Bioinformatics, 2024, **40(3)**, btae106
https://doi.org/10.1093/bioinformatics/btae106
Advance Access Publication Date: 23 February 2024
Original Paper

OXFORD

Structural bioinformatics

Optimizing representations for integrative structural modeling using Bayesian model selection

Shreyas Arvindekar [1], **Aditi S. Pathak** [1], **Kartik Majila** [1], **Shruthi Viswanath** [1,*]

[1]National Center for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India

*Corresponding author. National Center for Biological Sciences, Tata Institute of Fundamental Research, GKVK, Bellary Road, Bangalore, Karnataka 560065, India. E-mail: shruthiv@ncbs.res.in (S.V.)

Associate Editor: Pier Luigi Martelli

Abstract

Motivation: Integrative structural modeling combines data from experiments, physical principles, statistics of previous structures, and prior models to obtain structures of macromolecular assemblies that are challenging to characterize experimentally. The choice of model representation is a key decision in integrative modeling, as it dictates the accuracy of scoring, efficiency of sampling, and resolution of analysis. But currently, the choice is usually made *ad hoc*, manually.

Results: Here, we report NestOR (**Nest**ed Sampling for **O**ptimizing **R**epresentation), a fully automated, statistically rigorous method based on Bayesian model selection to identify the optimal coarse-grained representation for a given integrative modeling setup. Given an integrative modeling setup, it determines the optimal representations from given candidate representations based on their model evidence and sampling efficiency. The performance of NestOR was evaluated on a benchmark of four macromolecular assemblies.

Availability and implementation: NestOR is implemented in the Integrative Modeling Platform (https://integrativemodeling.org) and is available at https://github.com/isblab/nestor. Data for the benchmark is at https://www.doi.org/10.5281/zenodo.10360718.

1 Introduction

Bayesian integrative structure modeling enables the characterization of structures of large macromolecular assemblies by combining data from several experimental sources at different spatial resolutions, along with physical principles, statistics from previous structures, and prior models (Alber *et al.* 2007, Russel *et al.* 2012, Rout and Sali 2019, Sali 2021). First, information about the system is gathered. Second, a suitable representation is chosen for subunits; atomic and coarse-grained spherical bead representations being common choices. Input information is then translated into spatial restraints that form a Bayesian scoring function, which is then used to rank the sampled models. Third, structural sampling proceeds using the Markov Chain Monte Carlo algorithm. Finally, sampled models are analysed and validated using previously established protocols, which include assessments of sampling exhaustiveness, clustering, and estimation of model precision (Viswanath *et al.* 2017, Pasani and Viswanath 2021, Saltzberg *et al.* 2021, Arvindekar *et al.* 2022, Ullanat *et al.* 2022).

In this work, we present a method to objectively determine the optimal coarse-grained representation for a given system and input information. Model representation is the set of all variables whose values are described by modeling (Sali 2021). It includes the number of states of the system, the stoichiometry, as well as the coarse-graining of subunits. Coarse-graining is the mapping of system atoms to geometrical primitives, e.g. spherical beads. Subunits are coarse-grained to facilitate exhaustive sampling of models of large assemblies as it is prohibitively expensive to use an atomic representation. Commonly, each subunit is coarse-grained into spherical beads that represent one or more contiguous residues along the backbone (Viswanath and Sali 2019).

Representations are currently chosen *ad hoc* and/or by trial and error. However, choosing an appropriate representation is one of the most important decisions in modeling. The choice of representation dictates how accurately input information is translated to spatial restraints, how exhaustive and efficient the sampling of models is, and how useful the resulting models are, for subsequent biological analysis (Viswanath and Sali 2019). A sub-optimal representation can result in incorrect structural models that do not fit the input data well, and possibly require an infeasibly long time for exhaustive sampling of models (Fig. 1, Supplementary Fig. S1). A method to determine optimal coarse-grained representations based on sampling exhaustiveness and model precision has been previously described (Viswanath and Sali 2019). There, the optimal representation is defined as the highest-resolution representation for which sampling was exhaustive at a precision commensurate with the precision of representation. However, the method is computationally expensive. Here, we developed a fully automated, statistically rigorous method based on Bayesian model selection to optimize coarse-grained representation. The optimization is based on the fit to input information.

Bayesian model selection methods have been previously developed for optimizing the molecular dynamics force-field

Received: 12 December 2023; **Revised:** 3 February 2024; **Editorial Decision:** 9 February 2024; **Accepted:** 21 February 2024

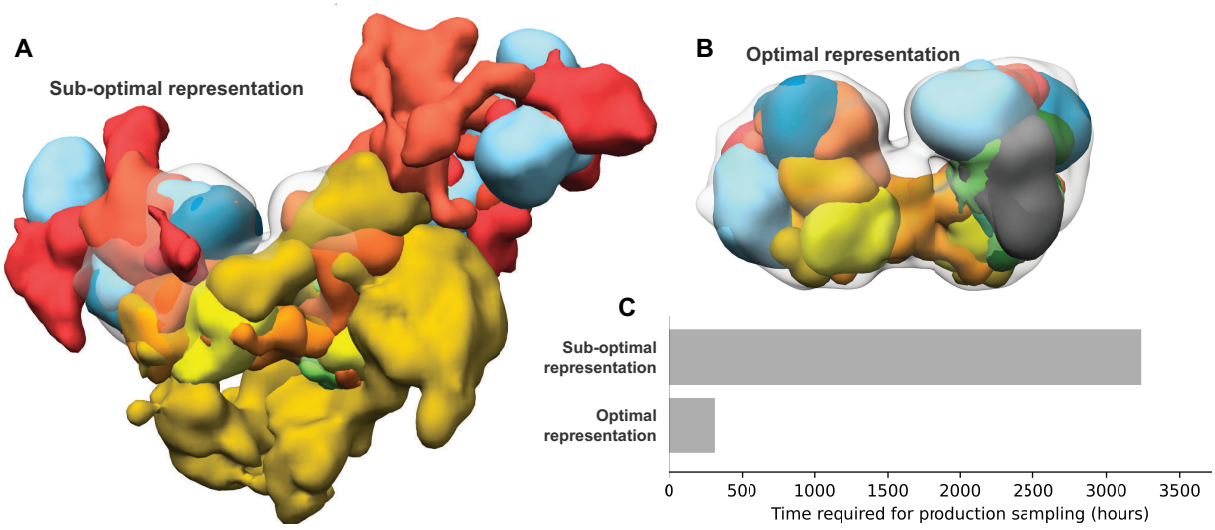


Figure 1. Effects of using sub-optimal representations for integrative modeling. Integrative models of the nucleosome deacetylase (NuDe) complex were produced using two coarse-grained representations, one comprising 1-residue per bead and another comprising 50-residues per bead. The former representation was sub-optimal based on the fit to EM map of the resulting models and the sampling efficiency. Localization probability densities (LPDs) of protein domains of the NuDe complex modeled using coarse-grained representations with (A) 1-residue beads (sub-optimal representation) and (B) 50-residue beads (optimal representation). The densities are superposed on the input EM map (grey, EMDB: 22904) contoured at the recommended threshold. The LPDs are contoured at 10% of their maximum threshold values. (C) Time for production sampling of models based on the two representations.

