ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理
Chinese Physics B

The determinant factors for map resolutions obtained using CryoEM single particle imaging method

Yihua Wang(王义华), Daqi Yu(余大启), Qi Ouyang(欧阳颀), Haiguang Liu(刘海广)

Citation: Chin. Phys. B . 2018, 27(12): 128702. **doi:** 10.1088/1674-1056/27/12/128702
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Applications of nanostructures in wide-field, label-free super resolution microscopy

Xiaowei Liu(刘小威), Chao Meng(孟超), Xuechu Xu(徐雪初), Mingwei Tang(汤明炜), Chenlei Pang(庞陈雷), Qing Yang(杨青)

Chin. Phys. B . 2018, 27(11): 118704. **doi:** 10.1088/1674-1056/27/11/118704

Structural biology revolution led by technical breakthroughs in cryo-electron microscopy

Chang-Cheng Yin(尹长城)

Chin. Phys. B . 2018, 27(5): 058703. **doi:** 10.1088/1674-1056/27/5/058703

Single-shot grating-based x-ray differential phase contrast imaging with a modified analyzer grating

Chen-Xi Wei(卫晨希), Zhao Wu(吴朝), Faiz Wali, Wen-Bin Wei(魏文彬), Yuan Bao(鲍园), Rong-Hui Luo(骆荣辉), Lei Wang(王磊), Gang Liu(刘刚), Yang-Chao Tian(田扬超)

Chin. Phys. B . 2017, 26(10): 108701. **doi:** 10.1088/1674-1056/26/10/108701

Investigation of noise properties in grating-based x-ray phase tomography with reverse projection method

Bao Yuan, Wang Yan, Gao Kun, Wang Zhi-Li, Zhu Pei-Ping, Wu Zi-Yu

Chin. Phys. B . 2015, 24(10): 108702. **doi:** 10.1088/1674-1056/24/10/108702

The laser-intensity dependence of the photoassociation spectrum of the ultracold $Cs_2(6S_{1/2}+6P_{1/2})$ 0_u^+ long-range molecular state

Jin Li, Feng Guo-Sheng, Wu Ji-Zhou, Ma Jie, Wang Li-Rong, Xiao Lian-Tuan, Jia Suo-Tang

Chin. Phys. B . 2013, 22(8): 088701. **doi:** 10.1088/1674-1056/22/8/088701

The determinant factors for map resolutions obtained using CryoEM single particle imaging method*

Yihua Wang(王义华)[1], Daqi Yu(余大启)[1], Qi Ouyang(欧阳颀)[1], and Haiguang Liu(刘海广)[2,†]

[1] *Key Laboratory for Artificial Microstructure and Mesoscopic Physics, Institute of Condensed Matter Physics, School of Physics, Center for Quantitative Biology School of Physics, The Peking–Tsinghua Center for Life Sciences at School of Physics, Peking University, Beijing* 100084, *China*

[2] *Complex Systems Division, Beijing Computational Science Research Centre, Beijing* 100193, *China*

(Received 12 July 2018; revised manuscript received 8 October 2018; published online 10 November 2018)

The CryoEM single particle structure determination method has recently received broad attention in the field of structural biology. The structures can be resolved to near-atomic resolutions after model reconstructions from a large number of CryoEM images measuring molecules in different orientations. However, the determining factors for reconstructed map resolution need to be further explored. Here, we provide a theoretical framework in conjunction with numerical simulations to gauge the influence of several key factors to CryoEM map resolutions. If the projection image quality allows orientation assignment, then the number of measured projection images and the quality of each measurement (quantified using average signal-to-noise ratio) can be combined to a single factor, which is dominant to the resolution of reconstructed maps. Furthermore, the intrinsic thermal motion of molecules has significant effects on the resolution. These effects can be quantitatively summarized with an analytical formula that provides a theoretical guideline on structure resolutions for given experimental measurements.

