Please cite this article in press as: Bowerman et al., BEES: Bayesian Ensemble Estimation from SAS, Biophysical Journal (2019), https://doi.org/10.1016/j.bpj.2019.06.024

Biophysical Journal
Computational Tool

Biophysical Society

BEES: Bayesian Ensemble Estimation from SAS

Samuel Bowerman,[1,*] Joseph E. Curtis,[2] Joseph Clayton,[1] Emre H. Brookes,[3] and Jeff Wereszczynski[1]
[1]Department of Physics and the Center for Molecular Study of Condensed Soft Matter, Illinois Institute of Technology, Chicago, Illinois; [2]National Institute of Standards and Technology Center for Neutron Research, National Institute of Standards and Technology, Gaithersburg, Maryland; and [3]University of Texas Health Science Center, San Antonio, Texas

ABSTRACT Many biomolecular complexes exist in a flexible ensemble of states in solution that is necessary to perform their biological function. Small-angle scattering (SAS) measurements are a popular method for characterizing these flexible molecules because of their relative ease of use and their ability to simultaneously probe the full ensemble of states. However, SAS data is typically low dimensional and difficult to interpret without the assistance of additional structural models. In theory, experimental SAS curves can be reconstituted from a linear combination of theoretical models, although this procedure carries a significant risk of overfitting the inherently low-dimensional SAS data. Previously, we developed a Bayesian-based method for fitting ensembles of model structures to experimental SAS data that rigorously avoids overfitting. However, we have found that these methods can be difficult to incorporate into typical SAS modeling workflows, especially for users that are not experts in computational modeling. To this end, we present the Bayesian Ensemble Estimation from SAS (BEES) program. Two forks of BEES are available, the primary one existing as a module for the SASSIE web server and a developmental version that is a stand-alone Python program. BEES allows users to exhaustively sample ensemble models constructed from a library of theoretical states and to interactively analyze and compare each model's performance. The fitting routine also allows for secondary data sets to be supplied, thereby simultaneously fitting models to both SAS data as well as orthogonal information. The flexible ensemble of K63-linked ubiquitin trimers is presented as an example of BEES' capabilities.

SIGNIFICANCE Small-angle scattering (SAS) is an increasingly popular method for probing the solution ensemble of flexible biomolecules. However, the interpretation of SAS data is nontrivial as theoretical models must be both complex enough to interpret the SAS experiments yet not overfit the limited experimental data. Here, we present the Bayesian Ensemble Estimation from SAS program, a publicly accessible Python code for fitting SAS data while penalizing models against overfitting. The program is available as both a stand-alone code and as a module on the SASSIE web server, allowing users with all levels of computational experience and resources to make use of the program.

Q1 Q2

INTRODUCTION

Biological molecules rely heavily on their conformational dynamics to conduct their cellular function, and the characterization of these flexible ensembles of states remains a key challenge in modern biophysics (1). As a result, many different experimental and computational techniques have been developed to probe and model configurational ensembles. Of these, small-angle scattering (SAS) measurements are an increasingly popular technique because of their relative ease of use and their ability to simultaneously probe the full solution ensemble (2,3). Moreover, SAS measurements are able to probe systems at room temperature, free from packing forces induced by the lattice and cryogenic effects of crystallography, and they can measure the solution of states in both equilibrium ensembles and time-dependent processes (4), such as protein and RNA folding (5,6), or the allosteric coupling of enzymatic activity and large-scale domain movement (7,8). However, the low-dimensional nature of SAS data can often cause the interpretation of scattering profiles to be relatively difficult, and reconstituting a three-dimensional molecular structure solely from scattering curves can often be misleading because multiple reconstitutions of varying shapes may result from the same scattering profile.

In contrast, model structures can also be identified from all-atom or coarse-grained simulations, and their calculated

Submitted February 7, 2019, and accepted for publication June 20, 2019.

*Correspondence: jwereszc@iit.edu

Samuel Bowerman's present address is Department of Biochemistry and Howard Hughes Medical Institute, University of Colorado Boulder, Boulder, Colorado 80309.

Editor: Jill Trewhella.

https://doi.org/10.1016/j.bpj.2019.06.024

Please cite this article in press as: Bowerman et al., BEES: Bayesian Ensemble Estimation from SAS, Biophysical Journal (2019), https://doi.org/10.1016/j.bpj.2019.06.024

scattering profiles can be compared against empirical curves (9–12). Because SAS profiles are measurements of the full solution ensemble and therefore may not be fully described by a single structural state, these in silico profiles can also serve as a basis set to construct an ensemble model through a linear combination of states (13–16). Although this ensemble reconstitution approach is conceptually straightforward, in practice, it can be quite difficult to identify the "best" ensemble model. For instance, it is not known a priori what the number of underlying states should be in the ensemble. It is also possible for ensemble models to overfit experimental data through the inclusion of too many underlying populations. Furthermore, altogether different combinations of states may yield similarly performing models, in respect to their goodness-of-fit values.

For these reasons, a Bayesian-based approach has many advantages over more traditional methods. For instance, Markov Chain Monte Carlo posterior sampling methods will not only estimate model parameters but will also allow for the direct assessment of their errors (17). Moreover, Bayesian formalism allows for the comparison of a population of models as a solution to parameterization rather than only identifying a single set of parameters (18–21). This is exceptionally useful for SAS modeling, in which information regarding the model is underdetermined. However, the ability to construct a large population of solutions can also be a disadvantage because both the computational resources to construct a complete array of model parameters, as well as tools for comparing models, can be daunting for many systems.