parameters, coarse-grained representations, and number of states, i.e. conformations, of macromolecules (Bonomi *et al.* 2017, Habeck 2023). (Potrzebowski *et al.* 2018) used model evidence to find the optimal set of states that describe SAS and NMR data. The model evidence was maximized by maximizing its lower bound using a fast variational Bayes approach that required a suitable analytical approximation to the posterior. (Ge and Voelz 2018, Voelz *et al.* 2021) developed BiCePs, a method that reweighs conformations generated from theoretical models based on experimental data. In BiCePs, model selection is performed using a Bayes factor-like term calculated using the computationally intensive method of free-energy perturbation. (Carstens *et al.* 2020) used Bayes factors to obtain the optimal number of chromatin states that explain ensemble Hi-C data. The Bayes factors are calculated by estimating the model evidence using the density of states method that involves sampling a series of annealed posteriors (Habeck 2012).

Here, we used the nested sampling algorithm to compute the model evidence (Skilling 2004, 2006, Ashton *et al.* 2022). It has been applied in astrophysics (Mukherjee *et al.* 2006, Shaw *et al.* 2007, Feroz and Hobson 2008, Higson *et al.* 2019, Ashton *et al.* 2022) and phylogenetics (Russel *et al.* 2019). To our knowledge, we present the first application of nested sampling in macromolecular structure and dynamics. Given an integrative modeling setup, our method, NestOR (**Nest**ed Sampling for **O**ptimizing **R**epresentation), determines the optimal representations from a given set of candidate representations by computing their model evidence and sampling efficiency. To assess the performance of NestOR independently, we prepared a benchmark of four macromolecular assemblies. The performance of the candidate representations was evaluated on this benchmark based on the results of full-length production sampling of models from each candidate representation. We show that NestOR obtains optimal representations for a system at a fraction of the cost required to assess each representation *via* full-length production sampling. The approach is general and could be used to optimize other aspects of representation such as the number of states, number of rigid bodies, and stoichiometry of subunits. Methods such as NestOR would aid in increasing the quality of deposited structures in the wwPDB, as well as contribute to efforts for Bayesian validation of integrative models and input data (Vallat *et al.* 2018, Sali 2021).

2 Materials and methods

We describe Bayesian model selection, nested sampling, and the NestOR algorithm. Subsequently, we describe the benchmarks and the metrics used for selecting the optimal representations.

2.1 Bayes factors and model evidence

Bayesian model selection considers the full prior distribution of parameters for model evaluation. Therefore, it allows for different priors on the same number of parameters to penalize parameters differently (Xie *et al.* 2011). The posterior probability of a statistical model M, given some data D, can be computed by Bayes theorem.

$$P(M|D) \propto P(D|M).P(M)\ldots \quad (1)$$

here P(M|D) is the posterior probability, P(D|M) is the likelihood of the model given the data, and P(M) is the prior on the model. Two models M_1 and M_2 can be compared by the ratio of their posterior probabilities, which is the Bayes factor, K (Jeffreys 1935)

$$K = \frac{P(M_1|D)}{P(M_2|D)} = \frac{P(D|M_1).P(M_1)}{P(D|M_2).P(M_2)}\ldots \quad (2)$$

For our case, the model selection is over coarse-grained representations, e.g. 20 residues-per-bead versus 1 residue-

per-bead. A structural model, M_R, in a representation R, is defined by the spatial coordinates of the beads. The set of all structural models in a representation defines its parameter space.

Since we are comparing representations R_1 and R_2, given D, our Bayes factor is

$$K = \frac{P(R_1|D)}{P(R_2|D)} = \frac{P(D|R_1)}{P(D|R_2)} \dots \quad (3)$$

assuming the two representations are equally likely, i.e. $P(R_1) = P(R_2)$.

$Z_1 = P(D|R_1)$ is the model evidence or marginal likelihood for R_1 and indicates the fit of R_1 to data D on average. $L(M_{R_1})$ and $\pi(M_{R_1})$ corresponds to the likelihood and prior associated with the model in representation R_1.

$$\begin{aligned} Z_1 = P(D|R_1) &= \int P(D|M_{R_1}).P(M_{R_1})dM_{R_1} \\ &= \int L(M_{R_1})\pi(M_{R_1}) \dots \end{aligned} \quad (4)$$

The model evidence naturally considers both goodness-of-fit and model complexity (Oaks *et al.* 2019). Averaging over the entire prior distribution provides a regularizing effect and prevents overfitting to very complex models. If a statistical model is expanded by adding parameters that increase its parameter space significantly, with a corresponding small increase in the high-likelihood region, the model evidence will be lowered, since the likelihood must be averaged over the larger parameter space (Oaks *et al.* 2019). Computing the model evidence for a representation, R_1, for large assemblies is intractable as the integral in (4) is over the high-dimensional parameter space of R_1.

2.2 Nested sampling

Methods to estimate model evidence, such as harmonic and generalized harmonic mean, are simple but may suffer from large errors. Sophisticated methods like thermodynamic integration, are more accurate and involve the sampling of several distributions that bridge from the prior to the posterior (Oaks *et al.* 2019). Here, we used nested sampling, a similar approach that involves sampling from a series of constrained priors (Skilling 2004, 2006).

(4) can be rewritten as (5) given the model parameters for the representation, $\theta = M_{R_1}$, the likelihood $L(\theta) = P(D|M_{R_1})$, and the prior $\pi(\theta) = P(M_{R_1})$

$$Z = \int L(\theta)\pi(\theta)d\theta \dots \quad (5)$$

The key idea in nested sampling is to convert the high dimensional integral in (4) and (5) to 1D over the prior mass, X, $0 \leq X \leq 1$. The prior mass, $X(L_i)$, is the proportion of prior with likelihood greater than a given threshold, L_i.

$$X(L_i) = \int_{L>L_i} \pi(\theta)d\theta \dots \quad (6)$$

(5) becomes (7) and the evidence can be obtained by numerical integration of the L versus X curve (Fig. 2A).

$$Z = \int_0^1 L(X)dX = \sum L(X)\Delta(X) \dots \quad (7)$$

2.3 NestOR

Given candidate representations for a system along with their modeling protocol, NestOR computes the model evidence, its associated uncertainty, and the sampling efficiency for each representation.

