_2$) has an MSMRD score ($2.50 \times 10^{-3}$) that is almost 2000 times worse. Furthermore, comparing the best $\chi^2$ association constants against the best MSMRD constants in these alternative pathways reveals that they differ by approximately 100 in $K_1$ and $10^3$ in $K_2$. In addition, as before, neither alternative pathway has a set of constants that score well under both metrics. Thus, using $\chi^2$ and MSMRD scores together, the correct pathway can be determined.

### 3.2. Robustness to noise

The simulated data sets include a realistic estimate of Gaussian noise found in experimental data sets at third-generation synchrotron sources (Williamson *et al.*, 2008), but the present simulation framework enables easy assessment of how robust the method is to much noisier data. As one example, ten noisy data sets were generated for the one-stage bovine IFN-$\gamma$ with the resolution-dependent Gaussian noise scaled up by a factor of two. The correct $A + B \rightarrow AB$ pathway was still the clear winner in all the data sets. It achieved a very good fine-grid $\chi^2$ score (an average of 1.23 across ten data sets, compared with 1.49 with standard noise) at a nearly correct association constant ($3.40 \times 10^6$, the same as with standard noise and near the simulated value of $3.43 \times 10^6$). It also achieved a good fine-grid MSMRD score ($3.41 \times 10^{-11}$ compared with $1.82 \times 10^{-14}$) with a good association constant ($3.86 \times 10^6$).

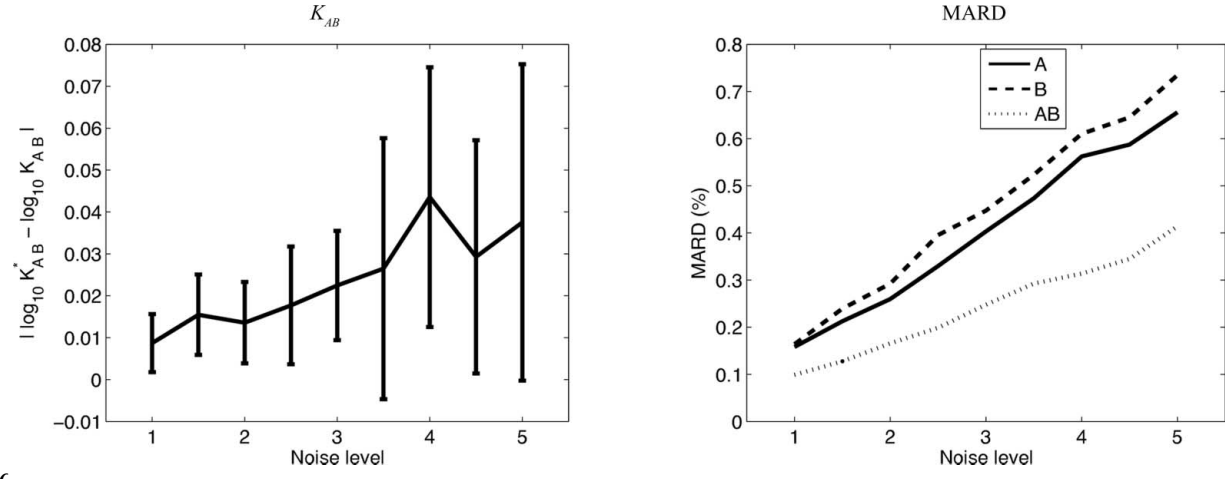


**Figure 6**
The effect of noise level (left) on the error in the association constant, assessed by the absolute difference in $\log_{10}K_{AB}$, and (right) on the reconstructed scattering curves, assessed by MARD. Values are for the bovine IFN-$\gamma$ best $\chi^2$ fine-grid point, averaged over ten data sets at each noise level. The association constant plot shows the means and standard deviations for the ten data sets at each noise level; only means are shown in the MARD plot, for clarity.

The performance of the method was then tested over a range of noise levels, increasing the Gaussian width up to five-fold, generating ten data sets for each noise level. The results were assessed in terms of identification of the association constant, as well as reconstruction of the underlying scattering curves of the monomers and oligomers. For association constants, the error was assessed with the absolute difference between the base-10 logs of the correct $K^*_{AB}$ and the inferred $K_{AB}$, *i.e.* $|\log_{10} K^*_{AB} - \log_{10} K_{AB}|$. For scattering curves, the evaluation is the MARD discussed above. Fig. 6 illustrates these error measures with respect to increasing noise (averaged over the ten data sets for each level). The figure shows that, as the noise increases, the best fine-grid points and reconstructions gradually become further away from the correct ones. Even at five times the noise, the errors in the association constants remain acceptable, approaching 10% (averaged across ten data sets), while the MARD values remain under 1% (0.6% for *A*, 0.7% for *B* and 0.4% for *AB*, averaged across ten data sets). Thus, we conclude that the method is indeed robust to such random noise. Robustness to some aspects of systematic noise (contamination with non-participating molecules) is discussed below.

