Biophysical Journal
Article

Biophysical Society

# Assessing Exhaustiveness of Stochastic Sampling for Integrative Modeling of Macromolecular Structures

Shruthi Viswanath,[1,*] Ilan E. Chemmama,[1,2] Peter Cimermancic,[1] and Andrej Sali[1,2,3,*]
[1]Department of Bioengineering and Therapeutic Sciences, [2]Department of Pharmaceutical Chemistry, and [3]Institute of Quantitative Biosciences, University of California San Francisco, San Francisco, California

ABSTRACT Modeling of macromolecular structures involves structural sampling guided by a scoring function, resulting in an ensemble of good-scoring models. By necessity, the sampling is often stochastic, and must be exhaustive at a precision sufficient for accurate modeling and assessment of model uncertainty. Therefore, the very first step in analyzing the ensemble is an estimation of the highest precision at which the sampling is exhaustive. Here, we present an objective and automated method for this task. As a proxy for sampling exhaustiveness, we evaluate whether two independently and stochastically generated sets of models are sufficiently similar. The protocol includes testing 1) convergence of the model score, 2) whether model scores for the two samples were drawn from the same parent distribution, 3) whether each structural cluster includes models from each sample proportionally to its size, and 4) whether there is sufficient structural similarity between the two model samples in each cluster. The evaluation also provides the sampling precision, defined as the smallest clustering threshold that satisfies the third, most stringent test. We validate the protocol with the aid of enumerated good-scoring models for five illustrative cases of binary protein complexes. Passing the proposed four tests is necessary, but not sufficient for thorough sampling. The protocol is general in nature and can be applied to the stochastic sampling of any set of models, not just structural models. In addition, the tests can be used to stop stochastic sampling as soon as exhaustiveness at desired precision is reached, thereby improving sampling efficiency; they may also help in selecting a model representation that is sufficiently detailed to be informative, yet also sufficiently coarse for sampling to be exhaustive.

## INTRODUCTION

Integrative structure determination is an approach for characterizing the structures of large macromolecular assemblies that relies on multiple types of input information, including from varied experiments, physical theories, and statistical analysis (1–4). Therefore, it maximizes the accuracy, precision, completeness, and efficiency of structure determination. Moreover, it can often produce a structure for systems that are refractive to traditional structure determination methods (5–11), such as x-ray crystallography, electron microscopy, and NMR spectroscopy. Integrative structure determination proceeds in four stages. First, all information that describes the system of interest, including data from wet lab experiments, statistical tendencies such as atomic statistical potentials (12–14), and physical laws such as molecular mechanics force fields (15,16), is collected. Second, a suitable representation for the system is chosen depending on the quantity and resolution of the available information. The available information is then translated into a set of spatial restraints on the components of the system. The spatial restraints are combined into a single scoring function that ranks alternative models based on their agreement with input information. Third, the alternative models are sampled using a variety of techniques, such as conjugate gradients, molecular dynamics, Monte Carlo (17), and divide-and-conquer message passing methods (18). The sampling generates an ensemble of models that are as consistent with the input information as possible. Finally, input information and output structures need to be analyzed to estimate structure precision and accuracy, detect inconsistent and missing information, and suggest more informative future experiments. Assessment begins with structural clustering of the modeled structures

---

*Submitted June 15, 2017, and accepted for publication October 2, 2017.*

*Correspondence: shruthi@salilab.org or sali@salilab.org

Shruthi Viswanath and Ilan E. Chemmama contributed equally to this work.

Peter Cimermancic's present address is Verily Inc., South San Francisco, California.

Editor: Amedeo Caflisch.

https://doi.org/10.1016/j.bpj.2017.10.005

© 2017 Biophysical Society.

produced by sampling, followed by assessment of the thoroughness of structural sampling, estimating structure precision based on variability in the ensemble of good-scoring structures, quantification of the structure fit to the input information, structure assessment by cross-validation, and structure assessment by data not used to compute it. Integrative modeling can iterate through these four stages until a satisfactory model is built.

A key challenge in integrative modeling of biomolecular structures is to map the complete ensemble of models consistent with the input information (good-scoring models) (1,2,19,20). The variation among the models in this ensemble quantifies the uncertainty of modeling (model precision). Because sampling large macromolecular systems is often necessarily stochastic, we can only aim to find representative good-scoring models. These representative models sample all good-scoring models at some precision, which we define as the sampling precision. Clearly, the sampling precision imposes a lower limit on the model precision. Therefore, exhaustive sampling of good-scoring models is a prerequisite for accurate modeling and assessment of model precision. Sampling is exhaustive at a certain precision when it generates all sufficiently good-scoring models at this precision. Importantly, sampling exhaustiveness and sampling precision are invariably intertwined. There is always a precision at which any sampling is exhaustive; for example, even a single structure provides an exhaustive sample at a precision worse than the scale of the system.

Accurate estimation of model precision is key in assessing an integrative structure. It is perhaps more important to assess the precision of a model than to compute a model in the first place. The reason is that the utility of a model is determined significantly by its precision. First, model precision provides an estimate of the aggregate uncertainty in the input information; second, it likely provides the lower bound on model accuracy; finally, applications of models strongly depend on their accuracy, with different applications having varied requirements for accuracy and precision (20–22). Further, only when model precision is estimated accurately, can the model be used to inform future choices, such as whether to gather more data, change the system representation, scoring functions, or sampling algorithms. Commonly used structural features for estimating model precision include the particle positions, distances, and contacts (5,6,23,24), although specific systems may benefit from the use of derived features, such as the distance to a membrane in a transmembrane assembly. Of particular interest are the features that have a single maximum in their probability distribution. The spread around the maximum describes how precisely the feature was determined by the input information.

Sampling convergence in Monte Carlo simulations for protein and RNA structure prediction has been assessed by checking for abundance of structures close to the lowest energy structure(s) (25–32). Convergence in molecular dynamics (MD) simulations has been measured by counting the number of structural clusters (33–35) and their relative populations (36–40), cosine of the principal components (41), distance between the free energy surfaces of different parts of the simulation (42), and drift in the free energies (43). Some methods assess convergence in MD simulations by comparing different trajectories via a difference in populations for each cluster (36–40). For example, models from a “reference” simulation are first clustered based on a predetermined cutoff (38), followed by assigning models from additional simulation to the nearest cluster in the reference simulation; thus, each simulation produces a histogram of populations of clusters that enables comparison of any two simulations.