2.4 Nested sampling for integrative structural modeling

First, we describe the algorithm to compute a single evidence estimate for a single representation, i.e. one nested sampling run (termed run henceforth) (Fig. 2B). Initially, integrative structural models are sampled from the modified prior (below) using the Replica Exchange MCMC algorithm in IMP and likelihoods are computed for these. These likelihoods form the initial set of "live points". At each iteration, new independent models are sampled from the modified prior using the Replica Exchange MCMC algorithm in IMP to generate a new set of likelihoods (Saltzberg *et al.* 2021, Arvindekar *et al.* 2022). The worst likelihood in the set of live points, L, is replaced with a sample from the new set, subject to the constraint that it is greater than L (sample from the constrained modified prior), and the evidence is incremented. This step is repeated till convergence (below).

2.5 Modified prior and likelihood

Nested sampling requires a likelihood and a prior distribution. In Bayesian integrative structural modeling with IMP, the likelihood is composed of restraints from the experimental data whereas the prior is composed of stereochemistry restraints. The latter is usually uninformative (many models configurations are equally probable) and could be dissimilar to the former. However, evidence computed by nested sampling is inaccurate if the prior is dissimilar to the posterior (Skilling 2004, Buchner 2023). Therefore, we use a modified prior comprising of the stereochemistry restraints and a subset of the restraints from the experimental data, e.g. 30% of the total crosslinks. The remaining restraints from experimental data inform the likelihood. Posterior repartitioning, i.e. redefining the prior and likelihood keeping their product unchanged, has been implemented in nested sampling methods for other applications (Ashton *et al.* 2022).

2.6 Convergence

Either of two conditions must be met for the run to converge. The first condition sets an upper limit on the number of consecutive times the Replica Exchange MCMC sampling fails to sample a new model from the constrained modified prior. The second condition sets an upper limit on the number of consecutive times the Replica Exchange MCMC sampling generates samples from a likelihood plateau.

$$\frac{L_{(i+1)}}{L_i} < \frac{X_i}{X_{(i+1)}} \dots \quad (8)$$

Besides, a run also terminates if it reaches the maximum number of allowed iterations. This is set to a high number to allow the sampling to converge through the first two conditions.