### 3.3. Robustness across ranges of association constants

The ability of the method to recover the contribution from a particular species depends on that species making a non-negligible contribution to the mixture scattering data. That in turn depends on the association constants. The present simulations used physiologically reasonable constants, selected to ensure non-negligible quantities of each molecular species at equilibrium. However, since there is a wide range of reasonable values for weak association, a set of one- and two-stage simulations was conducted with varying association constant pairs to assess the range of values suitable for the method. For each association constant or pair of association constants, the simulated value was compared with the best $\chi^2$ constants (results with MSMRD are similar and are not shown).

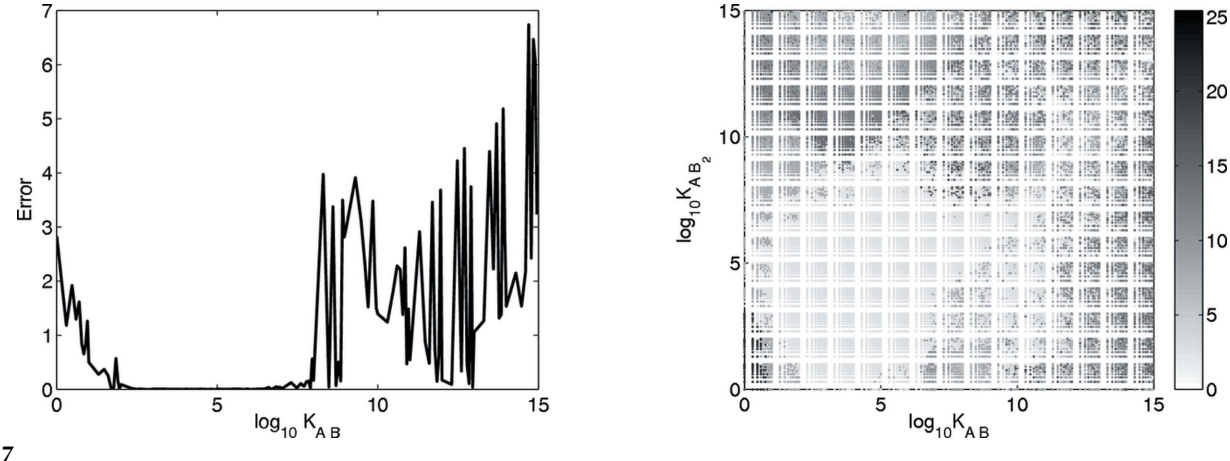


**Figure 7**
The error in inferring the simulated association constant for (left) one-stage bovine IFN-$\gamma$ and (right) two-stage BAF–emerin complex. The error for an association constant is the absolute log difference between the simulated and inferred association constants; for the two-stage case, the overall error is the square root of the sum of the squared errors.

Absolute log differences were used to assess the differences between the simulated and inferred values. For two association constants, the Euclidean distance $d_E$ was evaluated

$$d_E = \Big[(\log_{10} K^*_{AB} - \log_{10} K_{AB})^2 + \left(\log_{10} K^*_{AB_2} - \log_{10} K_{AB_2}\right)^2\Big]^{1/2}. \quad (19)$$

Fig. 7 shows the error over the range of association constant(s). For the one-stage bovine IFN-$\gamma$, the present method works best for values of $K_{AB}$ between 100 and $10^8$. For the two-stage BAF–emerin complex, the method works best (*i.e.* has an absolute log difference of around 2 or less) for most combinations over a broad range of $K_{AB}$ values between 10 and $10^{11}$ and $K_{AB_2}$ values between 100 and $10^9$. Poor scores for the one-stage association at low and high $K_{AB}$ values can be attributed to near-zero fractional masses of the initial or final components at those extremes. Likewise, for the two-stage association, poor scores for low $K_{AB}$ values can be attributed to the near-zero fractional mass of $AB$ (and hence $AB_2$) in such cases. The error is also large with high $K_{AB_2}$ values, owing to the very small amount of $AB$ remaining at equilibrium.

### 3.4. Robustness to monomers and complex size and shape

The performance of the present method was also studied on two other complexes that are quite different in molecular weight and structure from the two that have been discussed so far. While the main one-stage study, bovine IFN-$\gamma$, has monomers that are relatively small and close in molecular weight (14.2 and 13.3 kDa), the additional study, human calcineurin, has monomers that are larger and have very different molecular weights (43.6 and 18.8 kDa) and shapes. The main two-stage study, BAF–emerin complex, has monomers with weights of 5.7 and 10.1 kDa, while the additional study, HGH-receptor complex, has monomers with weights of 21.0 and 22.5 kDa and different shapes.