As mentioned above, testing for sampling exhaustiveness is the first step of the analysis and validation stage of our four-stage integrative modeling process, immediately following the sampling stage (2,4,7–9,44). Here, we present an objective and automated protocol that aims to estimate the precision at which sampling is exhaustive, thereby assessing sampling exhaustiveness for integrative structural modeling. As a proxy for assessing sampling exhaustiveness, we evaluate whether or not two independently and stochastically generated sets of models (model samples) are sufficiently similar. Model samples for assessment can be obtained, for example, from two independent simulations using random starting models or different random number generator seeds. The protocol for evaluating exhaustiveness includes two tests that consider the model scores, followed by two tests that consider the model structures.

There are at least two major limitations of our approach. First, sampling convergence is at best an approximation of sampling exhaustiveness. Although similarity between independent model samples does indicate sampling convergence, we can only hypothesize that the convergence of stochastic sampling at some precision also indicates sampling exhaustiveness at that precision, for scoring function landscapes like those used in integrative structure modeling (many dimensions, rugged, few major minima). This hypothesis is supported by all five examined cases of binary docking solutions enumerated at a specified precision. Accordingly, passing the proposed tests is a necessary, but not sufficient condition for exhaustive sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the native state. Second, our tests are also not applicable to methods whose sampling is not stochastic (e.g., a conjugate gradients minimization from a fixed unique starting point) or so expensive that they cannot generate a large enough sample of independent models.

The rest of the article is organized as follows. In Methods, we describe the four-part protocol for estimating sampling precision and assessing sampling exhaustiveness, including

its application to five illustrative cases of binary protein complexes. In Results, we demonstrate the protocol on the illustrative cases and validate it by comparing stochastic model samples with models from exhaustive enumeration using rigid docking (45,46). Parameters of the protocol, its applicability and uses, its shortcomings, the relationship between various kinds of precision in integrative modeling, relation to prior work, and future work are addressed in Discussion.

## METHODS

The protocol for estimating sampling precision and assessing sampling exhaustiveness (Fig. 1) consists of four tests that are increasingly stringent; each test needs to be passed before it makes sense to proceed to the next test. Given two model samples and their scores as input, the tests check 1) convergence of the model score, 2) whether model scores for the two model samples were drawn from the same parent distribution, 3) whether each structural cluster includes models from each sample proportionally to its size, and 4) whether there is sufficient structural similarity between the two model samples in each cluster. Next, each step in the flowchart (Fig. 1) is described in turn.

### Generating inputs for the protocol

The necessary input for the protocol is two model samples of approximately equal size and their scores. Each model sample consists of random, independently generated models. Both model samples must be generated using the same sampling method. In integrative structure modeling, we are generally not interested in all sampled models, but only in models that are good-scoring (i.e., those that are sufficiently consistent with input information) (2,4,7–9,44).

The precise definition of good-scoring models is left to the user and can be application-dependent. Example choices include all models scoring better than a threshold on the total score (2,4,7–9,44), or all models satisfying all types of input information within acceptable thresholds. For example, if a protein complex is modeled by fitting its components into an electron microscopy density map subject to cross-linking, excluded volume, and sequence connectivity, the corresponding scoring function can be a sum of the correlation coefficient between the EM map and a model as well as harmonic (Gaussian) restraints for chemical cross-links, pairs of overlapping atoms, and sequence connectivity; a good-scoring model can then be defined as a model that fits the EM density with a cross-correlation >0.80 and violates (e.g., a restraint value > 2 SD from the mean) a smaller number of harmonic restraints than expected for the corresponding Gaussian distributions (e.g., 5%).

The samples are usually generated by a stochastic sampling algorithm. One such algorithm is the Metropolis Monte Carlo scheme (47) that starts from a different configuration and/or random number seed for each run. For our purposes, a larger number of shorter runs is preferable over a smaller number of longer runs for two reasons. First, a larger number of runs benefits more easily from parallel execution than a smaller number of runs. Second, independent runs are guaranteed to result in uncorrelated models, whereas, additional care is needed to ensure the lack of correlation for models from a single run.

### Convergence of the best score

The first test assesses whether the best model score continues to improve as more models are sampled. This test operates on random subsets of the model scores of the two samples combined. Model score subsets of several sizes (e.g., 20, 40, 60, 80%, and the complete set) are each created several times (replicates). The best score in each subset is averaged across the replicates. Plotting the average best score for each model subset size shows whether the best score converges as the number of models is increased.

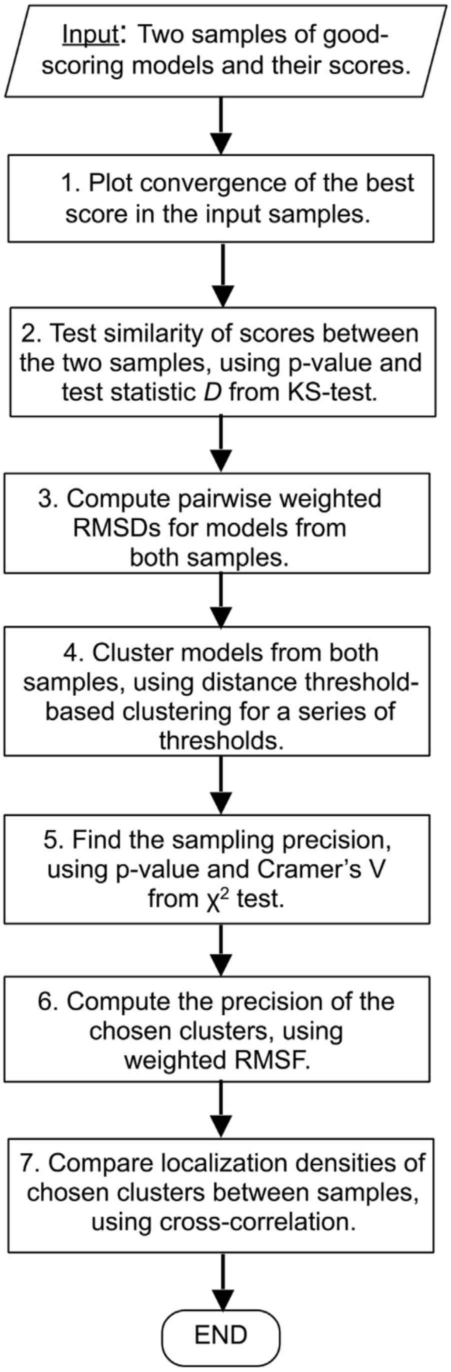
Input: Two samples of good-scoring models and their scores.
1. Plot convergence of the best score in the input samples.
2. Test similarity of scores between the two samples, using p-value and test statistic D from KS-test.
3. Compute pairwise weighted RMSDs for models from both samples.
4. Cluster models from both samples, using distance threshold-based clustering for a series of thresholds.
5. Find the sampling precision, using p-value and Cramer's V from χ2 test.
6. Compute the precision of the chosen clusters, using weighted RMSF.
7. Compare localization densities of chosen clusters between samples, using cross-correlation.
END

FIGURE 1 Flowchart of the protocol for estimating sampling precision and assessing sampling exhaustiveness.