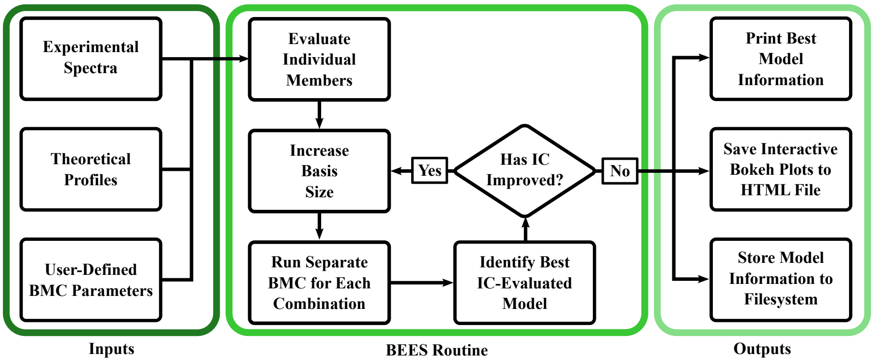
To this end, we previously developed an iterative Bayesian method to use SAS profiles, either of x rays (small-angle x-ray scattering [SAXS]) or of neutrons, to reweight the population of states from simulated models. This approach, which is an extension of the basis set-supported SAXS technique (13), compares solution ensembles of a variety of subensembles from a combination of potential scattering states. Originally, we used this method to fit ensembles of covalently linked ubiquitin trimers, and we observed that the algorithm could produce ensemble models that robustly resisted overfitting while also describing biochemically relevant behavior, including ensemble flexibility and burial or exposure of known ubiquitin-binding domain recognition sites, such as I36- or I44-centered hydrophobic patches (22–25).

Here, we present an update to this method as an open source program called Bayesian Ensemble Estimation from SAS (BEES). Two versions of this code have been developed. The primary version is an open-access module on the SASSIE web server (http://sassie-web.chem.utk.edu/sassie2/), which provides a graphical user interface for controlling the module (26,27). The BEES-SASSIE module is designed for users that are both new and experienced in biophysical modeling, and through SASSIE, it provides access to the computational resources required to calculate and analyze large combinations of states. The second, developmental version is a stand-alone Python code that is designed to be run from the command line and is intended for experienced computational scientists. We also provide two example use cases, one in which we fit profiles of K63-linked ubiquitin trimers to SAXS data alone and another in which we add a second data set to the fitting procedure.

MATERIALS AND METHODS

BEES algorithm

The BEES algorithm is designed to find the theoretical solution ensemble that uses the fewest number of populations to accurately describe the experimental data. This goal is similar to that of the Sparse Ensemble Search (SES) developed by Berlin et al. (25), but the BEES program uses a Bayesian Monte Carlo formalism (13), in contrast to the orthogonal matching pursuit employed by SES, to estimate the uncertainty of each fitted population weight. The BEES algorithm is briefly presented here (Fig. 1), but further details can be found in the supplemental text and elsewhere (22). In short, experimental data are gathered and postprocessed before using the BEES module. For example, users may wish to screen their data for low


Experimental Spectra
Theoretical Profiles
User-Defined BMC Parameters
Inputs
Evaluate Individual Members
Increase Basis Size
Yes
Has IC Improved?
No
Run Separate BMC for Each Combination
Identify Best IC-Evaluated Model
BEES Routine
Print Best Model Information
Save Interactive Bokeh Plots to HTML File
Store Model Information to Filesystem
Outputs

Web 3C

FIGURE 1 Workflow schematic of the BEES routine. Users supply empirical data and the collection of theoretical profiles for potential ensemble members as well as set several parameters associated with the Bayesian Monte Carlo (BMC) parameter search. After the performance of each individual theoretical state is evaluated, ensemble populations are fit by BMC routines conducted iteratively on increasing-sized subensembles until the addition of another member population does not improve the IC value and overfitting is observed. Alternatively, users can bypass the IC comparison step to compare all possible combinations of states. The routine then relays information regarding the resulting models to the command terminal (stand-alone version) or GUI (SASSIE web version) and further stores model information in several file locations for further review by users. To see this figure in color, go online.

Please cite this article in press as: Bowerman et al., BEES: Bayesian Ensemble Estimation from SAS, Biophysical Journal (2019), https://doi.org/10.1016/j.bpj.2019.06.024

Q-beam smearing effects or to extrapolate their scattering profile to I(0). Furthermore, a collection of candidate solution states are identified, possibly from a set of Protein Data Bank structures or a selection of structural states from molecular dynamics or Monte Carlo simulations. Theoretical profiles for these candidate solution states are then input to BEES, and they can be computed by stand-alone programs such as Crysol (28) or FoXS (29), in SASSIE via the "SasCalc" module (30), or from many other scattering prediction software (31–34).

Once initiated, the BEES routine reads the collection of theoretical scattering profiles for potential ensemble model members, and it first determines the goodness-of-fit values of each individual profile. It then identifies all possible sub-bases containing combinations of two theoretical profiles, and it conducts a Bayesian Monte Carlo routine on each combination to identify the population of states in each sub-basis. Each Monte Carlo routine is conducted using uniform priors and according to user-defined parameters: number of independent Monte Carlo parameter fittings per sub-basis, number of iterations per Monte Carlo fitting, and amount of population change per iteration. Notably, the BEES likelihood function (L) includes the ability to simultaneously fit the scattering profiles and an auxiliary set of measurements:

$$L = \mathrm{e}^{-\chi^2_{\mathrm{SAS}}/2.0} \bullet \mathrm{e}^{-\chi^2_{\mathrm{aux}}/2.0}. \tag{1}$$

Furthermore, the total model goodness of fit (χ^2_{total}) is calculated from the linear combination of the model scattering goodness of fit (χ^2_{SAS}) and the model goodness-of-fit to the auxiliary data set (χ^2_{aux}): $\chi^2_{total} = \chi^2_{SAS} + \chi^2_{aux}$. The likelihood function assumes normally distributed data and errors; therefore, users may be required to transform their auxiliary data before fitting with BEES to utilize alternative likelihood formalisms. For instance, users fitting measurements most appropriately described by a log-normal distribution may wish to express their auxiliary data as the logarithm of their measurements rather than attempting to directly fit the measurements themselves. However, it may not be possible for some forms of data (i.e., binomial data) to be transformed easily to a normal distribution. Because BEES is open source, users interested in fitting such data may modify the auxiliary likelihood function to a form more appropriate for their data.