In both cases, the present method inferred the correct pathway and association constants and reconstructed scattering curves that are very similar to the simulated ones. For the one-stage human calcineurin (Supplementary Table 5), the $\chi^2$ value averaged 1.08 over ten simulated data sets, with association constants averaging $4.24 \times 10^4$ (which was the simulated value). The resulting MARDs for the best $\chi^2$ association constant averaged 0.24% for $A$, 0.16% for $B$ and 0.22% for $AB$. As in our initial one-stage study, an alternative two-stage model yielding both $AB$ and $AB_2$ scored well by $\chi^2$ (1.28) but poorly by MSMRD ($1.06 \times 10^{-3}$), with substantial disagreement on the best association constants ($K_{AB} = 4.30 \times 10^4$, $K_{AB_2} = 5.35 \times 10^2$ for $\chi^2$, and $K_{AB} = 3.00 \times 10^4$, $K_{AB_2} = 5.10 \times 10^2$ for MSMRD). The $\chi^2$ score at the best MSMRD point and the MSMRD score at the best $\chi^2$ point were also worse. Several other pathways scored moderately well by $\chi^2$, but all of these could be eliminated by evaluating the MSMRD scores and the disagreement between the best association constants.

Similarly good results were seen for the HGH-receptor complex (Supplementary Table 6). The lowest $\chi^2$ was on average 0.98 at association constants averaging $K_{AB} = 8.43 \times 10^5$, $K_{AB_2} = 6.26 \times 10^4$ (which were the simulated values). The average MARDs across ten data sets at the lowest $\chi^2$ points were 0.08% for $A$, 0.09% for $B$, 0.10% for $AB$ and 0.10% for $AB_2$. Alternative models that extend the correct two-stage pathway with $AB_2 + A \rightarrow A_2B_2$ or $AB_2 + B \rightarrow AB_3$ third stages also have low $\chi^2$ scores (1.38 and 1.33, respectively), but poorer MSMRD scores and large disagreement on the best association constants.

### 3.5. Contaminated data

A frequent problem in the analysis of associating systems is the presence of 'incompetent protein' contaminants, either monomer protein that behaves similarly to ideal material during purification but does not participate in associations, or oligomers that do not dissociate (irreversible aggregates) (Xu, 2004). In both cases the protein appears in the initial concentrations but not in any complex. For example, we found in our previous work on homo-oligomers that the addition of 2% of another oligomeric form would lead to large $\chi^2$ values and incorrect association constants and reconstructions (Williamson *et al.*, 2008).

To test the robustness of the present method to such contaminants, a nonparticipating fraction of monomer $A$ was used as a contaminant in the one-stage bovine IFN-$\gamma$. Also, a nonparticipating $A_{13}$ aggregate was used in the two-stage BAF–emerin complex, using a single aggregated form to represent the total possible contribution from multiple aggregated forms. To construct an $A_{13}$ structure for this simulation, copies of $A$ were repeatedly docked together using the software *GRAMM-X* (Tovchigrechko & Vakser, 2006). Scattering curves from all forms were again simulated using *CRYSOL*. Data were simulated with off-grid values of 0.0047, 0.0113 and 0.0231 contaminant mass fraction in the initial mass of $A$, using the same association constants as before. Ten data sets were generated for each case with different random Gaussian noise.

First, the regular coarse- and fine-grid searches were performed on the simulated data with contaminants, assuming as in previous sections the absence of any contaminant (Supplementary Tables 7 and 8). All of the alternative (incorrect) association pathways were immediately eliminated owing to high $\chi^2$ or inconsistency between best $\chi^2$ and best MSMRD (not shown).

Using the correct association pathway for bovine IFN-$\gamma$, the $\chi^2$ values increase monotonically with contaminant fraction. As expected, in the presence of a nonparticipating monomer, the apparent association constants also shift towards smaller values. When the contaminant fraction increases to 0.0231 the $\chi^2$ score more than doubles, indicating a clear problem in the analysis. The MSMRD scores also increase significantly (although these scores do not have a standard baseline to reference).

The behavior of the BAF–emerin complex is similar. The $\chi^2$ scores also increase monotonically with contaminant fractions. The behavior of the MSMRD score is more variable, perhaps

**Table 2**
Fine-grid $\chi^2$ results for contaminated simulations with coarse-grid $\chi^2$ within 1.0 of the lowest scoring model.

Simulated association constants: Bovine IFN-$\gamma$, $K_1 = 3.43 \times 10^6$; BAF–emerin complex, $K_1 = 3.21 \times 10^5$, $K_2 = 4.23 \times 10^5$.