### Similarity of scores

The second test confirms that good-scoring models in the two model samples have similar score distributions (i.e., satisfy the data equally well). Specifically, the nonparametric Kolmogorov-Smirnov two-sample

test (48,49) tests the null hypothesis that the distributions of model scores in the two model samples were drawn from the same parent distribution. The $p$ value from the Kolmogorov-Smirnov two-sample test is a measure of the statistical significance of the difference between the two distributions. A $p$ value lower than the cutoff of significance (usually 0.05) indicates that the difference in the two score distributions is statistically significant.

Even a tiny difference between two distributions can be significant if the samples are large (50,51). Therefore, we additionally use an effect size measure for the Kolmogorov-Smirnov two-sample test. Conveniently, the Kolmogorov-Smirnov two-sample test statistic, $D$, is itself a proportion (48,49). The proportion ranges from 0 to 1, where 0 represents no difference between the two samples and 1 no overlap between the two samples. A value of 0.30 (medium effect size) or higher suggests that the two score distributions are different (48,49).

Finally, we conclude that the score distributions are similar if the difference is not statistically significant ($p$ value $> 0.05$) or if the difference is significant ($p$ value $< 0.05$) but its magnitude is small ($D < 0.30$).

## Computing pairwise root-mean-square deviations

The third test assesses whether models from each sample are present in each structural cluster proportionally to the sample size; when the sample sizes are equal, each cluster should contain approximately the same number of models from each sample. The test requires clustering models from both samples combined. It may be necessary to select sufficiently small random subsets of the two model samples, to make clustering computationally feasible.

The first step of clustering is to compute root-mean-square deviation (RMSD) values between all pairs of models from both samples combined (8):

$$RMSD_{i,j} = \left(\sum\nolimits_{k=1}^{b} n_k \left(\vec{x}_{i,k} - \vec{x}_{j,k}\right)^2 \Big/ \sum\nolimits_{k=1}^{b} n_k\right)^{1/2}$$

where $\vec{x}_{i,k}$ is the Cartesian coordinate of the $k$th of $b$ beads in model $i$, $n_k$ is the number of residues in bead $k$, and $n$ is the total number of models; other structural dissimilarity or similarity measures may be used.

## Finding the sampling precision

A stochastic sampling method does not enumerate all good-scoring models, but generates only a sample of them. Here, the sampling precision is defined as the radius of the clusters in the finest clustering for which each sample contributes models proportionally to its size (considering both significance and magnitude of the difference) and for which a sufficient proportion of all models occur in sufficiently large clusters (Fig. 2).

## Clustering models using several thresholds

To find the sampling precision, we evaluate increasingly coarser clusterings, obtained using the following threshold-based clustering method (33). For each model, we first find all neighboring models, defined as models whose RMSD distance (above) from the model is less than the input threshold. Initially all models are unclustered. The unclustered model with the maximum number of neighbors and its neighbors are added to form a new cluster, and the list of unclustered models is updated. The last step is repeated until no unclustered models remain. This clustering is performed for all thresholds sampling the interval between the minimum and maximum RMSDs in steps of 2.5 Å. The next three paragraphs describe the three criteria evaluated for each clustering.

*Significance*. To assess the significance of the difference between the proportions of each sample in the clusters, we use the $\chi^2$ test for homogeneity of proportions (52). This test evaluates the null hypothesis that the two model samples are distributed nearly equally (for equal-sized samples) or approximately in proportion to their sizes (for unequal sized samples) in all major clusters. The $p$ value from the test is a measure of the statistical significance of proportionate contributions to clusters from both samples. A $p$ value lower than the cutoff of significance (usually 0.05) indicates that the difference in the two distributions is statistically significant.

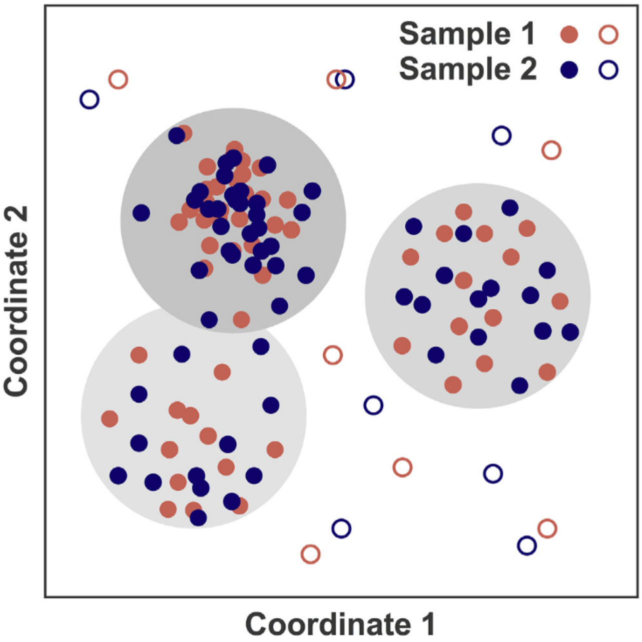

FIGURE 2 Conceptual representation of the $\chi^2$ test for sampling exhaustiveness, showing models in a 2D coordinate space. Two independent equal-sized random samples of good-scoring models are shown in red and blue. Models in the two samples are clustered together. The gray circles indicate cluster boundaries and the gray-scale indicates the density of models in the cluster. The size of the circles indicates the clustering threshold. The test assesses whether the proportion of models from the two samples (*red* and *blue*) is similar in each significant cluster. Note that some models are shown as open circles, indicating that these models belong to insignificant clusters (i.e., small clusters containing few models).

*Magnitude*. To assess the magnitude of the difference between the proportions of each sample in the clusters, we use an effect size measure for the $\chi^2$ test, Cramer's $V$ (53). This test measures the magnitude of the difference between the distributions of the two samples across clusters. Cramer's $V$ is defined as $\sqrt{\chi^2/n}$, where $\chi^2$ is the $\chi^2$ test metric and $n$ is the total number of models in both samples. A value of $V$ of at least 0.1 suggests that the difference between the two distributions is large; it corresponds approximately to a $p$ value of 0.05 for the case of two clusters and 500 models per sample.

*Population*. The calculation of the $p$ value and Cramer's $V$ requires that each sample has at least 10 expected models per cluster (54). Therefore, we remove all clusters containing $<10$ models from either sample. To allow us to proceed with the assessment, we also require that at least 80% of the models remain after this removal.