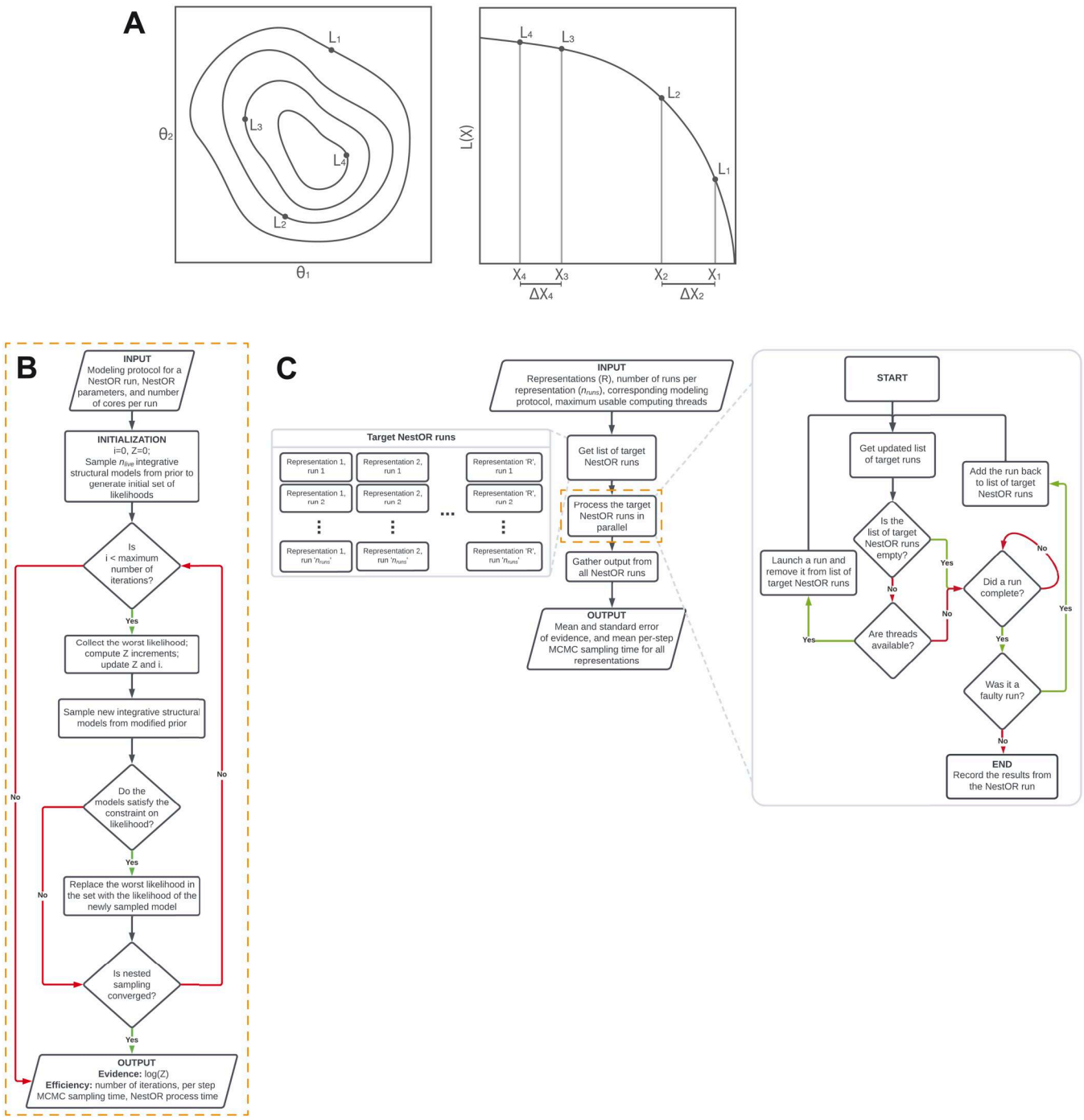


Figure 2. Schematic of nested sampling method for optimizing integrative model representation (NestOR). (A) A schematic representation showing the application of Nested Sampling (NS) to a 2D problem. The iso-likelihood contours for the points with likelihoods L_1, L_2, L_3, and L_4 are shown in the left panel. Their mapping to corresponding prior mass values, X_1, X_2, X_3, and X_4, respectively, is shown in the right panel. The panel to the right represents the L versus X plot for these points. (B) Flowchart describing an individual nested sampling run. Initialized with the modeling protocol, nested sampling parameters, and the number of cores per run, each NestOR run iteratively accumulates model evidence till nested sampling converges. Once converged, it returns the model evidence and measures of efficiency: time taken for a single MCMC step in IMP using the representation (per-step MCMC sampling time), and time taken by NestOR for the run (NestOR process time). (C) Flowchart describing the overall parallelized workflow of NestOR. Given an integrative modeling setup with candidate representations (R), their modeling protocol, the number of runs per representation (n_{runs}), and maximum usable threads, NestOR computes the mean model evidence and the mean per-step model sampling time for all candidate representations in parallel. The results of each independent run per representation, computed in the orange box; described in panel B, are aggregated to produce the mean values from the overall workflow in panel C.

2.7 Collating multiple nested sampling runs

Next, we describe how the results from multiple independent runs for each representation are collated (Fig. 2C). Given the available computing resources (e.g. *the* maximum number of cores), n_{runs} number of runs for each representation are executed in parallel. Runs that fail due to common sampling issues (e.g. large shuffle size), are relaunched. Finally, the mean and standard error, i.e. uncertainty, of the evidence for each representation is computed from the results of the independent runs. The sampling efficiency of a representation is

the average sampling time per Replica Exchange MCMC step across runs. We discuss choices for the number of runs for each candidate representation, and the number of live points for nested sampling and provide recommendations for the user (Supplementary Results).

2.8 Benchmark

We demonstrate the performance of NestOR on a benchmark of four assemblies (Table 1). For each assembly, six candidate representations were explored. These were coarse-grained spherical bead representations where each bead represented a fixed number of contiguous residues along the backbone: 1, 5, 10, 20, 30, and 50, referred to by the number of residues per bead (e.g. representation 50 is composed of 50-residue beads). Additionally, we compared a mixed-resolution representation for NuDe with extended 10-residues per bead representation for MTA1 regions with unknown structure and compact 50-residues per bead representation for other regions with unknown structure.

We performed full-length production sampling and analysis for all candidate representations using previous protocols (Viswanath *et al.* 2017, Brilot *et al.* 2021, Saltzberg *et al.* 2021, Arvindekar *et al.* 2022).

2.9 Metrics for assessing representations

We asked whether the representation enables efficient and exhaustive sampling of good-scoring models, whether the resulting models fit the input information sufficiently well, and whether the resulting model precision is high. The first two criteria for representations were introduced earlier (Viswanath and Sali 2019). We added the last criterion on model precision as a quantitative estimate of the information content of the model for subsequent biological analysis.

For each representation, the sampling efficiency was determined by the wall clock time for each independent production run. The sampling exhaustiveness was measured by the results of four statistical tests on two independent sets of model samples (Viswanath *et al.* 2017). The fit to crosslinks was measured by the average crosslink score in the final cluster of models normalized by the number of crosslinks. The fit to the EM map was measured by the cross-correlation of the localization density to the input EM map. Finally, the model precision is the average RMSD of the cluster models to the cluster centroid model in the final cluster of models (Viswanath *et al.* 2017). Additionally, for RNA polymerase II, we measured the accuracy of the integrative model ensemble with respect to the experimental structure (PDB 1WCM (Armache *et al.* 2005)). We computed the cross-correlation between the localization probability densities of the integrative models and the EM map simulated at 10 Å resolution from the experimental structure.

3 Results

First, we set parameter values for NestOR by comparing the performance of NestOR for different parameter values on the benchmark (Supplementary Results, Supplementary Figs S2–S5). We then demonstrate that NestOR identifies optimal representation(s) on the benchmark. Finally, we show that NestOR is efficient and robust to the choice of prior.

3.1 NestOR identifies optimal representations on a benchmark

We used NestOR to identify optimal representations from the candidate representations for each assembly. NestOR compares candidate representations using two criteria: the model evidence and the sampling time per MCMC step. An optimal representation has high model evidence (i.e. the resulting models fit the input information well) and requires less sampling time per step (i.e. the sampling is efficient). Two representations were ranked the same if their model evidences were similar, i.e. their errors on evidence overlapped, and their model sampling efficiency was similar, i.e. their sampling times were within a 3-fold range of each other.

We assessed the accuracy of NestOR by comparing the optimal representations obtained from it with those identified from full-length production sampling of models. To identify optimal representations from full-length sampling, the comparison of candidate representations was based on passing the score convergence tests for sampling exhaustiveness, the time required for one production run, the fit to input data, and the model precision for the major cluster (Section 2; Viswanath *et al.* 2017). A representation was rejected if it failed to sample a reasonably large number of good scoring models (at least one thousand) and failed to satisfy the sampling exhaustiveness tests for score convergence (Section 2; Viswanath *et al.* 2017). The most sampling-efficient representations were identified as those whose sampling time per run was within a 3-fold range of the most efficient representation;

Table 1. Description of the benchmark complexes.

System	Data used	Region for which the coarse-graining was optimized
γ-Tubulin small complex with an Spc110 monomer (gTuSC) (Brilot *et al.* 2021)	Structures from Cryo-EM and X-ray crystallography, EDC and DSS chemical crosslinks, and stereochemistry information	Regions of unknown structure in the Spc110 N-terminus (Spc1101-220)
RNA polymerase II (Saltzberg *et al.* 2021)	Structures from X-ray crystallography, BS3 and DSS chemical crosslinks, and stereochemistry information	Rpb7
MTA1-HDAC1-MBD3 complex (MHM) (Arvindekar *et al.* 2022)	Structures from X-ray crystallography and NMR spectroscopy, negative staining EM maps, BS3 and DSS chemical crosslinks, and stereochemistry information	Regions of unknown structure in all proteins
Nucleosome deacetylase complex (NuDe) (Arvindekar *et al.* 2022)	Structures from X-ray crystallography and NMR spectroscopy, negative staining EM maps, ADH, BS3, DSS and DMTMM chemical crosslinks, and stereochemistry information	Regions of unknown structure in all proteins

alternate ways to identify sampling-efficient representations can also be used. Finally, representations were considered equally suitable if their model precisions were within 10 Å of each other and their fit to input information was within 10% of each other. The uncertainty in the last two quantities is a result of the uncertainty in data, representation, and stochastic sampling (Schneidman-Duhovny *et al.* 2014, Viswanath and Sali 2019).