Keywords: Cryo-EM, resolution, noise, conformation heterogeneity

PACS: 87.64.Ee, 87.57.cf, 87.57.cm, 87.15.hp **DOI:** 10.1088/1674-1056/27/12/128702

1. Introduction

The Cryo-electron microscopy (CryoEM) single particle imaging method has become popular recently in the structural biology research community.[1] The basic idea of CryoEM is to measure the particles at all possible orientations and use computational model reconstruction algorithms to build a three-dimensional (3D) structure that best satisfies the overall measurements. Due to irradiation damage from high-energy electrons during the measurements, each particle or molecule can only tolerate certain amount of electron doses before the molecule deteriorates. Experimentally, each particle/molecule is only measured once at a given orientation that is nearly fixed in vitreous ice. The CryoEM single particle imaging method spreads the electron dosage to a large ensemble of molecules, and each scatters a tolerable number of electrons to form a magnified image. Meanwhile, the cryogenic environment protects the sample molecules, maintaining molecular integrity. Nonetheless, the model resolutions from CryoEM experiments were not close to those obtained from the x-ray crystallography method until the recent breakthrough in three aspects, namely: (i) the invention of a direct electron detector to allow accurate and fast measurement of electrons;[2] (ii) the development of data processing software, in particular the application of Bayesian algorithms in reconstructions, backed by high-performance computers;[3–5] and (iii) the advances in sample preparations that allow measurements at diverse orientations at improved signal levels using thin vitreous ice layers.[6–8] The fast readout rate of new direct electron detectors also enables measurements in movie modes that lead to the correction of molecular drift during data collection to sharpen the blurred signals.[9,10] For a long time, only the particles with high symmetries could be determined to high resolutions using CryoEM single particle imaging method, such as virus particles.[11–13] However, since the structural determination of the TRPV1 molecule at 3.4 Å,[14] many high-resolution structures of molecular complexes have been determined using the CryoEM single particle imaging method. This technology is enriching in the protein structure database, particularly with large molecular complexes.[15–21]

Despite the advances in CryoEM single particle imaging method, some fundamental questions remain. One question we would like to address here is regarding the determinant factors for model resolution. In the crystallography method, the concept and measures of resolution have been well established,[22,23] while they are still under investigation in CryoEM. In general, the resolutions for the maps determined using the CryoEM approach are estimated using model consistency; *i.e.*, by calculating the Fourier shell correlation (FSC) profiles and examining the point where the signal disappears.[24–27] Recently, several alternative methods have been developed to assess the map resolutions, such as the approach that checks the local details of the structure.[28]

*Project supported by the National Natural Science Foundation of China (Grant Nos. 11774011, 11434001, U1530401, and U1430237).

†Corresponding author. E-mail: hgliu@csrc.ac.cn

© 2018 Chinese Physical Society and IOP Publishing Ltd

For maps determined to near atomic levels (better than 3 Å), the reconstructed electron density maps can be visually inspected to check the model accuracy. Regardless of the different definitions in model resolution, the correct interpretation of the reconstructed models is subject to validation using complementary approaches, such as biochemical assays or single molecule fluorescence experiments. Putting aside the arguments on CryoEM map resolutions using different approaches, we would like to focus on the factors that determine the model resolution and hope to obtain a theoretical framework that guides the experiments to improve the resolution using optimized protocols for data collection.

In this work, we investigated four factors that influence model resolutions, namely, the number of projection measurements, the signal-to-noise ratio (SNR) of individual measurements, and the intrinsic flexibility of molecules. Early studies have provided important clues about how these factors may contribute to the model resolutions. For example, Henderson studied the resolution limits resulted from electron microscopy and provided a relation between the resolution and the number of projections.[29] Later, Rosenthal and Henderson formulated a more detailed equation (noted RH model hereafter) to estimate the desired number of projections for the different structure resolutions.[26] In the RH model, the electron dose, SNR, molecular symmetry, and an effective B-factor were considered.[26,30,31] The effective B-factor was found to be essential to fit to experimental data because it is used to model the combined effects of molecular drifting due to charging effects, molecular flexibility, errors in image processing, and so on, into a Gaussian envelope function that describes the signal falloff.[30,31] Based on this pioneer research, we would like to revisit these relations and validate the formulations using numerical simulations. Furthermore, it is known that many molecules undergo conformational changes to be functional. To solve structures at higher resolutions, the molecules can be locked in a particular conformation. For example, Subramaniam and coworkers used a cell-permanent inhibitor to stabilize β-galactosidase and obtained a CryoEM structure at 2.2 Å.[32] In another work, the same group obtained a structure of glutamate dehydrogenase to 1.8 Å after detailed projection classifications by sorting out the images that belong to the most populated conformation.[33] Here, we set out to investigate the effects of molecular intrinsic motion using a structure ensemble to simulate CryoEM single particle images, taking the heterogeneous conformation reality under consideration. Consequently, we proposed a framework using these factors to predict structure the resolutions. The numerical simulation results were used to estimate free parameters. The statistics from the resolved structures are consistent with the proposed model.