BEES also allows the user to define the χ^2_{SAS} metric in several ways. First, the standard χ^2 value can be determined from each individual scattering intensity and its associated error. However, this metric does not account for the highly correlated and oversampled nature of an SAS curve. As a result, the χ^2_{free} metric of Rambo and Tainer can also be used, which reduces the number of independent scattering points to the Shannon sampling limit (35):

$$\chi^2_{free} = \mathrm{med}\left\{\sum_{\mathrm{q}\epsilon\mathrm{S}} \frac{\left(\mathrm{I_{obs}(q)} - \mathrm{I_{exp}(q)}\right)^2}{\sigma^2_{\mathrm{exp}}(\mathrm{q})}\right\}, \tag{2}$$

where S represents the randomly selected set of q values that are separated into the appropriate number of Shannon channels (determined by the experimentally observed $\mathrm{D_{max}}$ value), and the reported χ^2_{free} value is the median value determined from a large number of independently selected sets of values. Here, we have opted for 3001 sets of values from past experience with the metric. Use of the χ^2_{free} metric is the recommended mode of operation, but a third approach is available to the user in which the number of independent scattering points is defined a priori and is used to scale a reduced χ^2 value:

$$\chi^2_{SAS} = \frac{n}{N_q} \sum_{q=1}^{N_q} \frac{\left(I_{obs}(q) - I_{exp}(q)\right)^2}{\sigma^2_{exp}(q)}, \tag{3}$$

where N_q is the number of scattering points, and n is the number of user-defined independent scattering points.

Once the ensemble of states for each two-member sub-basis has been identified, the best two-member state is selected in accordance to the information criteria (IC) selected by the user (36–38) (see Comparing model performances with IC for more details). If the IC value of the best two-member state is worse than that of any single theoretical profile, then the module reports the best single profile as the most likely model. However, if the IC value of this two-member state is instead an improvement over all individual profiles, then the BEES module conducts the Bayesian Monte Carlo routine on every three-member sub-basis, and the best three-state IC value is similarly compared to the two-state ensemble. This iterative increase in sub-basis size and comparison of IC values is conducted until either the IC metric does not improve or every possible combination of states is considered. Alternatively, users also have the option to override the IC comparison and force the construction of all combinations of subensembles. Once the desired number of models has been identified, the BEES module will also calculate each model's "relative performance" metric to determine its likelihood over the best IC-identified model (Comparing model performances with IC) (39):

$$RP(m) = \mathrm{e}^{(\mathrm{IC_m} - \mathrm{IC_o})/2}, \tag{4}$$

where $RP(m)$ and IC_m are the relative performance and IC values of model m, and IC_o is the minimal IC value of all observed models. The relative performance metric is more commonly known as the relative likelihood of a model. Here, we opt for the changed nomenclature to assist nonexperts in the interpretation of the metric as well as to avoid confusion with the likelihood function used by the Bayesian Monte Carlo fitting routine. Although the relative performance provides a quantitative result, it is admittedly an approximation of the more rigorous Bayes factor (37,40). As such, it is intended to be interpreted loosely and to assist the user in applying their intuition toward the performance of alternative ensembles to the best identified one.

Once the best model has been identified, BEES outputs information regarding ensemble members of the IC-identified model, its model population weights, goodness-of-fit information for the full ensemble model and each individual, and the IC value of the model. Beyond the best identified model, information regarding every model identified for each sub-basis is also saved. Plots of the model ensemble fit to the experimental data, along with the associated residual errors, are automatically created once the fitting routine is completed. These plots are included in a multitab Hypertext Markup Language (HTML) page, which provides graphical and table presentations to allow users the ability to compare different models and performances.

Comparing model performances with IC

The rigorous comparison of theoretical ensembles to experimental data requires creating models that are rich enough to describe the underlying physical structures that generated the data while simultaneously avoiding overfitting. However, it is imperative that the final model does not achieve a strong goodness-of-fit value through inclusion of an arbitrary number of parameters (here, the number of scattering profiles). As a result, the BEES method enforces that the "best model" must be a balance between optimizing the goodness-of-fit metric and minimizing the number of underlying scattering states. To this end, the module utilizes IC to penalize model goodness-of-fit values according to their ensemble size. Users have the option to use one of three different IC metrics during fitting—the Akaike IC (AIC), the Bayesian IC (BIC), or the deviance IC (DIC) (36–38,41):

$$AIC = 2k - 2 \bullet \log\left(\widehat{L}\right), \tag{5}$$

$$BIC = \log(n) \bullet k - 2 \bullet \log\left(\widehat{L}\right), \tag{6}$$

$$DIC = 2P - 2 \bullet \log(L(\overline{w})). \tag{7}$$

Here, k is the number of model parameters (number of scattering states), $\widehat{L}$ is the maximal observed likelihood value during the Bayesian Monte

Carlo parameter fitting, n is the number of points in the experimental data set, $L(\overline{w})$ is the likelihood of the model from the posterior-averaged weights, and P is the estimated number of free parameters. In the DIC metric, P is determined by the difference between $\log(L(\overline{w}))$ and the average observed log $(L(w_i))$ over the course of the Monte Carlo iterations.

Each IC metric rewards models with improved experimental fits (higher values of $\widehat{L}$ and $L(\overline{w})$) and penalizes those with more free parameters (higher values of k and P). The BIC is closely related to the AIC; however, it is derived from Bayesian principles rather than the frequentist foundation of the AIC. In each metric, smaller values are indicative of better model performance, with the defining separation between them being the strength of the penalty term. In the AIC, the penalty is always double the number of states, whereas the BIC penalty will become increasingly larger for a larger number of data points. In reality, both metrics are an approximate way to identify the best model, and the AIC may be more prone to false positive estimations (including too many states), whereas the BIC metric may be more prone to false negatives (rejecting too many states), depending on the number of experimental data points. The penalty of the DIC for free parameters is directly tied to the variance of the sampled likelihood during the Monte Carlo routine, so it can either be more permissive or restrictive than the BIC or AIC, depending on the system. In the K63-linked ubiquitin trimer discussed in this manuscript, the DIC was found to be less discriminatory than the other IC metrics (Supporting Materials and Methods, Section S3). AIC and BIC may converge upon the same solution as is the case with the K63 example presented here.