| Contamination | $K_1$ | $K_2$ | $\chi^2$ | $c_A$† | $c_B$† |
|---|---|---|---|---|---|
| Bovine IFN-$\gamma$, $A + B \rightarrow AB$ | | | | | |
| 0.0000 | $4.80 \times 10^6 \pm 6.3 \times 10^5$ | | $1.61 \pm 0.1$ | $4.10 \times 10^{-3} \pm 6.0 \times 10^{-4}$ (×9) | $2.00 \times 10^{-3} \pm 0.0$ (×1) |
| 0.0047 | $3.59 \times 10^6 \pm 3.6 \times 10^5$ | | $1.51 \pm 0.1$ | $5.30 \times 10^{-3} \pm 9.5 \times 10^{-4}$ | N/A |
| 0.0113 | $3.41 \times 10^6 \pm 1.3 \times 10^5$ | | $1.43 \pm 0.1$ | $1.13 \times 10^{-2} \pm 6.7 \times 10^{-4}$ | N/A |
| 0.0231 | $3.39 \times 10^6 \pm 1.7 \times 10^5$ | | $1.46 \pm 0.1$ | $2.34 \times 10^{-2} \pm 5.2 \times 10^{-4}$ | N/A |
| Bovine IFN-$\gamma$, $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ | | | | | |
| 0.0000 | $3.94 \times 10^{13} \pm 9.6 \times 10^{13}$ | $1.01 \times 10^{11} \pm 2.0 \times 10^{11}$ | $1.51 \pm 0.1$ | $5.18 \times 10^{-2} \pm 4.2 \times 10^{-2}$ (×6) | $1.08 \times 10^{-2} \pm 6.8 \times 10^{-3}$ (×4) |
| 0.0047 | $8.81 \times 10^{14} \pm 2.5 \times 10^{15}$ | $2.70 \times 10^{12} \pm 7.6 \times 10^{12}$ | $1.45 \pm 0.2$ | $5.39 \times 10^{-2} \pm 4.2 \times 10^{-2}$ | N/A |
| 0.0113 | $5.39 \times 10^{12} \pm 1.1 \times 10^{13}$ | $1.55 \times 10^9 \pm 3.2 \times 10^9$ | $1.42 \pm 0.2$ | $1.37 \times 10^{-2} \pm 8.6 \times 10^{-3}$ | N/A |
| 0.0231 | $1.26 \times 10^{14} \pm 2.6 \times 10^{14}$ | $1.94 \times 10^{11} \pm 4.1 \times 10^{11}$ | $1.46 \pm 0.1$ | $3.11 \times 10^{-2} \pm 1.6 \times 10^{-2}$ | N/A |
| Bovine IFN-$\gamma$, $A + B \rightarrow AB$, $AB + A \rightarrow A_2B$ | | | | | |
| 0.0000 | $1.02 \times 10^{10} \pm 3.2 \times 10^{10}$ | $1.33 \times 10^6 \pm 3.1 \times 10^6$ | $1.60 \pm 0.1$ | N/A | $1.70 \times 10^{-2} \pm 2.2 \times 10^{-2}$ |
| 0.0047 | $3.91 \times 10^{12} \pm 1.2 \times 10^{13}$ | $6.63 \times 10^9 \pm 2.1 \times 10^{10}$ | $1.55 \pm 0.3$ | N/A | $1.66 \times 10^{-2} \pm 2.3 \times 10^{-2}$ |
| 0.0113 | $7.25 \times 10^6 \pm 5.1 \times 10^6$ | $6.65 \times 10^3 \pm 9.1 \times 10^3$ | $1.40 \pm 0.0$ | $1.04 \times 10^{-3} \pm 8.4 \times 10^{-4}$ | N/A |
| 0.0231 | $4.59 \times 10^{10} \pm 1.5 \times 10^{11}$ | $8.75 \times 10^5 \pm 2.7 \times 10^6$ | $1.19 \pm 0.4$ | $2.24 \times 10^{-2} \pm 7.3 \times 10^{-4}$ (×9) | $8.00 \times 10^{-2} \pm 0.0$ (×1) |
| BAF–emerin complex, $A + B \rightarrow AB$, $AB + B \rightarrow AB_2$ | | | | | |
| 0.0000 | $5.44 \times 10^5 \pm 4.3 \times 10^3$ | $8.00 \times 10^5 \pm 0.0$ | $1.78 \pm 0.1$ | N/A | $6.60 \times 10^{-3} \pm 5.2 \times 10^{-4}$ |
| 0.0047 | $6.87 \times 10^5 \pm 9.2 \times 10^5$ | $1.04 \times 10^6 \pm 1.6 \times 10^6$ | $1.74 \pm 0.0$ | $1.00 \times 10^{-2} \pm 0.0$ (×8) | $8.50 \times 10^{-3} \pm 7.1 \times 10^{-4}$ (×2) |
| 0.0113 | $3.13 \times 10^5 \pm 1.2 \times 10^4$ | $4.08 \times 10^5 \pm 2.2 \times 10^4$ | $1.49 \pm 0.1$ | $1.18 \times 10^{-2} \pm 9.2 \times 10^{-4}$ | N/A |
| 0.0231 | $3.60 \times 10^5 \pm 1.7 \times 10^4$ | $4.92 \times 10^5 \pm 3.0 \times 10^4$ | $1.56 \pm 0.1$ | $2.09 \times 10^{-2} \pm 8.8 \times 10^{-4}$ | N/A |

† The search considers only $A$ or $B$ contaminant. Rows with values for both $c_A$ and $c_B$ are the result of different identified contaminants for different simulations (number of times in parentheses).

because the $A_{13}$ contaminant used here has a disproportionate effect on the $I(0)$ values. For the 0.0231 contaminant fraction, even the coarse-grid search is unable to identify the nearest grid point. Here again, a significantly increased $\chi^2$ and disagreement between the best $\chi^2$ and best MSMRD association constants indicates problems for the 0.0113 and 0.0231 contaminant fractions. The increasing presence of the $A_{13}$ contaminant shifts the association constants to larger values, forming more of the larger complexes.