2. Method and theory

2.1. Existing theoretical framework

The model originally proposed by Rosenthal and Henderson (the RH model) and later elaborated by Liao and Frank connects several key factors in the CryoEM method using the following equation:[26]

$$N(k) = \frac{C\sigma^2}{1/n_k \sum_{r=1}^{n_k} |F_T^r(k)|^2} k \exp\left(\frac{Bk^2}{2}\right), \tag{1}$$

where N is number of particles (projections) that are needed to reach the resolution defined by the Fourier frequency k; σ^2 is the variance of the noise, $F_T^r(k)$ is the Fourier intensity of the map in the resolution shell $[k, k+\Delta k]$, B is the effective temperature factor, and C is a scaling constant.

The SNR (defined as the ratio between variances of signals and noises) at a resolution shell k is effectively represented by $\left(1/n_k \sum_{r=1}^{n_k} |F_T^r|^2\right)/\sigma^2$, where the numerator resembles the rotational average of the intensity (similar to the small/wide angle scattering intensity) and the denominator is the noise level. The second term on the right side of Eq. (1), k, describes the linear dependency of $N(k)$ on the resolution shell k because the number of data points in each two-dimensional (2D) projection is at the order of k^2, while the desired number of data points in 3D increases with k^3. The last term is the Gaussian falloff to account factors, such as structural fluctuations or misalignment during data processing. We carried out numerical simulations to examine the effects of each component. Based on the results, we revise the formula to provide an improved prediction formula on map resolutions.

2.2. Data simulation

The structure of GroEL (PDB ID: 1XCK,[34]) was used as the model system in the numerical simulations (see Fig. 1). We present two theoretical formulations to investigate the factors that influence the map resolutions, the Gaussian noise (GN) model and the Thermal Fluctuation (TF) model, as shown in Fig. 1.

The Gaussian noise (GN) model was proposed to describe the dependency of the model resolution on the number of projections and the noise variance.

$$N(k) = A_{\mathrm{GN}} \sigma^2 k \mathrm{e}^{Bk^2}, \tag{2}$$

where $N(k)$ is the desired number of particles (projections) to collect confident signal levels at a frequency k; σ^2 is the variance of the Gaussian noise; and B is a scaling parameter.

The TF model describes how thermal fluctuations of the molecules influence the model resolution. RMSD (root-mean-square-deviation) is one quantity that measures the difference between the structures, and here we used the mean square of the pairwise RMSD of an structure ensemble to quantify the structure fluctuations and to mimic the Debye–Waller factor.

The following formula is proposed to relate the thermal fluctuation levels, number of projection images and achievable map resolution:

$$N(k) = A_{\mathrm{TF}} k \, \mathrm{e}^{C\langle RMSD^2\rangle k^2} \tag{3}$$

where $\langle RMSD^2\rangle$ is the mean square of the *RMSD* values obtained by pairwise comparison within the structure ensemble; and k is the spatial frequency. In both Eqs. (2) and (3), A_{GN}, A_{TF}, B, and C are free parameters.

For the Gaussian noise model, SPIDER 22.03[35] packages were used to generate the simulation data. The atomic model of GroEL was first converted to a density map (voxel size = $(0.86\ \text{Å})^3$), and then projection images were simulated at orientations generated with the successive orthogonal rotation sample approach.[36] The noises were incorporated according to the desired SNR following a Gaussian distribution after the contrast transfer function (CTF) for defocus of range from 1.0 μm to 3.0 μm was convoluted to the simulated projections. In this case, the CTF was not modulated by the envelope function. The image simulation process is summarized in Fig. 2(a). More specific procedure in mathematical phrase is as the following: (i) a 3D map is rendered by voxelization of an atomic model; (ii) generate projection images at orientations that uniformly distributed in the SO3 space; (iii) convert the projection images to Fourier space using FFT, then multiplies the CTF function for a random defocus level between 1.0 μm to 3.0 μm; (iv) apply inverse Fourier transform to get the CTF convoluted projection image; and (v) add desired Gaussian noise to get the final projection image.