## Computing precision of clusters

For defining clusters and visualization, any threshold equal to or worse than the sampling precision can be chosen. The sampling precision is the smallest clustering threshold at which sampling is exhaustive; choosing a larger threshold will result in fewer, larger clusters, and may be preferable for analysis and/or visualization.

Although the sampling precision limits the maximum radius of a cluster (Fig. 2), models could be more tightly distributed inside a cluster. To

quantify the actual spread of models in clusters, we define the cluster precision as the weighted root-mean-square fluctuation (RMSF) of all models in the cluster. Weighted RMSF accounts for differing sizes of beads often used to represent integrative structures (8,55). It is computed using

$$\left\langle RMSF^2 \right\rangle^{1/2} = \left[ \left( \sum_{n=1}^{b} n_k \sum_{i=1}^{n} \left( \vec{x}_{i,k} - \langle x_k \rangle \right)^2 \right) \Big/ n \left( \sum_{k=1}^{b} n_k \right) \right]^{1/2}.$$

The cluster precision is ~1.4 times the sampling precision, reflecting the general relationship between RMSD and RMSF (8,55).

## Computing localization densities and their cross correlation

The final test involves computing the cross correlation between the model densities from each sample, for each cluster. The density maps are created at a resolution equal to the threshold used for defining clusters (above). The cross-correlation coefficients between the maps are calculated using the software UCSF Chimera (56).

## Validation of the protocol

We illustrate our protocol by relying on five binary protein complexes of known structure from the ZDOCK Benchmark 4.0 (57), spanning a range of docking difficulty, and 5–7 simulated distance restraints per complex. We modeled the structure of each complex by stochastic sampling as implemented in an integrative modeling platform (IMP; Supporting Material). We assessed the sampling exhaustiveness protocol based on a comparison of stochastic sampling with exhaustive enumeration, as follows.

The quality of the sampling exhaustiveness protocol is quantified by the fraction of good-scoring models from exhaustive enumeration (below) that are located within any sufficiently large cluster of the good-scoring models from the tested sampling, for the clustering threshold equal to the tested sampling precision; an enumerated model is located in a cluster, if its distance to the cluster center is within the tested sampling precision.

Fast-Fourier transform-based protein docking algorithms (45,58–60) efficiently construct models of binary protein complexes by enumerating all possible rigid rotations and translations on a uniform 3D grid. The set of all models (57) produced by ~1.2 Å and 6° uniform sampling on an FFT grid was used. Good-scoring models from enumeration were identified as in stochastic sampling (models for which at least 90% of cross-links span a C$\alpha$-C$\alpha$ distance of <12 Å). For each good-scoring ZDOCK model, its distance to the nearest major cluster center from IMP was calculated.

The distribution of models from stochastic sampling in IMP cannot be compared directly to enumerated models computed by ZDOCK. The ZDOCK models are enumerated on a uniform grid, whereas IMP samples the posterior probability of models and therefore produces a nonuniform model distribution. In addition, ZDOCK and IMP use different representations (atomic and coarse-grained, respectively).

## RESULTS

We demonstrate the sampling exhaustiveness protocol on an example from the Protein Data Bank (PDB), 1AVX. The remaining four examples are described in Figs. S1–S5.

There are 3369 good-scoring models for PDB: 1AVX (1896 in sample 1 and 1473 in sample 2). The score convergence test shows that the best score does not continue to improve significantly with an increase in the number of models sampled (Fig. 3 *A*; to visualize the relatively rapid convergence in model scores, see Fig. S6). The two score distributions are similar to each other, as shown by the overlap in the score histograms and the insignificant *p* value and small *D* value from the Kolmogorov-Smirnov two-sample test (Fig. 3 *B*).

Next, exhaustiveness is examined at varying thresholds between the minimum and maximum RMSDs of 0.43 and 42.93 Å (Fig. 3 *C*; Table 1). Based on the *p* value, Cramer's *V*, and the population of models in the contingency table, the $\chi^2$ test is satisfied from the threshold of 12.93 Å onwards (Table 1). Hence, the sampling precision is 12.93 Å. In general, stricter (smaller) clustering thresholds result in many small clusters, which are ignored (Table 1, *last column*; Fig. 3 *C*). In contrast, more lenient (larger) clustering thresholds result in fewer, larger clusters that are more likely to be retained in the analysis. For example, for the lowest clustering threshold of 0.43 Å, each model is in its own-cluster and hence all clusters are small and eliminated from the contingency table. In contrast, for thresholds >25.43 Å (Fig. 3 *C*), only one cluster containing all models remains.

Finally, we chose the sampling precision as the clustering threshold for visualizing clusters. Inspection of the cluster populations (Fig. 3 *D*) shows that they are similar for the two samples. The sampling precision is ~1.4 times the cluster precision, as expected from the general relationship between RMSD and RMSF (Methods; (55)). The agreement between the localization densities for samples 1 and 2 (Fig. 3, *E* and *F*) is demonstrated by the high cross-correlation coefficient of 0.99 for each cluster.

## Validation by comparison to exhaustive enumeration

The sampling exhaustiveness protocol was validated by showing that 99.2% of the good-scoring ZDOCK models were within an IMP cluster for PDB: 1AVX (Fig. 4); the corresponding fraction was 100% for the other four examples (Fig. S5). For PDB: 1AVX, out of 510 good-scoring ZDOCK models, 506 were within the sampling precision of the center of a significant cluster and the distances for the other four models were less than one grid spacing further away (Fig. 4). Similarly, the largest distances between good-scoring ZDOCK and IMP models were 1.52, 3.96, 1.56, and 0.96 Å short of their sampling precisions for PDB: 1I2M, 1SYX, 2IDO, and 7CEI, respectively (Fig. S5). In conclusion, the sampling exhaustiveness protocol neither overestimates nor underestimates the sampling precision, for the five examined cases.