In both cases, the presence (or suspicion) of an incorrect analysis (particularly disagreement between the best $\chi^2$ and best MSMRD values) signals the need for a more sophisticated analysis. We have developed a convex quadratic optimization method specifically to deal with problems arising from nonparticipating contaminants.

Grid searches extended to include a contaminant fraction were performed for all cases. The coarse contaminant fraction grid dimension ranged from 0 to 0.1 in steps of 0.01. Fine-grid searches (including contaminant fraction) were then performed for all pathways with a $\chi^2$ value for the extended coarse-grid search within 1.0 of the best $\chi^2$ pathway (note that the MSMRD value cannot be used to assess the quality of these searches because scattering curves are only generated upon applying the quadratic optimization). The fine

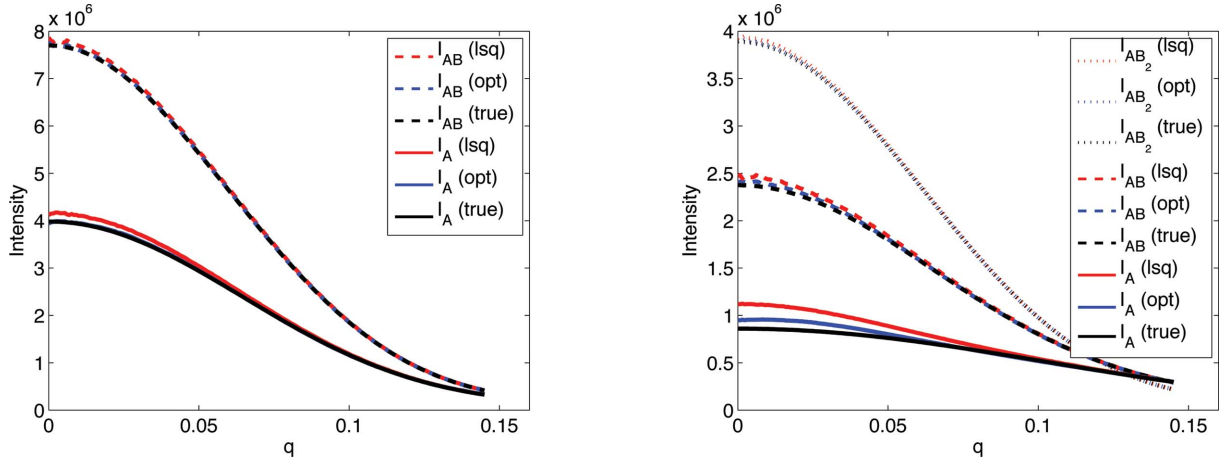


**Figure 8**
Simulated intensities compared with reconstructed ones computed by the quadratic program (opt) and the initial least squares $\tilde{O}_0$ (lsq), for one 0.0231 contaminant fraction data set of (left) bovine IFN-$\gamma$ and (right) BAF–emerin complex. The $I_B$ reconstruction, which is independent of contaminant, is not shown.

**Table 3**
MARDs (%) for contaminated reconstructions.

| Contaminant fraction | Method | $I_A$ | $I_B$ | $I_{AB}$ | $I_{AB_2}$ |
|---|---|---|---|---|---|
| Bovine IFN-$\gamma$ | | | | | |
| 0.0000 | LSQ | 0.83 ± 0.2 | 0.28 ± 0.2 | 0.41 ± 0.1 | – |
| | OPT | 0.27 ± 0.1 | 0.24 ± 0.1 | 0.22 ± 0.1 | – |
| 0.0047 | LSQ | 0.53 ± 0.1 | 0.17 ± 0.0 | 0.17 ± 0.0 | – |
| | OPT | 0.20 ± 0.0 | 0.17 ± 0.0 | 0.09 ± 0.0 | – |
| 0.0113 | LSQ | 1.13 ± 0.1 | 0.16 ± 0.0 | 0.34 ± 0.0 | – |
| | OPT | 0.42 ± 0.0 | 0.16 ± 0.0 | 0.14 ± 0.0 | – |
| 0.0231 | LSQ | 2.28 ± 0.1 | 0.17 ± 0.0 | 0.69 ± 0.0 | – |
| | OPT | 0.83 ± 0.0 | 0.17 ± 0.0 | 0.29 ± 0.0 | – |
| BAF–emerin complex | | | | | |
| 0.0000 | LSQ | 0.08 ± 0.0 | 0.64 ± 0.1 | 0.20 ± 0.0 | 0.20 ± 0.0 |
| | OPT | Not feasible | | | |
| 0.0047 | LSQ | 2.32 ± 0.0 | 0.08 ± 0.0 | 0.85 ± 0.0 | 0.56 ± 0.0 |
| | OPT | 1.91 ± 0.0 | 0.08 ± 0.0 | 0.77 ± 0.1 | 0.53 ± 0.0 |
| 0.0113 | LSQ | 4.45 ± 0.1 | 0.08 ± 0.0 | 0.90 ± 0.1 | 0.42 ± 0.1 |
| | OPT | 2.41 ± 0.2 | 0.08 ± 0.0 | 0.56 ± 0.1 | 0.27 ± 0.1 |
| 0.0231 | LSQ | 8.72 ± 0.1 | 0.08 ± 0.0 | 1.41 ± 0.1 | 0.60 ± 0.1 |
| | OPT | 1.31 ± 0.1 | 0.08 ± 0.0 | 0.31 ± 0.1 | 0.21 ± 0.0 |