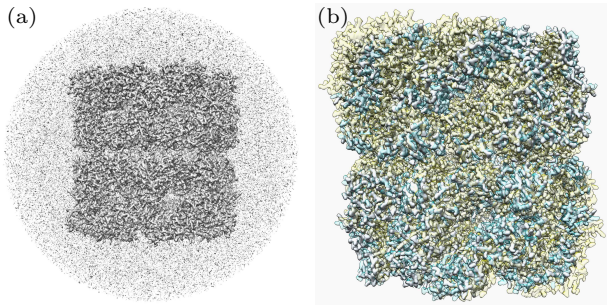


Fig. 1. (color online) The two models used for CryoEM single particle imaging data simulations. (a) Gaussian noise model: noises were added to the simulation data to study the influence of noise on model resolution. (b) Thermal fluctuation model: a structure ensemble was first generated to mimic the structure heterogeneity. Each single particle projection was simulated based on a randomly picked structure from the ensemble. Three representative models for GroEL are shown in panel (b): the original structure is in gray; the yellow and blue structures were generated using normal mode perturbations.

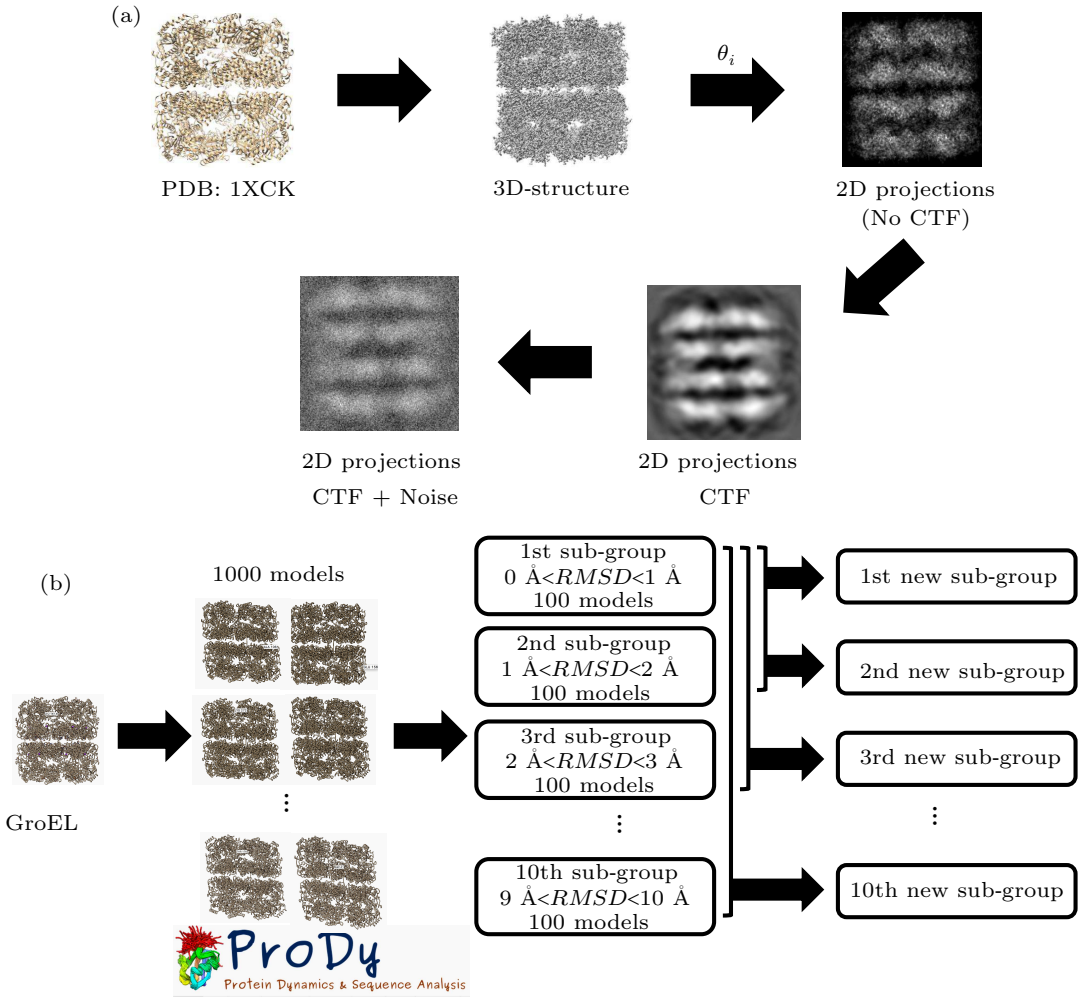


Fig. 2. (color online) The workflow for single particle imaging data stimulation. (a) The protocol for the projection data simulation with defocused lenses and Gaussian noises from a single 3D structure. (b) The procedure for structure ensemble generation to model the heterogeneity of biomolecular conformations.