## DISCUSSION

### Summary of the protocol

Accurate assessment of model uncertainty in integrative modeling necessitates that sampling is exhaustive at a

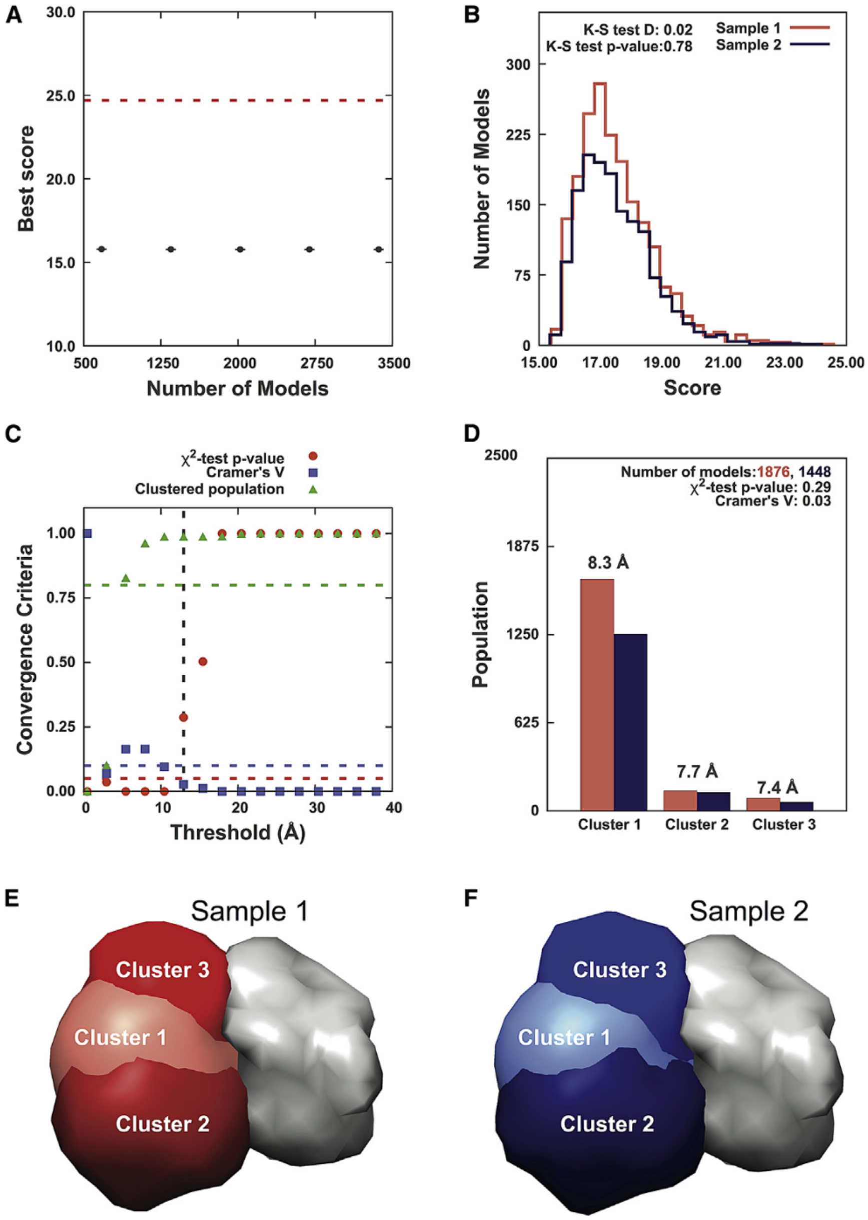

FIGURE 3 Results for sampling exhaustiveness protocol for PDB: 1AVX. (*A*) Shown here are results of test 1, convergence of the model score, for the 3369 good-scoring models; the scores do not continue to improve as more models are computed essentially independently. The error bar represents the SD of the best scores, estimated by repeating sampling of models 10 times. The red dotted line indicates a lower bound on the total score. (*B*) Shown here are results of test 2, testing similarity of model score distributions between samples 1 (*red*) and 2 (*blue*); the difference in distribution of scores is not significant (Kolmogorov-Smirnov two-sample test $p$ value $> 0.05$) and the magnitude of the difference is small (the Kolmogorov-Smirnov two-sample test statistic $D$ is 0.02); thus, the two score distributions are effectively equal. (*C*) Shown here are results of test 3, containing three criteria for determining the sampling precision ($y$ axis), evaluated as a function of the RMSD clustering threshold (33) ($x$ axis). First, the $p$ value is computed using the $\chi^2$ test for homogeneity of proportions (52) (*red dots*). Second, an effect size for the $\chi^2$ test is quantified by the Cramer's $V$ value (*blue squares*). Third, the population of models in sufficiently large clusters (containing at least 10 models from each sample) is shown as green triangles. The vertical dotted gray line indicates the RMSD clustering threshold at which three conditions are satisfied ($p$ value $> 0.05$ (*dotted red line*), Cramer's $V < 0.10$ (*dotted blue line*), and the population of clustered models $> 0.80$ (*dotted green line*)), thus defining the sampling precision of 12.93 Å. (*D*) Populations of sample 1 and 2 models in the clusters are obtained by threshold-based clustering using the RMSD threshold of 12.93 Å. Cluster precision is shown for each cluster. (*E* and *F*) Shown here are results of test 4: comparison of localization probability densities of models from sample 1 (*red*) and sample 2 (*blue*) in each cluster. The density map of the receptor, which is kept fixed through the simulation, is shown in gray. All densities were visualized at a threshold equal to one-third the maximum. The cross-correlation of the density maps of the two samples is 0.99 for each of the three clusters.

precision sufficient for assessing model uncertainty. In this article, we introduce a protocol for determining the sampling precision of integrative structural models computed by a stochastic sampling algorithm. The protocol requires two samples of independently and stochastically generated sets of models and their scores. It includes two tests for convergence of the score and two tests for convergence of the structures. The tests for score convergence assess whether the scores in the two samples are from similar distributions. The tests for structural convergence rely on structural clustering of the models, followed by assessing whether the models in the two samples are distributed similarly across the clusters. The five illustrative cases demonstrate the relative accuracy of the sampling exhaustiveness protocol (Figs. 3 and 4; Figs. S1–S5). Below, we discuss the parameters used in the protocol, and its applicability, shortcomings, and relationship among various kinds of precision in integrative modeling; we then address overfitting in integrative modeling, relation to prior work, and future work.

## Parameters

All parameters used in the protocol are listed next; their values are chosen based on rules-of-thumb in statistics literature. First, the significance cutoff for the *KS* test is 0.05 and the magnitude cutoff for the KS statistic, $D$, is 0.3, the latter corresponding to medium effect size (49).

**TABLE 1 Three Criteria for Determining the Sampling Precision for PDB: 1AVX, Evaluated as a Function of the Clustering Threshold**

| Threshold in Ångstroms | $p$ Value | Cramer's $V$ | Population of Models in Contingency Table [%] |
|---|---|---|---|
| 0.4 | 0.0 | 1.0 | 0.0 |
| 2.9 | 0.0 | 0.1 | 10.1 |
| 5.4 | 0.0 | 0.2 | 82.7 |
| 7.9 | 0.0 | 0.2 | 96.2 |
| 10.4 | 0.0 | 0.1 | 98.7 |
| 12.9 | 0.3 | 0.0 | 98.7 |
| 15.4 | 0.5 | 0.0 | 98.7 |
| 17.9 | 1.0 | 0.0 | 98.8 |
| 20.4 | 1.0 | 0.0 | 99.8 |
| 22.9 | 1.0 | 0.0 | 100.0 |
| 25.4 | 1.0 | 0.0 | 100.0 |
| 27.9 | 1.0 | 0.0 | 100.0 |
| 30.4 | 1.0 | 0.0 | 100.0 |
| 32.9 | 1.0 | 0.0 | 100.0 |
| 35.4 | 1.0 | 0.0 | 100.0 |
| 37.9 | 1.0 | 0.0 | 100.0 |
| 40.4 | 1.0 | 0.0 | 100.0 |
| 42.9 | 1.0 | 0.0 | 100.0 |

The three criteria are 1) $p$ values and 2) Cramer's $V$, both from the $\chi^2$ test; and 3) the population of models in the contingency table after eliminating small clusters.