contaminant grid then ranged from the point below the identified coarse-grid contaminant fraction to that above it, with a step size of 0.001. The grid searches were performed considering either an $A$ or $B$ homo-oligomeric contaminant (but not both). Optimized scattering intensities were then computed for the best $\chi^2$ fine-grid association constants by solving the quadratic program with constraints and parameter values as presented in *Methods*, §2.

Table 2 summarizes the fine-grid contaminant search results. For the one-stage bovine IFN-$\gamma$ contaminated with nonparticipating $A$, three pathways passed the $\chi^2$ cut-off: the correct model and the same two alternatives that were found in the baseline studies. While it is hard to distinguish the three solely on the basis of $\chi^2$, the intensity reconstruction optimization procedure found no feasible solution for the alternative models but successfully yielded scattering curves for the correct model, in all ten data sets. For the two-stage BAF–emerin complex contaminated with the $A_{13}$ aggregate, only the correct model passed the $\chi^2$ filter and its intensity reconstruction optimization was successful. For both cases and at all contaminant levels, the identified fine-grid association constants and contaminant fractions are close to the simulated values (bovine IFN-$\gamma$: $K_1 = 3.43 \times 10^6$; BAF–emerin complex: $K_1 = 3.21 \times 10^5$ and $K_2 = 4.23 \times 10^5$) and, for the higher contaminant fractions, notably closer than the values obtained in the contaminant-free searches.

Scattering intensities optimized using the quadratic program (labeled OPT/opt) were compared with simulated intensities (labeled TRUE/true) and those computed by least squares (labeled LSQ/lsq), both visually (Fig. 8) and by calculating MARD (Table 3). Here the quadratic program is consistently successful. MARD scores are substantially improved for the optimized reconstructions, with the greatest improvement at the higher contaminant fractions, although even the lower ones benefit, presumably as a result of the added constraints. Examining the scattering curve reveals that the greatest deviations from the simulated data and the greatest improvement come at small $q$ values. Note that $I_B$ is an independent vector in the intensity matrix, and thus the MARD values are the same for the two methods. The reconstructed scattering curve for the contaminating molecule (not shown) was not a close approximation to the true curve, probably because of the extremely small fraction of contaminant in the solution.

As a final test, contaminant grid searches were carried out on uncontaminated data (0.0000 entries in Tables 2 and 3). This approach did not perform as well as the contaminant-free search on uncontaminated data. As expected, fitting the additional contaminant parameter drives the association constants somewhat away from their best values.

### 3.6. Application of contaminant methods to homo-oligomers

Contamination with aggregates proved to be a problem for our earlier method for characterizing homo-oligomers (Williamson *et al.*, 2008). Thus, the present contaminant search and reconstruction were performed on the case studied previously: octameric purE from *Escherichia coli* (PDB code 1qcz; Mathews *et al.*, 1999), under a monomer–tetramer–octamer association with a 2% mass fraction of a 16-mer as contaminant. The best association constants resulting from the contaminant search were $K_{12} = 4.00 \times 10^{12}$, $K_{23} = 1.25 \times 10^1$, close to the simulated association constants $K_{12} = 2.87 \times 10^{12}$, $K_{23} = 1.29 \times 10^1$, although the identified contaminant fraction was higher than simulated, at 6.6%. The association model found by the previous method (Williamson *et al.*, 2008) was $K_{12} = 3.46 \times 10^{12}$, $K_{23} = 1.00 \times 10^1$, also close to the simulated association constants. However, the present reconstructed monomer scattering curve is much better than the previous one, whose $\chi^2$ is four times worse. The optimized monomer intensity curve is much closer to the simulated curve than that computed by least squares (after a contaminant search) or that found without contaminant search [as done by Williamson *et*