3.1.1 gTuSC

The score tests for sampling exhaustiveness indicated that sampling converged for all representations. Representations 5, 10, 20, 30, and 50, i.e. representations 5–50 were equally efficient, whereas representation 1 was much slower (Fig. 3A). The fit to crosslinks and model precision for all representations were about the same based on the above criteria. Overall, representations 5–50 were identified as optimal. NestOR produces representation 1 as the representation with the highest model evidence. But it is much slower than other representations (Fig. 3A). The second-best representation in terms of model evidence, five, is significantly more efficient for sampling. Therefore, representation five can be considered optimal in terms of both model evidence and sampling efficiency.

3.1.2 RNA polymerase II

The score tests indicate that sampling did not converge for representation 1, i.e. <1000 good-scoring models were obtained. Representations 20–50 are the most efficient, whereas representations 5–50 have similar model precision (Fig. 3B). All the representations fit the data equally well. Thus, representations 20–50 were optimal. The optimal representations from the benchmark resulted in better cross-correlations to the experimental structure (Supplementary Fig. S6). NestOR shows that representations 5, 10, 20, and 50 are optimal as they have similar model evidence and sampling efficiency (Fig. 3B).

3.1.3 MHM

The score tests for sampling exhaustiveness indicated that the sampling had converged for all representations. Representations 10–50 are more efficient than others, whereas representations 5, 10, 30, and 50 have similar model precisions (Fig. 3C). All these representations fit the data equally well. Thus, representations 10, 30, and 50 were optimal. From NestOR, representations 30 and 50 are optimal as they have the highest evidence and sampling efficiency (Fig. 3C).

3.1.4 NuDe

The score tests indicate sampling did not converge for representation 1. Representations 10–50 and the mixed representation are the most efficient (Fig. 3D). Representations 5–50 and the mixed representation have similar model precisions and fit to data. Therefore, representations 10–50 and the mixed representation were optimal. From NestOR, representations 30 and 50 are optimal as they have the best evidence and sampling efficiency (Fig. 3D).

In summary, several coarse-grained representations are equally good based on the above metrics, and more than one optimal representation can exist for a given system. We visualized integrative models for gTuSC and RNA polymerase II at different coarse-grained representations (Supplementary Fig. S7).

The representations returned by NestOR are among the optimal representations for a given system. The only exception was RNA polymerase II where one of the three optimal representations from NestOR was not optimal based on the benchmark. In this case, multiple representations had similar evidence and the errors in evidence were large. This could be because the differences between candidate representations were small for this system: the representation of only one of twelve proteins was varied and the positions of ten proteins were fixed throughout.

In particular, for systems with extended proteins, such as the disordered Spc110 N-terminus in gTuSC, the model evidence favors finer 1- and 5-residue representations (Fig. 3) (Brilot *et al.* 2021). Whereas, for more compact and globular systems such as NuDe and MHM, the model evidence favors coarser 20- and 50-residue representations (Fig. 3). These results highlight that model evidence balances the representation complexity with the fit to data to determine the optimal representation for a given system and data.

3.2 NestOR is efficient

A comparison of the total time for full-length production sampling for all candidate representations with the time required to run NestOR shows that NestOR takes a fraction of the time required by the former (Fig. 4). Further, the time for production sampling shown here is exclusive of the time for analysis, which could add several more hours. Overall, NestOR is efficient and takes a few hours on a modern multicore workstation (Fig. 4).

3.3 NestOR is robust to the choice of prior

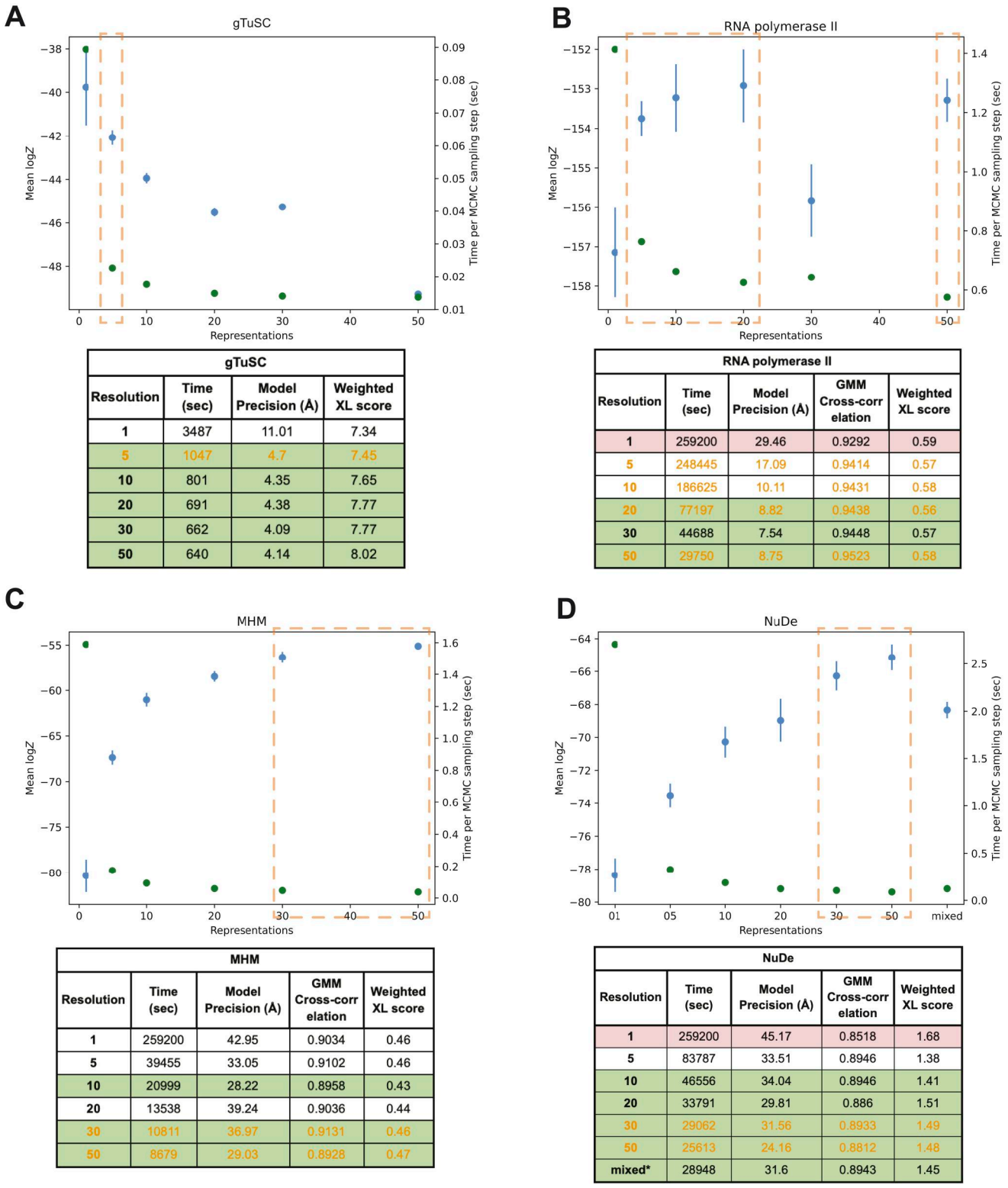
NestOR requires the restraints to be split into two mutually exclusive sets—for prior and likelihood estimation (Section 2). We examined the effect of choice of prior on the estimated evidence. We partitioned the input crosslink restraints for MHM and NuDe into prior and likelihood estimation sets thrice at random and compared the evidence estimated by NestOR. The trend in evidence remained largely conserved, although the actual values of computed evidence were slightly different for these runs (Fig. 5), which could be attributed to the uncertainties associated with the input data, representation, and stochastic sampling (Schneidman-Duhovny *et al.* 2014).