Second, due to the inability of the $\chi^2$ test to handle small expected cell counts, we eliminate clusters with $<$10 models for either sample from the contingency table for the test, as recommended (54). Third, because we are eliminating these small clusters, we additionally check if the population of models remaining in the contingency table from both samples is $>$80%. Finally, the significance cutoff for the $\chi^2$ test is 0.05 and the magnitude cutoff on Cramer's $V$ is 0.1, the latter corresponding approximately to a $p$ value of 0.05 for the case of two clusters and 500 models per sample.

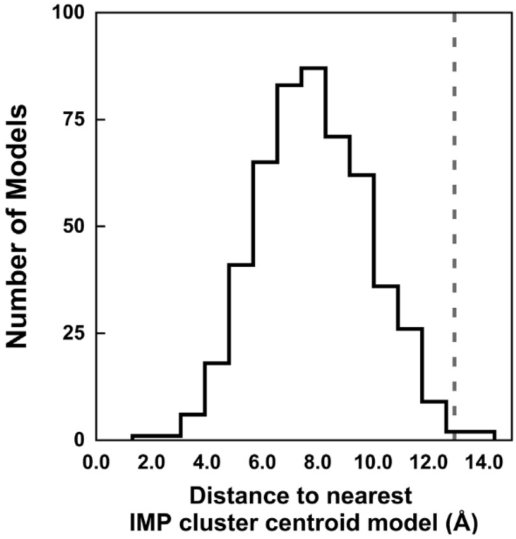

FIGURE 4 Histogram showing the distribution of distance (measured by weighted ligand RMSD) of a good-scoring PDB: 1AVX model from enumeration (ZDOCK) to the nearest cluster centroid model from stochastic sampling (IMP). The dotted line indicates the sampling precision for the IMP model sample determined by the sampling exhaustiveness protocol.

## Applicability and uses

The sampling exhaustiveness protocol is broadly applicable to a range of sampling methods, a range of clustering or binning methods, features of models other than model scores, and models other than macromolecular structures, and it can be used dynamically during sampling to stop as soon as desired sampling precision is reached, as follows.

First, any stochastic sampling method that generates a large number of independent model samples is appropriate. Metropolis Monte Carlo sampling can satisfy this requirement of independence by 1) sampling models from multiple independent trajectories (e.g., starting from different random initial configurations) and 2) sampling models at sufficiently distant intervals on a single trajectory, such that samples are effectively uncorrelated with each other. The sampling exhaustiveness protocol can only compare model samples produced by the same sampling algorithm (e.g., samples from uniform sampling and importance sampling are clearly not directly comparable).

Second, any one of the variety of clustering or binning methods for grouping models based on their similarity could be used instead of the distance threshold-based clustering. In principle, even a uniform grid could be applied to bin the models. This clustering is used as a relatively rapid method to assign most models to a relatively small number of groups of similar precision. As a result, we can easily quantify the sampling precision across the entire space of models and convey the results in terms of a small number of model clusters. In contrast, for example, k-means clustering generally results in clusters of varying precision, thus obfuscating the relationship between the cluster precision and sampling precision.

Third, any quantity of interest, such as radius of gyration and distance to a membrane, can be tested in the same manner as the model scores here.

Fourth, the protocol is applicable to stochastic sampling of any kind of a model, not just a structural model.

Fifth, and finally, the protocol can be applied to estimate sampling precision dynamically during a simulation, so that sampling is stopped as soon as desired sampling precision is reached, maximizing sampling efficiency. Assessment of exhaustiveness is particularly important for modeling large systems with many degrees of freedom, where exhaustive sampling of representative good-scoring solutions is particularly difficult.

## Critique

In the absence of enumeration, exhaustiveness of stochastic sampling cannot be proved with complete certainty. Therefore, we suggest that even a statistical test such as the one

proposed here is better than no test. As a proxy for assessing exhaustiveness, our protocol evaluates whether two independent random model samples are similar to each other (Introduction). Our tests are not applicable to methods that do not generate independent random samples (e.g., a conjugate gradients minimization from a fixed unique starting point), or are so expensive that they cannot generate a large enough sample of independent models. Further, passing the proposed tests is a necessary, but not sufficient, condition for exhaustive sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the native state. Nevertheless, based on the five examples, we argue that convergence of stochastic sampling at some precision often also indicates sampling exhaustiveness at that precision.

## Precision in integrative modeling

In this article, we used the model (ensemble) precision, sampling precision, and cluster precision. In addition, the data precision (uncertainty) reflects the experimental noise (systematic and random error) (4); and the representation precision can be defined, for example, by the diameter of the largest primitive (Gaussian, bead) used to represent the system. We now discuss these five precisions in the context of each other.

First, the sampling precision imposes a lower limit on the model precision. The shape of the scoring function landscape at precisions better than the sampling precision is not sampled accurately by definition; thus, any features of the model landscape more precise than the sampling precision are unlikely to be estimated accurately.

Second, because the model ensemble is divided into one or more clusters, the model precision is always equal to or worse than any cluster precision.

Third, for the final description of the model ensemble, it only makes sense to cluster the models using a clustering threshold that is equal to or larger than the sampling precision (due to the first point above; see Fig. 4).

Fourth, and lastly, the sampling precision is in turn limited by the representation precision and data precisions. Although the model, sampling, and cluster precisions, as defined here, are directly comparable to each other, the representation and data precisions are defined on different scales. Nevertheless, qualitatively speaking, the sampling precision cannot be significantly higher than the representation and data precisions; moreover, it is likely not beneficial to use a representation with a precision that is significantly higher than the data precision.

## Addressing overfitting in integrative structure modeling

Overinterpretation of the data (overfitting) is a frequent concern in any modeling. For example, a single high-resolution atomic model may fit an EM density map at intermediate resolution well; proposing such a model as the solution is often a case of overfitting because there are likely many other atomic models that also fit the data equally well. Our sampling exhaustiveness test provides a potential insurance against overfitting. When a test is passed, overfitting is not a problem (at the sampling precision) because all models (at this precision) that are consistent with the data are provided in the output model ensemble.