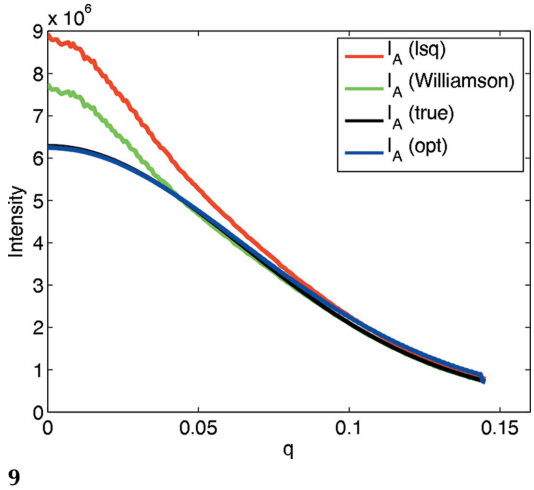


**Figure 9**
Reconstructed pure monomer intensity from a monomer–tetramer–octamer association contaminated with a 16-mer. $I_A$ (Williamson) is computed using the original grid search with no contaminant fraction and subsequent intensity reconstruction, as done by Williamson *et al.* (2008).

*al.* (2008)], especially at low $q$ (Fig. 9). Thus, the contaminant search plus the quadratic program reconstruction produce a curve that closely approximates the true one, while the contaminant-free and least-squares reconstructions introduce substantial error. Note again that the least-squares curve is just one of the infinitely many satisfying solutions, and thus it is not too surprising that it is actually much worse. The curves for the tetramer and octamer are not plotted, since for both methods they are extremely similar to the true curves. These results demonstrate that the present method can also be profitably applied to homo-oligomers in the presence of contaminants.

## 4. Discussion

We have presented a method to infer an association model (pathway and association constants), along with the underlying scattering curves of the initial components and intermediate and final complexes, from solution scattering data for a set of equilibrium mixtures undergoing hetero-association with different initial component concentrations. The method searches over possible association models and contaminant fractions, reconstructing the underlying scattering curves either by a least-squares method in the absence of 'incompetent protein' contaminants or by a convex quadratic program in their presence. The model and scattering curves are evaluated in terms of how well they can then reconstruct de-noised input data. Two complementary scores are used: a $\chi^2$ to assess the overall fit between the data and the association model combined with reconstructed scattering, and the MSMRD to assess the consistency between the association model stoichiometry and the reconstructed scattering. The convex quadratic program provides an optimization-based method for the difficult problem of reconstructing the underlying scattering curves in the presence of either nonparticipating monomers or irreversible aggregates.

In a variety of simulated test cases covering one- and two-stage association pathways, this approach correctly determined the pathway, accurately estimated the association constants with generally less than 2% error and accurately reconstructed the scattering curves to within an average deviation of less than 0.25%. While such accuracy cannot be expected for all experimental scattering data, the potential for such accurate evaluation exists in the most favorable cases. The good accuracy for reconstructing the scattering curve bodes well for the application of three-dimensional structural modeling based on the reconstructed scattering curves. The $\chi^2$ and MSMRD were found to be effective as complementary metrics. Cases where an alternative model with an extra association step obtained a fairly good $\chi^2$ value could be ruled out by a greater MSMRD and inconsistency between the best-scoring association constants under one metric *versus* the other. The method was also found to be amenable to a range of association constants, Gaussian noise levels, different complex sizes and shapes, and contaminants.

The range of association constants that were found acceptable for the method (Fig. 7) compares well with the range of $10^4$–$10^9$ routinely available from analytical ultracentrifugation (Lebowitz *et al.*, 2002), while also revealing the molecular weight of each complex [*via* $I(0)$ calculations] calibrated by the molecular weights of the initial components. At the same time, the SAS method provides complex scattering curves that can serve as the basis for three-dimensional reconstruction. In addition, this range of affinities is explored with the same fixed set of initial concentrations used in the earlier simulation. The initial concentrations could also be adjusted upwards to explore weaker interactions (limited by the solubility of the proteins) and downwards to explore stronger ones (limited by the strength of observed scattering). The strongest beamlines at third-generation sources can generate accurate scattering profiles at concentrations as low as 0.05 mg ml$^{-1}$ (Williamson & Friedman, unpublished results), a fact that also aids in the reduction of noise from interparticle interference (see below).

At realistic contaminant levels, the present method is able to reconstruct the scattering curves quite accurately, a result not possible by previous methods which assumed an absence of contaminants. While by no means perfect, the objective and set of constraints chosen here yield good solutions in practice. Smoothness is taken as the primary objective, and the potential for over-smoothing is mitigated by a counter-balancing constraint from the $\chi^2$ constraint. Other constraints could potentially be incorporated in order to encode shape characteristics and relationships between the different forms. It is not possible to determine adequately the exact contaminant fraction or its scattering curve, but the incorporation of additional constraints could help. Extensions to other forms of contamination and systematic noise may be amenable to analogous techniques.