4 Discussion

Here, we report NestOR, a fully objective Bayesian model selection method to identify the optimal representation from a set of candidate representations for a given integrative modeling setup. NestOR obtains optimal representations for a system at a fraction of the cost required to assess each representation via full-length production sampling. We discuss parameter settings for NestOR, its advantages, disadvantages, alternate implementations, and future directions.

4.1 Recommended parameters

It is recommended that users initially start with 5 independent runs (Supplementary Results, Supplementary Figs S1 and S2). The number of live points should ideally be at least the number of free parameters (Ashton *et al.* 2022). For integrative modeling, the number of free parameters is less than



gTuSC

Resolution	Time (sec)	Model Precision (Å)	Weighted XL score
1	3487	11.01	7.34
5	1047	4.7	7.45
10	801	4.35	7.65
20	691	4.38	7.77
30	662	4.09	7.77
50	640	4.14	8.02

RNA polymerase II

Resolution	Time (sec)	Model Precision (Å)	GMM Cross-corr elation	Weighted XL score
1	259200	29.46	0.9292	0.59
5	248445	17.09	0.9414	0.57
10	186625	10.11	0.9431	0.58
20	77197	8.82	0.9438	0.56
30	44688	7.54	0.9448	0.57
50	29750	8.75	0.9523	0.58

MHM

Resolution	Time (sec)	Model Precision (Å)	GMM Cross-corr elation	Weighted XL score
1	259200	42.95	0.9034	0.46
5	39455	33.05	0.9102	0.46
10	20999	28.22	0.8958	0.43
20	13538	39.24	0.9036	0.44
30	10811	36.97	0.9131	0.46
50	8679	29.03	0.8928	0.47

NuDe

Resolution	Time (sec)	Model Precision (Å)	GMM Cross-corr elation	Weighted XL score
1	259200	45.17	0.8518	1.68
5	83787	33.51	0.8946	1.38
10	46556	34.04	0.8946	1.41
20	33791	29.81	0.886	1.51
30	29062	31.56	0.8933	1.49
50	25613	24.16	0.8812	1.48
mixed*	28948	31.6	0.8943	1.45

Figure 3. Performance of NestOR on the benchmark. Candidate coarse-grained representations of uniform resolutions (1-, 5-, 10-, 20-, 30-, 50- residues per bead) are compared for each system (A. gTuSC, B. RNA polymerase II, C. MHM and D. NuDe). In addition, a mixed resolution representation was evaluated for NuDe. The output of NestOR, i.e. the mean of log model evidence and its standard error (blue), and the mean time per Replica Exchange MCMC step (green) is plotted for each system. Based on these two criteria, the optimal representation(s) inferred from NestOR are highlighted in orange dashed boxes. The tables accompanying each plot show the results from full-length production sampling for each candidate representation for each system: the time required per independent sampling run, model precision, and fit to data based on the average crosslink score in the major cluster, and the cross-correlation of the EM map with the localization densities of the major cluster. The optimal representations based on the results from full-length production sampling are highlighted in green, whereas representations for which sampling was not exhaustive in the given time are in red. All times are on an AMD Ryzen Threadripper 3990X 64-Core Processor with 256 GB RAM and 2.2 GHz clock speed. Four computing threads were used for each system, except for gTuSC where six threads were used.

$3n + 6m$, where n is the number of flexible beads with three degrees of freedom each and m is the number of rigid bodies with six degrees of freedom each. We used 50 live points for NestOR on the benchmark; higher numbers of live points did not change the ranking of representations based on evidence (Supplementary Results, Supplementary Figs S3 and S4). In some cases, the evidence is associated with overlapping errors, making it difficult to rank representations. The errors on evidence can be decreased by increasing the number of live points, number of runs per representation, and number of models sampled per iteration, in this order.

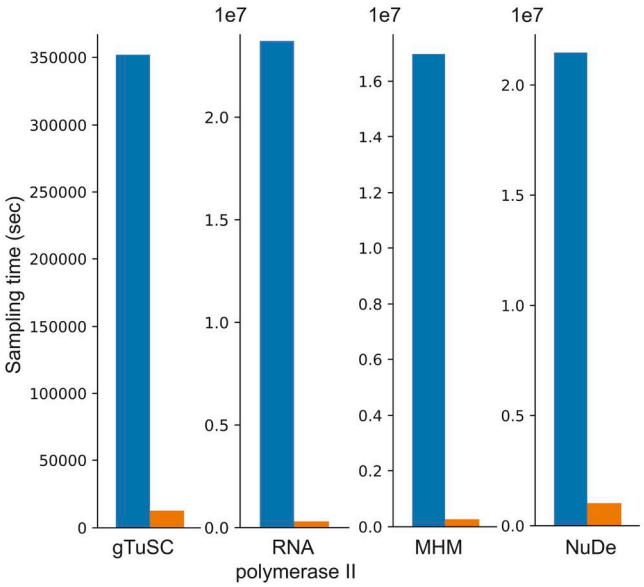


Figure 4. NestOR efficiency. The total time required for full-length production sampling of models using all candidate representations for each system (blue) is compared with the total time required by NestOR (orange). Production sampling consisted of 50 (28) independent Replica Exchange MCMC runs for gTuSC, MHM, and NuDe (RNA polymerase II). NestOR was run with previously set parameters (5 runs, 50 live points, 50 RE-MCMC steps per iteration) for each candidate representation till a convergence criterion was met. All times are on an AMD Ryzen Threadripper 3990X 64-Core Processor with 256 GB RAM and 2.2 GHz clock speed.