## Relation to prior work

The methods most related to that in this article, applied in the context of MD simulations, are those in (36–39) (also used in (40)). In (36,37), models from multiple MD simulations are combined and compared in terms of their relative populations. In (38), a new simulation is compared against a reference simulation, by clustering models from the reference simulation based on a predetermined cutoff. The models of the new simulation are then assigned to the nearest cluster from the reference simulation. Thus, each simulation produces a histogram of populations across clusters and any two simulations can be compared by the difference in their populations for each cluster. In (39), this method is expanded by computing the number of independent samples in an MD trajectory as a way of assessing the sampling quality. The number of independent samples in an MD simulation is determined by comparing the observed variance in the population of a cluster to the expected analytical variance from an independent and identically distributed sample, for various subsample sizes.

Our protocol additionally determines the significance and magnitude of the difference in population distributions across clusters, using the $\chi^2$ test. More importantly, our protocol also determines the sampling precision objectively, by applying the $\chi^2$ test for a range of clustering thresholds (Figs. 3 and 4). Moreover, we test both score convergence and convergence of structural coordinates (Fig. 1). A few minor differences exist in our respective clustering methods as well: 1) similarly to (36,37), we cluster models from all simulations, potentially producing a more comprehensive set of clusters, in contrast to clustering only models from the reference simulation (38,39); and 2) our cluster centers are chosen based on the density of models close to the cluster center, in contrast to choosing cluster centers randomly (38), choosing clusters of uniform probability (39), or choosing cluster centers based on average linkage with a similarity cutoff (36,37). Finally, our statistical test applies to independent samples from a stochastic algorithm such as Monte Carlo sampling, whereas some other methods do not require the samples to be independent (36–40). Preliminary versions of our sampling exhaustiveness protocol have been already used in several integrative modeling applications (9–11,61–63). Earlier, sampling

exhaustiveness for integrative modeling was estimated, at best, by manual visual inspection of localization densities of clusters (7,8,44).

## Future directions

Future directions include expanding this protocol to establish more detailed tests for exhaustiveness. For instance, it will be useful to determine not just the sampling precision for the entire macromolecular system, but also the sampling precision for different components of the system (e.g., proteins, domains) separately. Such more detailed information would be useful in the analysis stage of the iterative four-stage integrative modeling process (2,4) to determine, for instance, what representations to change and what input data to reexamine to improve the sampling precision for the entire system.

Structures of macromolecular systems are increasingly computed by integrative modeling that relies on various types of experimental data and theoretical information (20). However, validation of integrative models and data is a major open research challenge. It is particularly timely because of the Worldwide Protein Data Bank effort to expand the scope of its archive to integrative structures (20). We suggest that a sampling exhaustiveness protocol, such as the one described here, is the first assessment applied to all integrative models.

## Availability

Benchmark data and code used in this article are freely available at http://salilab.org/sampcon. The code relies on our open source IMP package (http://integrativemodeling.org).

## SUPPORTING MATERIAL

Supporting Materials and Methods, six figures, and one table are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)31090-1.

## AUTHOR CONTRIBUTIONS

S.V., I.E.C., and A.S. designed research. S.V. and I.E.C. performed research. P.C. contributed computational tools. S.V., I.E.C., and A.S. analyzed data. S.V., I.E.C., and A.S. wrote the manuscript.

## ACKNOWLEDGMENTS

The authors thank the members of their research group for useful suggestions and Dr. Benjamin Webb for helping to implement the method in IMP.

This work was supported by the National Institutes of Health (NIH) (P01 GM105537, R01 GM083960, and P41 GM109824 to A.S.) and the National Science Foundation (NSF) (graduate research fellowship 1650113 to I.E.C.). Molecular graphics images were produced using the UCSF Chimera package from the Computer Graphics Laboratory, University of California, San Francisco, California (supported by NIH P41 RR-01081).

## REFERENCES

1. Ward, A. B., A. Sali, and I. A. Wilson. 2013. Biochemistry. Integrative structural biology. *Science.* 339:913–915.
2. Russel, D., K. Lasker, …, A. Sali. 2012. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10:e1001244.
3. Webb, B., K. Lasker, …, A. Sali. 2014. Modeling of proteins and their assemblies with the integrative modeling platform. *In* Methods in Molecular Biology. Y. Chen, ed. Humana Press, London, UK, pp. 277–295.
4. Schneidman-Duhovny, D., R. Pellarin, and A. Sali. 2014. Uncertainty in integrative structural modeling. *Curr. Opin. Struct. Biol.* 28:96–104.
5. Alber, F., S. g, …, M. P. Rout. 2007. The molecular architecture of the nuclear pore complex. *Nature.* 450:695–701.
6. Lasker, K., F. Förster, …, W. Baumeister. 2012. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl. Acad. Sci. USA.* 109:1380–1387.
7. Shi, Y., J. Fernandez-Martinez, …, B. T. Chait. 2014. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteomics.* 13:2927–2943.
8. Robinson, P. J., M. J. Trnka, …, R. D. Kornberg. 2015. Molecular architecture of the yeast mediator complex. *eLife.* 4:e08719.
9. Fernandez-Martinez, J., S. J. Kim, …, M. P. Rout. 2016. Structure and function of the nuclear pore complex cytoplasmic mRNA export platform. *Cell.* 167:1215–1228.e25.
10. Upla, P., S. J. Kim, …, J. Fernandez-Martinez. 2017. Molecular architecture of the major membrane ring component of the nuclear pore complex. *Structure.* 25:434–445.
11. Viswanath, S., M. Bonomi, …, E. G. Muller. 2017. The molecular architecture of the yeast spindle pole body core determined by Bayesian integrative modeling. *Mol. Biol. Cell.* https://doi.org/10.1091/mbc.E17-06-0397.
12. Sippl, M. J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–883.
13. Shen, M. Y., and A. Sali. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507–2524.
14. Viswanath, S., D. V. Ravikant, and R. Elber. 2013. Improving ranking of models for protein complexes with side chain modeling and atomic potentials. *Proteins.* 81:592–606.
15. Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225–11236.
16. Brooks, B. R., C. L. Brooks, 3rd, …, M. Karplus. 2009. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* 30:1545–1614.
17. Metropolis, N., and S. Ulam. 1949. The Monte Carlo method. *J. Am. Stat. Assoc.* 44:335–341.
18. Lasker, K., M. Topf, …, H. J. Wolfson. 2009. Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. *J. Mol. Biol.* 388:180–194.
19. Alber, F., B. T. Chait, …, A. Sali. 2008. Integrative structure determination of protein assemblies by satisfaction of spatial restraints. *In* Protein-Protein Interactions and Networks: Identification, Characterization and Prediction. A. Panchenko and T. Przytycka, eds. Springer, London, UK, pp. 99–114.
20. Sali, A., H. M. Berman, …, J. D. Westbrook. 2015. Outcome of the first wwPDB hybrid/integrative methods task force workshop. *Structure.* 23:1156–1167.
21. Schwede, T., A. Sali, …, I. A. Wilson. 2009. Outcome of a workshop on applications of protein models in biomedical research. *Structure.* 17:151–159.
22. Baker, D., and A. Sali. 2001. Protein structure prediction and structural genomics. *Science.* 294:93–96.