As discussed in the *Introduction*, we have focused only on the contributions from the modeled molecular species – initial components, higher-affinity intermediate and final complexes, and possibly static contaminants. Experimental scattering data also contain contributions from interparticle interference, arising from the lowest affinity, typically most transient, protein–protein complexes. A study of and extension to handle interparticle interference remains very interesting future work, which is likely to increase the power and applicability of this approach. There are several possible ways in which the method could be extended to account for this non-ideality. Some weak interparticle interactions of a different stoichiometry from the primary modeled association may become factored out as noise in the low-rank approximation or as residuals in the modeling. Other interparticle interactions may be captured as a form of explicit contaminant. Alternatively, data from dilution series towards zero concentrations (where these interactions become vanishingly small) could be collected and incorporated into the model. When the weak interactions are of the same stoichiometry as the modeled associations, they are linearly inseparable and thus cannot be directly accounted for in a 'bottom-up' analysis like that presented here. However, 'top-down' structural information could be exploited to constrain the scattering curves according to the structural characteristics of the monomers and complexes, a modification of our approach for

NMR (Potluri *et al.*, 2006; Martin *et al.*, 2011; Chandola *et al.*, 2011). Perhaps an iterative approach could even be employed, starting with an initial factorization as presented here, and then iteratively alternating between inferring a structure based on the current factorization and improving the derived model and curves based on the current structural information.

In the test cases, pure $A$ and pure $B$ were included as two of the samples. This suggests an alternative strategy to use the intensity curves of these pure samples to reduce the number of unknowns (removing known intensity column vectors for $A$ and $B$ in $\tilde{O}$) in the computations. However, when contaminants are present, there may be no such thing as a 'pure' sample. Likewise, this approach works with a self-associating system which does not contain pure monomers even at the lowest concentration. In preliminary studies (not shown), we have found that, even without using pure $A$ and pure $B$ in the set of samples, the correct model can still be obtained as long as the equilibrium mixtures contain sufficiently diverse concentrations of species.

This work was supported in part by the National Science Foundation (grant No. IIS-0502801 to CBK, AMF and BAC, and grant No. CCF-0915388 to CBK), along with the National Institutes of Health (grant No. R01 GM-65982 to Bruce Randall Donald, Duke University.

## References

Attri, A. K. & Minton, A. P. (2005). *Anal. Biochem.* **346**, 132–138.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernadó, P., Pérez, Y., Blobel, J., Fernández-Recio, J., Svergun, D. I. & Pons, M. (2009). *Protein Sci.* **18**, 716–726.
Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Chandola, H., Yan, A. K., Potluri, S., Donald, B. R. & Bailey-Kellogg, C. (2011). *J. Comput. Biol.* **12**, 1757–1775.
Chen, L., Hodgson, K. O. & Doniach, S. (1996). *J. Mol. Biol.* **261**, 658–671.
Codreanu, S. G., Thompson, L. C., Hachey, D. L., Dirr, H. W. & Armstrong, R. N. (2005). *Biochemistry*, **44**, 10605–10612.
Dervichian, D. G., Fournet, G. & Guinier, A. (1952). *Biochim. Biophys. Acta*, **8**, 145–149.
Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering.* New York: Plenum Press.
Guinier, A. & Fournet, G. (1955). *Small-Angle Scattering of X-rays.* New York: Wiley.
Kameyama, K. & Minton, A. P. (2006). *Biophys. J.* **90**, 2164–2169.
Lebowitz, J., Lewis, M. S. & Schuck, P. (2002). *Protein Sci.* **11**, 2067–2079.
Martin, J. W., Yan, A. K., Bailey-Kellogg, C., Zhou, P. & Donald, B. R. (2011). *Protein Sci.* **20**, 970–985.
Mathews, I. I., Kappock, T. J., Stubbe, J. & Ealick, S. E. (1999). *Structure*, **7**, 1395–1406.
Potluri, S., Yan, A. K., Chou, J. J., Donald, B. R. & Bailey-Kellogg, C. (2006). *Proteins*, **65**, 203–219.
Segel, D. J., Bachmann, A., Hofrichter, J., Hodgson, K. O., Doniach, S. & Kiefhaber, T. (1999). *J. Mol. Biol.* **288**, 489–499.
Segel, D. J., Fink, A. L., Hodgson, K. O. & Doniach, S. (1998). *Biochemistry*, **37**, 12443–12451.
Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). *Biophys. J.* **80**, 2946–2953.
Svergun, D. I. & Stuhrmann, H. B. (1991). *Acta Cryst.* A**47**, 736–744.
Tovchigrechko, A. & Vakser, I. A. (2006). *Nucleic Acids Res.* **34**, W310–W314.
Velazquez-Campoy, A., Leavitt, S. A. & Freire, E. (2004). *Methods Mol. Biol.* **261**, 35–54.
Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.
Williamson, T. E., Craig, B. A., Kondrashkina, E., Bailey-Kellogg, C. & Friedman, A. M. (2008). *Biophys. J.* **94**, 4906–4923.
Xu, Y. (2004). *Biophys. Chem.* **108**, 141–163.