The model with the minimal IC value can be interpreted as the most likely, best performing model. Although it may be tempting to accept this model and reject all others, there is a possibility that one of these other models might actually be more accurate to the true nature of the system, even though each one possesses a weaker IC value. The probability that a model is, in fact, a better assessment of the data can be calculated by comparing the model IC values to the lowest IC value, as previously stated (Eq. 4; (39)).

Because the BIC and AIC apply different penalties to the number of states, they may also produce different relative performance values for the same set of models. When fewer than seven independent data points are contained within the scattering curve, BIC-based calculations of relative model performances will yield distributions with more models being comparable to the lowest IC-selected one, in comparison to using the AIC metric for evaluating model quality. That is, if the number of independent data points is seven or fewer, then more models will have a relative performance closer to 1.0 than if evaluated by AIC. On the other hand, if the number of observed data points is greater than eight, then more models will have relative performances closer to 0.0 if they are evaluated by the BIC in place of the AIC. In the end, the choice of IC evaluation is up to the user, and it may sometimes be appropriate to use each of them to determine upper and lower bounds for relative model performances.

RESULTS

Here, we describe a sample usage of BEES and its resulting data. The necessary data files for this test set are included in the Supporting Materials and Methods. Users can thereby recreate the analyses presented here by unpacking the archive locally and uploading the relevant files for each case to the BEES module in the SASSIE web server or by following the shell scripts provided alongside the standalone version (https://github.com/WereszczynskiGroup/BEES/tree/master/examples). In the first example, we model the populations of states of K63-linked ubiquitin trimers using clusters identified from accelerated molecular dynamics trajectories (22). In the second example, we showcase the effects of simultaneously fitting the SAS spectra and an auxiliary data set by including simulated measurements of an interdomain distance and angle.

Building ensembles of SAS data

BEES requires the user to supply the experimental scattering curve along with theoretical scattering curves for candidate structures. In addition to providing this data, users must also define the D_{max} of the molecule, which can be determined from the experimental profile using pre-existing software (42). Here, a D_{max} of 83.6 Å was determined using the SHANUM program of the ATSAS package (43). Furthermore, five Monte Carlo walkers were used for each sub-basis ensemble, and each walker was conducted for 10,000 iterations. The first 1000 iterations were neglected when determining the model populations so as to remove any influence of the randomly selected initial values from the final result. Parallel processing can also be used (here, six processors were used), but using multiple processors will only enhance the speed of the calculation and has no effect on the final result (see Supporting Materials and Methods for more information). In addition, the full array of subensembles has been calculated to display the depth of analysis available. In this example, truncation of the algorithm via the IC parameter would save a significant amount of computational time without effecting the best IC-identified model; however, models with lower χ^2_{free} would not have been observed. At the conclusion of the BEES routine, the best identified model is reported (Fig. 2), and an interactive plot interface is created (Fig. 3).

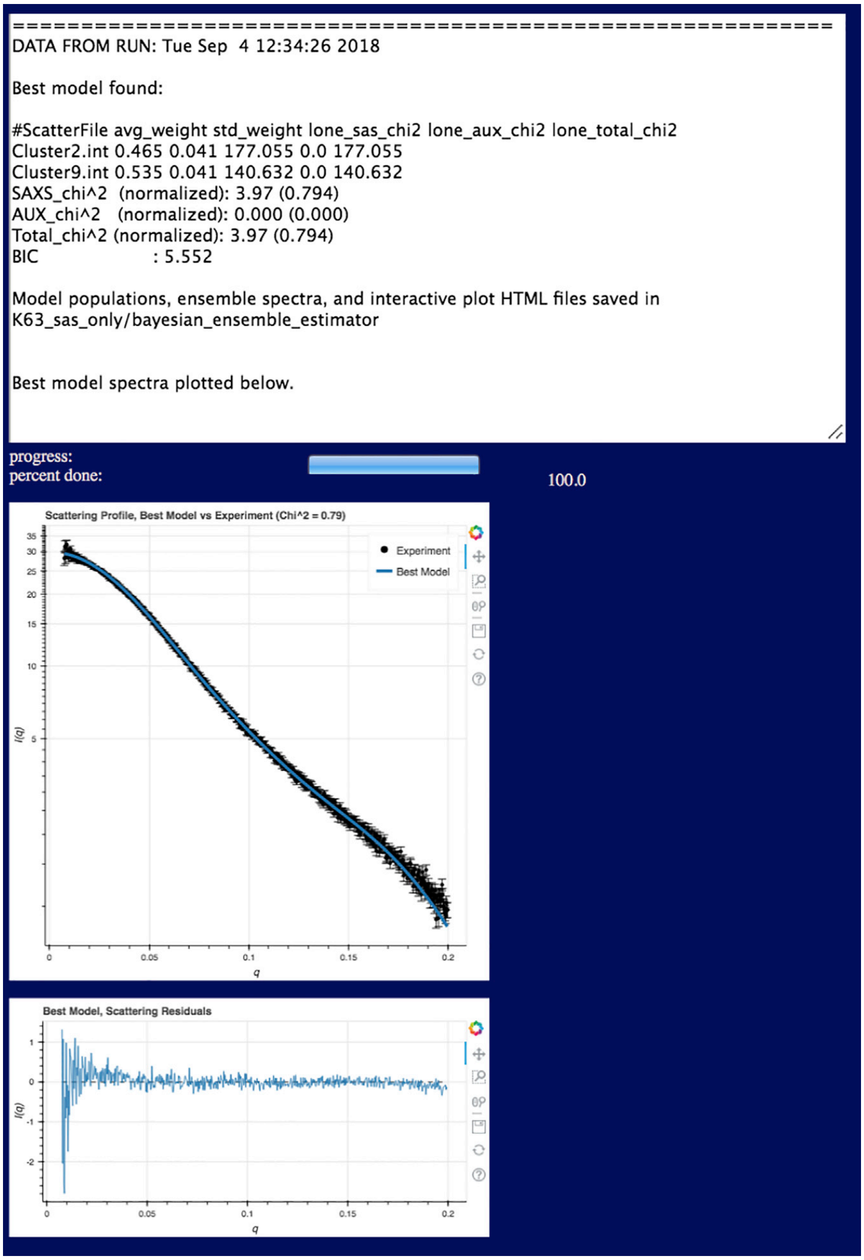
In this example, the best model is a two-state solution that is approximately equal parts clusters 2 and 9. This model has a χ^2_{free} of 0.79 and a BIC value of 5.55. Although this is the best model according to BIC comparisons, roughly 50 models of varying sizes possess better χ^2_{free} values, and the model with the best goodness of fit ($\chi^2_{free} = 0.74$) is a four-member state comprised of clusters 2 (~45%), 4 (~22%), 10 (~15%), and 11 (~18%). This lowest χ^2_{free} model has an IC value of 8.47, which yields a relative performance of 0.23 when compared to the IC-identified two-state model. As such, the improved χ^2_{free} value of this model is unwarranted as it is likely the result of overfitting by too many basis members. Indeed, inspection of the model performance histogram (Fig. 2 *B*, *top*) shows that the best performing models are largely two-state solutions, but some three-state solutions perform moderately well. Furthermore, many of the two- and three-state solutions are a significant improvement over each of the single-state models.