23. Tjong, H., K. Gong, …, F. Alber. 2012. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res.* 22:1295–1305.

24. Loquet, A., N. G. Sgourakis, …, A. Lange. 2012. Atomic model of the type III secretion system needle. *Nature.* 486:276–279.

25. Abagyan, R., and M. Totrov. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* 235:983–1002.

26. Zhang, Y., D. Kihara, and J. Skolnick. 2002. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. *Proteins.* 48:192–201.

27. Shen, Y., O. Lange, …, A. Bax. 2008. Consistent blind protein structure generation from NMR chemical shift data. *Proc. Natl. Acad. Sci. USA.* 105:4685–4690.

28. Roy, A., A. Kucukural, and Y. Zhang. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5:725–738.

29. Song, Y., F. DiMaio, …, D. Baker. 2013. High-resolution comparative modeling with RosettaCM. *Structure.* 21:1735–1742.

30. Bhattacharya, D., and J. Cheng. 2015. De novo protein conformational sampling using a probabilistic graphical model. *Sci. Rep.* 5:16332.

31. Zhang, Z., C. E. Schindler, …, M. Zacharias. 2015. Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta. *PLoS One.* 10:e0125941.

32. Yesselman, J. D., and R. Das. 2016. Modeling small noncanonical RNA motifs with the Rosetta FARFAR server. *Methods Mol. Biol.* 1490:187–198.

33. Daura, X., K. Gademann, …, A. E. Mark. 1999. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed. Engl.* 38:236–240.

34. Daura, X., W. F. van Gunsteren, and A. E. Mark. 1999. Folding-unfolding thermodynamics of a $\beta$-heptapeptide from equilibrium simulations. *Proteins.* 34:269–280.

35. Smith, L. J., X. Daura, and W. F. van Gunsteren. 2002. Assessing equilibration and convergence in biomolecular simulations. *Proteins.* 48:487–496.

36. Okur, A., L. Wickstrom, …, C. Simmerling. 2006. Improved efficiency of replica exchange simulations through use of a hybrid explicit/implicit solvation model. *J. Chem. Theory Comput.* 2:420–433.

37. Okur, A., D. R. Roe, …, C. Simmerling. 2007. Improving convergence of replica-exchange simulations through coupling to a high-temperature structure reservoir. *J. Chem. Theory Comput.* 3:557–568.

38. Lyman, E., and D. M. Zuckerman. 2006. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.* 91:164–172.

39. Lyman, E., and D. M. Zuckerman. 2007. On the structural convergence of biomolecular simulations by determination of the effective sample size. *J. Phys. Chem. B.* 111:12876–12882.

40. Grossfield, A., S. E. Feller, and M. C. Pitman. 2007. Convergence of molecular dynamics simulations of membrane proteins. *Proteins.* 67:31–40.

41. Hess, B. 2002. Convergence of sampling in protein simulations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65:031910.

42. Son, W. J., S. Jang, and S. Shin. 2008. A simple method of estimating sampling consistency based on free energy map distance. *J. Mol. Graph. Model.* 27:321–325.

43. Neale, C., W. F. Bennett, …, R. Pomès. 2011. Statistical convergence of equilibrium properties in simulations of molecular solutes embedded in lipid bilayers. *J. Chem. Theory Comput.* 7:4175–4188.

44. Luo, J., P. Cimermancic, …, J. Ranish. 2015. Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol. Cell.* 59:794–806.

45. Chen, R., and Z. Weng. 2002. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins.* 47:281–294.

46. Chen, R., L. Li, and Z. Weng. 2003. ZDOCK: an initial-stage protein-docking algorithm. *Proteins.* 52:80–87.

47. Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika.* 57:97–109.

48. Siegal, S. 1956. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York, NY.

49. McCarroll, D. 2016. Simple Statistical Tests for Geography. CRC Press, Swansea University, UK.

50. Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev. Camb. Philos. Soc.* 82:591–605.

51. Greenland, S., S. J. Senn, …, D. G. Altman. 2016. Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31:337–350.

52. McDonald, J. H. 2009. Handbook of Biological Statistics. Sparky House, Baltimore, MD.

53. Cramer, H. 1946. Mathematical Methods of Statistics. Princeton University Press, Princeton, NJ.

54. Cochran, W. G. 1954. Some methods for strengthening the common $\chi^2$ tests. *Biometrics.* 10:417–451.

55. Kuzmanic, A., and B. Zagrovic. 2010. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophys. J.* 98:861–871.

56. Pettersen, E. F., T. D. Goddard, …, T. E. Ferrin. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25:1605–1612.

57. Hwang, H., T. Vreven, …, Z. Weng. 2010. Protein-Protein Docking Benchmark version 4.0. *Proteins.* 78:3111–3114.

58. Comeau, S. R., D. W. Gatchell, …, C. J. Camacho. 2004. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics.* 20:45–50.

59. Tovchigrechko, A., and I. A. Vakser. 2006. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* 34:W310–W314.

60. Viswanath, S., D. Ravikant, and R. Elber. 2014. DOCK/PIERR: web server for structure prediction of protein-protein complexes. *Methods Mol. Biol.* 1137:199–207.

61. Saltzberg, D. J., H. B. Broughton, …, A. Sali. 2016. A residue resolved Bayesian approach to quantitative interpretation of hydrogen deuterium exchange from mass spectrometry: application to characterizing protein-ligand interactions. *J. Phys. Chem. B.* 121:3493–3501.

62. Wang, X., I. E. Chemmama, …, Y. Ye. 2017. The Proteasome-interacting Ecm29 protein disassembles the 26S proteasome in response to oxidative stress. *J. Biol. Chem.* 292:16310–16320.

63. Zhou, C. Y., C. I. Stoddard, …, G. J. Narlikar. 2017. Regulation of Rvb1/Rvb2 by a Domain within the INO80 chromatin remodeling complex implicates the yeast Rvbs as protein assembly chaperones. *Cell Reports.* 19:2033–2044.