We recommend repartitioning the posterior to improve the overlap between prior and likelihood distributions for nested sampling (Section 2). We recommend biasing the stereochemistry prior by a small subset of the experimental restraints that are sensitive to coarse-graining, such as crosslinking restraints. This allows the evidence to be estimated from the majority of the experimental restraints. Restraints that are often not sensitive to small changes in the coarse-graining, such as shape restraints on low-resolution EM maps, can be used to inform the likelihood along with other experimental restraints.

4.2 Advantages, disadvantages, and considerations

NestOR is fast and parallelized and takes a fraction of the time required by production runs. The use of Bayesian model selection via model evidence naturally balances representation complexity (prior) with goodness of fit (likelihood). It is especially important to avoid overfitting to a particular set of data when performing inference with noisy, sparse, and ambiguous data (Voelz *et al.* 2021). Unlike annealing-based methods, one does not need to tune an annealing schedule to bridge between the prior and posterior. NestOR can be used to assess uniform-resolution as well as mixed-resolution coarse-grained representations (Fig. 3). Apart from the coarse-graining, other aspects of representation such as the number of rigid bodies and stoichiometry can also be optimized using the approach.

There are three disadvantages of the method. First, it is difficult to compare setups where the difference in the evidence is likely to be small, such as, when the difference between candidate representations is small, e.g. the representation of only one of several proteins is altered, and/or when the data on the regions for which representations are being optimized is sparse. In the former case, the major contribution to the evidence will be from the regions whose representation is not changing. In the latter case, the evidence estimates will be noisy (high error bars). This was seen in RNA polymerase II, where the representation of only one of twelve proteins was varied across candidates (Fig. 2). The remaining two disadvantages are general to nested sampling; evidence estimates are inaccurate if the prior and likelihood distributions are dissimilar and if there are plateaus in the likelihood (Skilling 2004, 2006, Ashton *et al.* 2022).

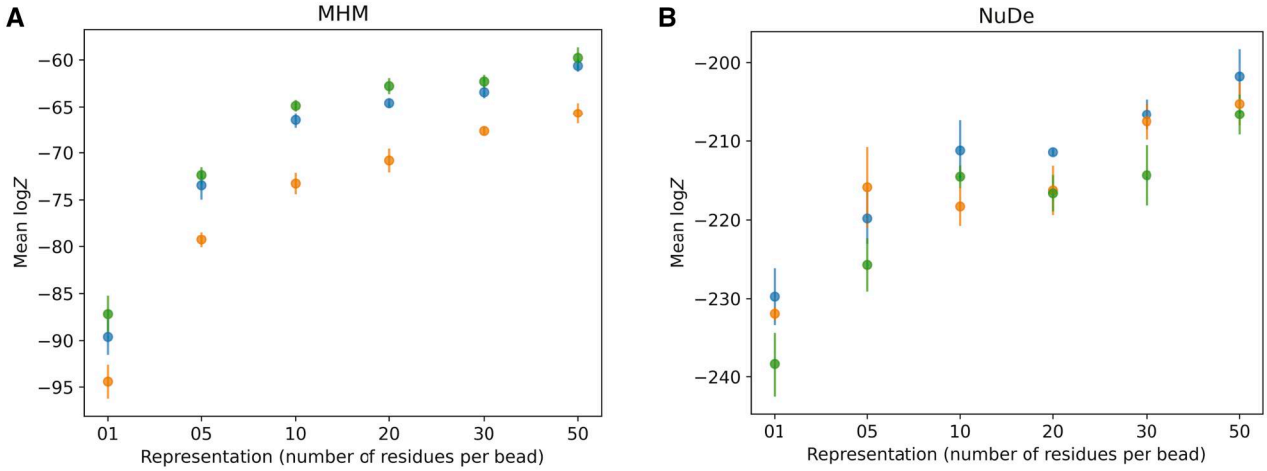


Figure 5. Robustness to the choice of prior. NestOR outputs, i.e. the evidence estimates and associated uncertainties, were compared for three different priors (orange, green, blue), on two systems, (A) MHM, and (B) NuDe. Each prior comprised a random subset of 30% of a set of input crosslinks, in addition to stereochemistry restraints.

One needs to consider these points before using NestOR. First, NestOR only evaluates the set of user-provided candidate representations based on their model evidence; it does not search for optimal representations. It is assumed that the input representations are compatible with the experimental data available, and therefore the restraints used, for modeling. For example, the forward models and restraints for NMR and SAXS do not support coarse-grained representations. Second, the restraints used for evidence calculation should account for the effect of the choice of representation on the forward model. In the present work, the uncertainty in the coarse-grained representation is accounted for, in the forward model of the Bayesian XLMS (Shi *et al.* 2014) and EM restraints (Bonomi *et al.* 2019) that were used, through the structural uncertainty parameter, σ, in the former and the formulation of model GMMs in the latter. The forward model for other restraints may need to be modified to account for this.

The predictive value of an integrative model is determined by both, its resolution and its precision (global and local). Often, the highest resolution and precision are desired. However, the highest resolution representation for a region may not result in the most precise model. The precision is limited, in part, by the quality, quantity, and coverage of available information for the region being modeled. For example, the use of one residue per bead for modeling a large macromolecular assembly with sparse structural information is not ideal, even if one requires residue-level predictions. It is often infeasible to exhaustively sample models at this representation. Further, even exhaustive sampling might result in imprecise models owing to the uncertainty in the input information. NestOR compares candidate representations based on their fit to data and model complexity and penalizes overly complex representations that are not justified by the data.

4.3 Alternative implementations and future improvements

Alternate ways to obtain live points in each iteration can be used. Multiple live points can be sampled from the constrained prior at each iteration instead of a single one; although one replacement per iteration is considered optimal (Ashton *et al.* 2022). An existing live point can be used to evolve a new point using MCMC. Existing live points can be used to generate live points using genetic algorithms. More sophisticated approaches that adaptively vary the number of live points can be used (Higson *et al.* 2019). The widths, ΔX can also be sampled instead of using an approximation (Skilling 2006, Ashton *et al.* 2022).

Optimizing the number of rigid bodies, stoichiometry, number of states, and other aspects of representation are future avenues that need to be explored. In a future Bayesian implementation, representations and models can be sampled simultaneously.