Building ensembles with auxiliary data

Some users may desire to use BEES to build theoretical solution states by fitting solely to SAS data and then use

Please cite this article in press as: Bowerman et al., BEES: Bayesian Ensemble Estimation from SAS, Biophysical Journal (2019), https://doi.org/10.1016/j.bpj.2019.06.024

Web 3C

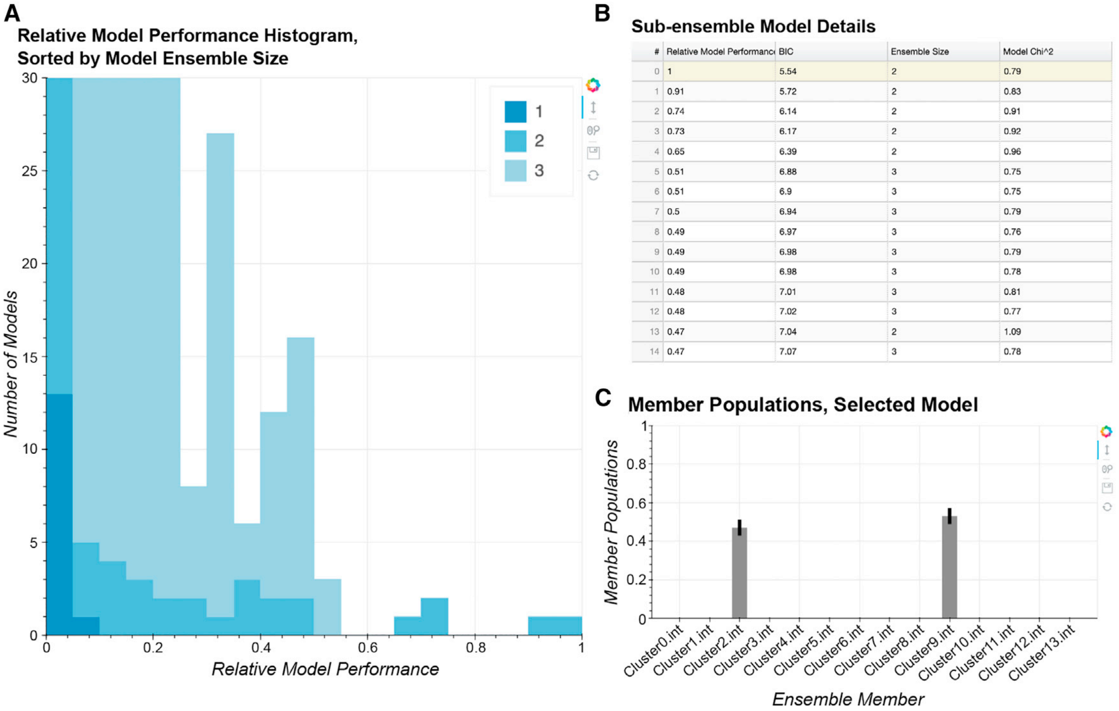


```
================================================================
DATA FROM RUN: Tue Sep  4 12:34:26 2018

Best model found:

#ScatterFile avg_weight std_weight lone_sas_chi2 lone_aux_chi2 lone_total_chi2
Cluster2.int 0.465 0.041 177.055 0.0 177.055
Cluster9.int 0.535 0.041 140.632 0.0 140.632
SAXS_chi^2  (normalized): 3.97 (0.794)
AUX_chi^2   (normalized): 0.000 (0.000)
Total_chi^2 (normalized): 3.97 (0.794)
BIC                     : 5.552

Model populations, ensemble spectra, and interactive plot HTML files saved in
K63_sas_only/bayesian_ensemble_estimator


Best model spectra plotted below.
```

FIGURE 2 Example of output from BEES program as used on SASSIE web GUI. (*Top*) Text output displaying the contributing populations of the best IC-identified ensemble and the associated error in population estimates as well as goodness of fit for each member. Total model goodness of fit and IC value are also printed by the module. (*Middle*) Ensemble scattering profile of the best identified model, shown in blue, is fit to the experimental spectrum, shown in black. (*Bottom*) Residual errors of the best model against the experiment are shown. To see this figure in color, go online.

these states to predict the measurements of future experiments. However, others may already possess such data and may prefer to create models that are consistent with both these measurements as well as the observed SAS profiles. For example, an experimenter may desire to simultaneously model both a scattering profile and a catalog of NMR-derived distances. For the benefit of this class of users, we have included this functionality within BEES. To demonstrate how including such data might affect the modeling results, we discuss here an extension of the previous triubquitin example in which we provide a simulated data set that contains the ensemble-averaged center-of-mass distance between distal monomers and the angle formed by the trimer arrangement (Fig. 4). These data were created by taking the ensemble-averaged measures of the best model from the previous example with the inclusion of a Gaussian noise factor, resulting in a target distance of 53.0 ± 1.6 Å and a target angle of 117.7 ± 8.3°. Whereas these data represent a simplified case of auxiliary measurements, the BEES code is capable of handling any auxiliary data that can be represented as a collection of points with associated Gaussian error estimates, as previously described (BEES algorithm). Inputs to the BEES routine are identical to the previous example, with the exception of the auxiliary data set.

With the addition of the distance and angle measurements, we find a shift in the best IC-identified model.



B

Sub-ensemble Model Details

#	Relative Model Performanc	BIC	Ensemble Size	Model Chi^2
0	1	5.54	2	0.79
1	0.91	5.72	2	0.83
2	0.74	6.14	2	0.91
3	0.73	6.17	2	0.92
4	0.65	6.39	2	0.96
5	0.51	6.88	3	0.75
6	0.51	6.9	3	0.75
7	0.5	6.94	3	0.79
8	0.49	6.97	3	0.76
9	0.49	6.98	3	0.79
10	0.49	6.98	3	0.78
11	0.48	7.01	3	0.81
12	0.48	7.02	3	0.77
13	0.47	7.04	2	1.09
14	0.47	7.07	3	0.78

FIGURE 3 Panels of the third tab of the BEES output HTML file ("Compare All Models"), which contains (*A*) the relative performances histogram as well as (*B*) a table of all the constructed ensemble models and their relative performance, ensemble size, selected IC metric, and goodness-of-fit values. (*C*) Selecting a particular model in the table will also visualize the constituent populations on the bar graph panel (best identified model selected here). The full interactive HTML file with high-resolution plots can be accessed by downloading the "K63_sas_only_plots.html" file from the example files contained within the Supporting Materials and Methods. A similar file for the inclusion of auxiliary data can be found in "K63_with_aux.html," also included in the Supporting Materials and Methods example files. To see this figure in color, go online.

Although still a two-state solution, the contributing members are now clusters 3 (43 ± 5%) and 4 (56 ± 5%). This model yields a χ^2_{total} of 0.80, with a χ^2_{SAS} of 0.96 and a χ^2_{aux} of 0.38. As was the case in the last example, there are a plethora of models containing three or more members in which better goodness of fits are observed, and the best goodness-of-fit model is a mixture of clusters 2, 4, and 11 and has a χ^2_{total} of 0.65. Although this model is arguably a better fit to the data than the two-state ensemble of clusters 3 and 4, the IC value of this model is larger because of the addition of a third population. As such, this model is only the eighth most probable model and possesses a relative performance of 0.63.

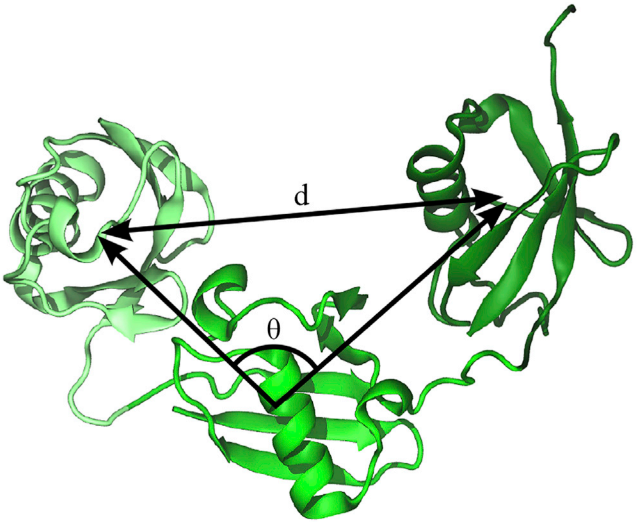


FIGURE 4 A visual representation of the two auxiliary measurements included in the second BEES routine. Both the distal monomer separation distance (d) and angle (θ) are measured in accordance to each monomer's center of mass. To see this figure in color, go online.

When we inspect the 10 best ensembles, we once again find the best model from the previous example, which possesses a χ^2_{total} of 0.81, a χ^2_{SAS} of 0.81, and a χ^2_{aux} of 0.83. Differences between the exact values of the χ^2_{SAS} metric in this example and the previous example are a result of the random-sampling nature of the (2) free metric, but these values are statistically indistinguishable. Similarly, the total goodness of fit in the clusters 3 and 4 ensemble is comparable to the ensemble containing clusters 2 and 9. As both models are two-state solutions, this results in very similar IC metrics and a relative performance value of 0.94, which suggests that neither model is significantly more accurate than the other. However, the 3 + 4 ensemble significantly outperforms the 2 + 9 ensemble in the context of the distance and angle measurements, whereas the 2 + 9 ensemble is a better fit to the scattering curve.

DISCUSSION

Here, we have presented the BEES program and highlighted its use with two example use cases. In the first example, we used the module to reweight the states of K63-linked triubiquitin that were obtained from accelerated molecular dynamics simulations. The BEES module identified a two-state solution as the model that best balanced the fit to experimental data with the fewest number of states. However, the analysis also found a plethora of models that had improved goodness of fits to the experimental scattering profile, but each of these models had more ensemble members than the two-state solution. The BEES module provides users with a convenient interface to both find and compare these other candidate ensembles with the IC-identified best state. This allows researchers the option to either rigorously trust the IC statistics to identify the most appropriate scattering model or to use the "ensemble of ensembles" constructed by the BEES module to guide their understanding of data sets separate from the fitting procedure.

The second use case discussed here demonstrated how BEES performs when simultaneously fitting populations to both SAXS and auxiliary data (here, simulated distance and angle measurements). In this example, the best identified model was still a two-state solution. However, a three-member ensemble was observed to have a better goodness of fit, but the improvement to $\chi^2{}_{\text{total}}$ was not sufficient to also improve the IC parameter, yielding a relative performance of 0.63. Because the two-state solution has strong agreement with both measurements ($\chi^2{}_{\text{free}}$, $\chi^2{}_{\text{aux}} < 1.0$), this relative performance value suggests that a conservative estimate for the solution ensemble would favor the two-state model over the χ^2 three-state case. However, the performance is of high enough quality that this ensemble could also be considered as a solution for future measurements. In this way, we emphasize that the relative performance metric should aid the intuition of researchers rather than completely replace it.