Methods like NestOR will aid in improving the quality of integrative structures archived in the wwPDB (Vallat *et al.* 2018, Sali 2021). This effort is timely and relevant, given the imminent increase in the number of deposited integrative structures and the merging of the PDB-DEV database for integrative model deposition (https://pdb-dev.wwpdb.org) with the wwPDB. Bayesian model selection-based methods like NestOR will be essential for ongoing efforts on Bayesian validation of all structures and data in the PDB (https://cdn.rcsb.org/wwpdb/docs/about/advisory/wwpdbac2021-report.pdf). Finally, they will also contribute to new metamodeling efforts to combine models by suggesting criteria that make the deposited models as well as the modeling methods more rigorous (Sali 2021).

Acknowledgements

We thank lab members Muskaan Jindal, Arijit Das, and Sakshi Shigvan, for useful comments on the manuscript. We are grateful to Andrej Šali, Barak Raveh, and Greg Voth for their comments on the work presented here. We thank Ben Webb for his help with integrating NestOR with IMP. Molecular graphics images were produced using the UCSF Chimera and UCSF ChimeraX packages from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR001081, NIH R01-GM129325, and National Institute of Allergy and Infectious Diseases).

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was supported by the following grants: Department of Atomic Energy (DAE) TIFR grant [RTI 4006] and Department of Science and Technology (DST) Science Engineering and Research Board (SERB) grant [SPG/2020/000475] from the Government of India to S.V.

Data availability

The benchmark of four macromolecular assemblies characterized using IMP is available at https://www.doi.org/10.5281/zenodo.10360718. NestOR scripts and usage instructions are available at https://github.com/isblab/nestor and will be integrated into IMP (https://integrativemodeling.org).

References

Alber F, Dokudovskaya S, Veenhoff LM *et al.* Determining the architectures of macromolecular assemblies. *Nature* 2007;**450**:683–94.

Armache K-J, Mitterweger S, Meinhart A *et al.* Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. *J Biol Chem* 2005;**280**:7131–4.

Arvindekar S, Jackman MJ, Low JKK *et al.* Molecular architecture of nucleosome remodeling and deacetylase Sub-complexes by integrative structure determination. *Protein Sci* 2022;**31**:e4387.

Ashton G, Bernstein N, Buchner J *et al.* Nested sampling for physical scientists. *Nat Rev Methods Primers* 2022;**2**:39.

Bonomi M, Hanot S, Greenberg CH *et al.* Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* 2019;**27**:175–88.e6.

Bonomi M, Heller GT, Camilloni C *et al.* Principles of protein structural ensemble determination. *Curr Opin Struct Biol* 2017;**42**:106–16.

Brilot AF, Lyon AS, Zelter A *et al.* CM1-driven assembly and activation of yeast γ-tubulin small complex underlies microtubule nucleation. *Elife* 2021;**10**:e65168.

Buchner J. Nested sampling methods. *Stat Surv* 2023;**17**:169–215.

Carstens S, Nilges M, Habeck M. Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc Natl Acad Sci USA* 2020;**117**:7824–30.

Feroz F, Hobson MP. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Monthly Notices R Astronomical Soc* 2008; **384**:449–63.

Ge Y, Voelz VA. Model selection using BICePs: a bayesian approach for force field validation and parameterization. *J Phys Chem B* 2018;**122**:5610–22.

Habeck M. Evaluation of marginal likelihoods via the density of states. In: *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, La Palma, Canary Islands, JMLR: W & CP, 2012, 486–94.

Habeck M. Bayesian methods in integrative structure modeling. *Biol Chem* 2023;**404**:741–54.

Higson E, Handley W, Hobson M *et al.* Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *Stat Comput* 2019;**29**:891–913.

Jeffreys H. Some tests of significance, treated by the theory of probability. *Math Proc Camb Phil Soc*1935;**31**:203–22.

Mukherjee P, Parkinson D, Liddle AR. A nested sampling algorithm for cosmological model selection. *ApJ* 2006;**638**:L51–4.

Oaks JR, A Cobb K, N Minin V *et al.* Marginal likelihoods in phylogenetics: a review of methods and applications. *Syst Biol* 2019; **68**:681–97.

Pasani S, Viswanath S. A framework for stochastic optimization of parameters for integrative modeling of macromolecular assemblies. *Life* 2021;**11**:1183.

Potrzebowski W, Trewhella J, Andre I. Bayesian inference of protein conformational ensembles from limited structural data. *PLoS Comput Biol* 2018;**14**:e1006641.

Rout MP, Sali A. Principles for integrative structural biology studies. *Cell* 2019;**177**:1384–403.

Russel D, Lasker K, Webb B *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* 2012;**10**:e1001244.

Russel PM, Brewer BJ, Klaere S *et al.* Model selection and parameter inference in phylogenetics using nested sampling. *Syst Biol* 2019; **68**:219–33.

Sali A. From integrative structural biology to cell biology. *J Biol Chem*2021;**296**:100743.

Saltzberg DJ, Viswanath S, Echeverria I *et al.* Using integrative modeling platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci* 2021;**30**:250–61.

Schneidman-Duhovny D, Pellarin R, Sali A. Uncertainty in integrative structural modeling. *Curr Opin Struct Biol* 2014;**28**:96–104.

Shaw JR, Bridges M, Hobson MP. Efficient bayesian inference for multimodal problems in cosmology. *Monthly Notices R Astronomical Soc* 2007;**378**:1365–70.

Shi Y, Fernandez-Martinez J, Tjioe E *et al.* Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol Cell Proteomics* 2014;**13**:2927–43.

Skilling J. Nested sampling. *AIP Conf Proc* 2004;**735**:395–405.

Skilling J. Nested sampling for general Bayesian computation. *Bayesian Anal* 2006;**1**:833–59.

Ullanat V, Kasukurthi N, Viswanath S. PrISM: precision for integrative structural models. *Bioinformatics* 2022;**38**:3837–9.

Vallat B, Webb B, Westbrook JD *et al.* Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* 2018;**26**:894–904.e2.

Viswanath S, Chemmama IE, Cimermancic P *et al.* Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophys J* 2017;**113**:2344–53.

Viswanath S, Sali A. Optimizing model representation for integrative structure determination of macromolecular assemblies. *Proc Natl Acad Sci USA* 2019;**116**:540–5.

Voelz VA, Ge Y, Raddi RM. Reconciling simulations and experiments with BICePs: a review. *Front Mol Biosci* 2021;**8**:661520.

Xie W, Lewis PO, Fan Y *et al.* Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst Biol* 2011; **60**:150–60.