BEES seeks to identify the theoretical ensemble of states that uses the fewest number of populations to accurately describe the experimentally measured solution ensemble. In doing so, BEES is biased toward fitting the minimal amount of information contained within the experimental data so as to avoid potential overfitting. In contrast, other methods, such as genetic algorithms and maximal entropy approaches, will seek to use the full information of each scattering point (14,15,44). Whereas these methods may be susceptible to overfitting when ensemble size is not properly restricted (45,46), BEES is also susceptible to underfitting when utilizing SAXS data alone. As a result, the most accurate model of the underlying physical solution ensemble is likely one that is of a size between the smallest and largest ensembles identified by these different methods. This is especially true for the case of intrinsically disordered proteins, in which the high flexibility of the molecule would likely result in a large number of sampled conformations. In contrast, a protein with well-folded domains connected by flexible linkers may be better described with a minimal number of model states as the underlying nature of the molecule may be that of distinguishable "open" and "closed" states or differences between interacting domains, rather than the highly degenerate unfolded behavior of intrinsically disordered proteins. Furthermore, accurate use of any of these fitting methods is reliant on high-quality theoretical profiles; inaccurate theoretical states will likely lead to incorrect models. Therefore, users should be careful when selecting scattering calculator programs and parameters, and special attention should be paid to accurately accounting for hydration layer effects (47), especially when the maximal q-value being measured or modeled extends beyond 0.2 Å^{-1}, as previously described (33).

Several methodologies for modeling SAS data currently exist, and BEES can be considered a member of the "sample and select" family of methods (46), along with methods such as ensemble optimization method (14), SES (25), and many others (46). As has been previously noted (46), sample and select methods may yield models of high fit (low χ^2 values) but may not necessarily be accurate to the "true" physical ensemble being probed by the experimental data, especially when the experimental data is low dimensional or noisy (see Supporting Materials and Methods, Section S2). Therefore, users should be cautious to quickly interpret biochemical properties from a single ensemble model constructed from SAS data alone. To help users quickly analyze a wide collection of potential models, BEES provides users with the means to compare a wide family of models through the relative performance metric, which can be used to analyze alternative models to the "best fit" one according to their quality of fit and ensemble member differences. In this way, users can access these alternative models of high relative performance to infer alternative biochemical properties from the "best fit" ensemble and then further analyze their accuracy and discrepancies using orthogonal biochemical measurements. Alternatively, the bias of sample and select algorithms toward high fit but not highly accurate models may be ameliorated in BEES by including these high-resolution biochemical data as auxiliary fitting data. Measurements such as NMR chemical shifts or single-molecule fluorescence resonance energy transfer-derived distances could provide localized, high-resolution data that can more readily distinguish between individual states than the low-dimensional SAS curve data alone, whereas the SAS data can provide the information required to accurately model the overall size and shape of the molecule.

In summary, we have shown that BEES can be used to construct ensemble models of scattering data from a library of candidate states, and the iterative algorithm of BEES quantitatively resists overfitting of the data from the addition of unnecessary populations. The program is available as a module on SASSIE (https://sassie-web.chem.utk.edu/sassie2/) as well as in a stand-alone form (https://github.com/WereszczynskiGroup/BEES). BEES is designed for use by both new and expert users of computational ensemble modeling, and the graphical user interface (GUI)-based module for the SASSIE web platform provides structural and computational biophysicists with the resources necessary to construct molecular models in a Bayesian-based manner. Furthermore, BEES provides visual tools for quickly interpreting not only the quality of the best IC-identified model but also for the full ensemble of sub-basis models available from the candidate populations. This feature allows users to inspect many different potential solutions and to compare their ability to model both SAS and auxiliary data sets. In this way, BEES serves the intuition of structural researchers in building ensembles of states for their systems of interest.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2019.06.024.

AUTHOR CONTRIBUTIONS

S.B. and J.W. designed the BEES routine. S.B. and J.C. wrote the code of the BEES routine. S.B., J.E.C., and E.H.B. designed and wrote the SASSIE web implementation of BEES. S.B., J.C., and J.W. analyzed the data sets. S.B. and J.W. wrote the first manuscript draft, and all authors contributed to editing of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Susan Krueger for valuable discussions in designing the plotting interface.

E.H.B.'s work is supported by National Science Foundation grant number OAC-1740097 and NIH grant GM120600. S.B., J.C., and J.W. are supported by the National Institute of General Medical Sciences of the NIH under award number R35GM119647. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. This work benefited from CCP-SAS software developed through a joint Engineering and Physical Sciences Research Council (EP/K039121/1) and National Science Foundation (CHE-1265821) grant, as well as interactions and data collection at the Biophysics Collaborative Access Team, which is supported by National Institute of General Medical Sciences grant P41GM103622. This work used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant number ACI-1548562 (48). Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Q3

REFERENCES

1. Henzler-Wildman, K., and D. Kern. 2007. Dynamic personalities of proteins. *Nature.* 450:964–972.
2. Boldon, L., F. Laliberte, and L. Liu. 2015. Review of the fundamental theories behind small angle X-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev.* 6:25661.
3. Trewhella, J. 2016. Small-angle scattering and 3D structure interpretation. *Curr. Opin. Struct. Biol.* 40:1–7.
4. Graceffa, R., R. P. Nobrega, …, T. C. Irving. 2013. Sub-millisecond time-resolved SAXS using a continuous-flow mixer and X-ray microbeam. *J. Synchrotron Radiat.* 20:820–825.
5. Nasedkin, A., M. Marcellini, …, J. Davidsson. 2015. Deconvoluting protein (Un)folding structural ensembles using X-ray scattering, nuclear magnetic resonance spectroscopy and molecular dynamics simulation. *PLoS One.* 10:e0125662.
6. Plumridge, A., A. M. Katz, …, L. Pollack. 2018. Revealing the distinct folding phases of an RNA three-helix junction. *Nucleic Acids Res.* 46:7354–7365.
7. Cross, P. J., R. C. Dobson, …, E. J. Parker. 2011. Tyrosine latching of a regulatory gate affords allosteric control of aromatic amino acid biosynthesis. *J. Biol. Chem.* 286:10216–10224.
8. Fetler, L., E. R. Kantrowitz, and P. Vachette. 2007. Direct observation in solution of a preexisting structural equilibrium for a mutant of the allosteric aspartate transcarbamoylase. *Proc. Natl. Acad. Sci. USA.* 104:495–500.
9. Howell, S. C., X. Qiu, and J. E. Curtis. 2016. Monte Carlo simulation algorithm for B-DNA. *J. Comput. Chem.* 37:2553–2563.
10. Datta, S. A., J. E. Curtis, …, A. Rein. 2007. Conformation of the HIV-1 Gag protein in solution. *J. Mol. Biol.* 365:812–824.
11. Chen, P. C., and J. S. Hub. 2014. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* 107:435–447.
12. Hub, J. S. 2018. Interpreting solution X-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.* 49:18–26.
13. Yang, S., L. Blachowicz, …, B. Roux. 2010. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. USA.* 107:15757–15762.
14. Tria, G., H. D. Mertens, …, D. I. Svergun. 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ.* 2:207–217.
15. Pelikan, M., G. L. Hura, and M. Hammel. 2009. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* 28:174–189.
16. Schneidman-Duhovny, D., M. Hammel, …, A. Sali. 2016. FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* 44:W424–W429.
17. Hines, K. E. 2015. A primer on Bayesian inference for biophysical systems. *Biophys. J.* 108:2103–2113.
18. Fisher, C. K., A. Huang, and C. M. Stultz. 2010. Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.* 132:14919–14927.
19. Voelz, V. A., and G. Zhou. 2014. Bayesian inference of conformational state populations from computational models and sparse experimental observables. *J. Comput. Chem.* 35:2215–2224.
20. Ge, Y., and V. A. Voelz. 2018. Model selection using BICePs: a bayesian approach for force field validation and parameterization. *J. Phys. Chem. B.* 122:5610–5622.
21. Potrzebowski, W., J. Trewhella, and I. Andre. 2018. Bayesian inference of protein conformational ensembles from limited structural data. *PLoS Comput. Biol.* 14:e1006641.
22. Bowerman, S., A. S. J. B. Rana, …, J. Wereszczynski. 2017. Determining atomistic SAXS models of tri-ubiquitin chains from bayesian analysis of accelerated molecular dynamics simulations. *J. Chem. Theory Comput.* 13:2418–2429.
23. Winget, J. M., and T. Mayor. 2010. The diversity of ubiquitin recognition: hot spots and varied specificity. *Mol. Cell.* 38:627–635.
24. Reyes-Turcu, F. E., J. R. Horton, …, K. D. Wilkinson. 2006. The ubiquitin binding domain ZnF UBP recognizes the C-terminal diglycine motif of unanchored ubiquitin. *Cell.* 124:1197–1208.

Please cite this article in press as: Bowerman et al., BEES: Bayesian Ensemble Estimation from SAS, Biophysical Journal (2019), https://doi.org/10.1016/j.bpj.2019.06.024

25. Berlin, K., C. A. Castañeda, …, D. Fushman. 2013. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J. Am. Chem. Soc.* 135:16595–16609.

26. Brookes, E. H., N. Anjum, …, M. Pierce. 2015. The GenApp framework integrated with Airavata for managed compute resource submissions. *Concurr. Comput.* 27:4292–4303.

27. Perkins, S. J., D. W. Wright, …, J. E. Curtis. 2016. Atomistic modelling of scattering data in the collaborative computational project for small angle scattering (CCP-SAS). *J. Appl. Cryst.* 49:1861–1875.

28. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. Crysol– a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28:768–773.

29. Schneidman-Duhovny, D., M. Hammel, and A. Sali. 2010. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* 38:W540–W544.

30. Watson, M. C., and J. E. Curtis. 2013. Rapid and accurate calculation of small-angle scattering profiles using the golden ratio. *J. Appl. Cryst.* 46:1171–1177.

31. Stovgaard, K., C. Andreetta, …, T. Hamelryck. 2010. Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models. *BMC Bioinformatics.* 11:429.

32. Ravikumar, K. M., W. Huang, and S. Yang. 2013. Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J. Chem. Phys.* 138:024112.

33. Virtanen, J. J., L. Makowski, …, K. F. Freed. 2011. Modeling the hydration layer around proteins: applications to small- and wide-angle x-ray scattering. *Biophys. J.* 101:2061–2069.

34. Chen, P. C., and J. S. Hub. 2015. Interpretation of solution x-ray scattering by explicit-solvent molecular dynamics. *Biophys. J.* 108:2573–2584.

35. Rambo, R. P., and J. A. Tainer. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature.* 496:477–481.

36. Akaike, H. 2011. Akaike's information criterion. *In* International Encyclopedia of Statistical Science. M. Lovric, ed. Springer, p. 25.

37. Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Stat.* 6:461–464.

38. Spiegelhalter, D. J., N. G. Best, …, A. van der Linde. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B Stat. Methodol.* 64:583–639.

39. K. P. Burnham and D. R. Anderson, eds 2002. Information and Likelihood Theory: A Basis for Model Selection and Inference, Second Edition. Springer, pp. 76–123.

40. Kass, R. E., and A. E. Raftery. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–795.

41. Gelman, A., J. B. Carlin, …, D. B. Rubin. 2004. Bayesian Data Analysis, Second Edition. Chapman and Hall/CRC.

42. Franke, D., M. V. Petoukhov, …, D. I. Svergun. 2017. *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Cryst.* 50:1212–1225.

43. Konarev, P. V., and D. I. Svergun. 2015. A posteriori determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ.* 2:352–360.

44. Różycki, B., and E. Boura. 2014. Large, dynamic, multi-protein complexes: a challenge for structural biology. *J. Phys. Condens. Matter.* 26:463103.

45. Trewhella, J., A. P. Duff, …, A. E. Whitten. 2017. 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. *Acta Crystallogr. D Struct. Biol.* 73:710–728.

46. Ravera, E., L. Sgheri, …, C. Luchinat. 2016. A critical assessment of methods to recover information from averaged data. *Phys. Chem. Chem. Phys.* 18:5686–5701.

47. Henriques, J., L. Arleth, …, M. Skepö. 2018. On the calculation of saxs profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.* 430:2521–2539.

48. Towns, J., T. Cockerill, …, N. Wilkins-Diehr. 2014. XSEDE: accelerating scientific discovery. *Comput. Sci. Eng.* 16:62–74.