To simulate the heterogeneity in the structures, we first obtained a set of diverse structures to form a structure ensemble based on the GroEL structure. Without the loss of generality, the structural ensemble was generated using the normal mode perturbation approach. The 'ProDy' program based on an anisotropic elastic network model was used to compute the normal modes and the eigenvalue spectrum.[37] Since the functional relevant motions are highly collective, three normal modes corresponding to the lowest frequencies were used to generate the perturbed structures. Specifically, the original structure (gray colored in Fig. 1(b)) was perturbed along the three normal modes (accounting to approximately 20.8% of total fluctuations), with deformation amplitudes ranging from 1 to 100, and an ensemble of 1000 structures was generated around the original structure, with a maximum *RMSD* of approximately 10 Å compared to the original structure. The *RMSD* values of generated structures with respect to the original structure were used to group the structures into 10 bins using 1 Å as the bin size. The structure ensembles were then compiled by drawing structures from the 10 bins using the following approach: the first group includes structures randomly drawn from the first bin (*RMSD* $<$ 1 Å compared to the original structure). The second group is composed of structures randomly drawn from the first two bins, *etc*. This procedure ensures that the structure diversity levels are different in 10 groups. The structure preparation procedure is summarized in Fig. 2(b). The average structure deviation from the original structure and the mean square of the pairwise *RMSD* within each group, which is used to measure the structure diversity, are summarized in Table 1. This allows us to examine the dependence of the model resolution on the structure heterogeneity (or thermal fluctuation levels).

Table 1. The characteristics of the structure ensembles after incorporating the thermal fluctuations. In each group, there were 100 structures selected using the protocol described in the main text. In the table, $\langle RMSD\rangle^{a}$: Average *RMSD* with respect to the original model. $\langle RMSD^2\rangle^{b}$: The mean square of the pairwise *RMSD* within the ensemble.

Group ID	$\langle RMSD\rangle^{a}$/Å	$\langle RMSD^2\rangle^{b}$/Å^2
1	0.79	0.90
2	1.24	1.65
3	1.62	3.16
4	2.08	5.86
5	2.52	10.78
6	3.02	17.35
7	3.54	23.03
8	4.02	27.88
9	4.53	32.51
10	4.99	35.80

The projection image simulation procedure is essentially the same as previously described for the case of a single 3D structure, except that images forming a dataset were simulated based on randomly selected 3D structures from the corresponding group. Consequently, each simulated dataset has the following controlled parameters: the number of projections, SNR, and the structure heterogeneity due to thermal fluctuations.

2.3. Resolution cutoff and curve fitting

The resolution determination was based on the Fourier shell correlation (FSC) implemented in SPIDER with the gold standard rule at the cutoff level of FSC = 0.143. The simulation data was split into two half subsets randomly, and each reconstruction was carried out using back projections with known orientations or using the iterative reconstruction methods based on Bayesian maximum likelihood algorithm implemented in the Relion 1.4 package.[4] In the cases that the standard reconstruction procedures were carried out to build the electron density maps, each dataset was processed five times and the best resolutions were used in the final analysis.

The resolutions of the reconstructed maps were determined for various combinations of factors that were considered in this work. The value of each factor was systematically scanned in practical ranges so that the quantitative relationships could be studied using a parameter fitting to theoretical formula. The free parameters used in the GN and TF models were obtained by the nonlinear Least Squares (Curve Fitting) module in MATLAB.

2.4. The survey of resolutions of experimental models

A statistical survey was carried out on the structures determined using CryoEM deposited in the EMDB database (http://www.ebi.ac.uk/pdbe/). The retrieved information includes the molecular weight, detector type, resolution determination criteria, molecular symmetry, number of projections and the microscope operating voltages.

We focused on the models that fulfill the following requirements: (I) Models that were deposited between 2017/01/01 and 2018/05/17; (II) Micrographs were recorded using direct electron detection technology; (III) The reported resolution is determined with gold standard rule at FSC = 0.143; (IV) Models without higher symmetry; (V) Molecule weight of all models are ranged from 0.5 MDa to 1.0 MDa. As a result, 86 models from EMDB were selected to for the map resolution statistics.

3. Results

The RH model describes the dependency of the resolution on the number of projections and the SNR, molecular symmetry, and other factors under the umbrella of the *b*-factor. We focused on the study of three parameters by simplifying the formula to the GN and TF models, as described in Section 2. By varying the parameters that quantify these determinant factors, the influence on the resolution of the reconstructed structures

were systematically assessed. The simulated GroEL datasets were used to estimate the free parameters by optimizing the fitting of the analytical formula to the data points obtained from the simulations.

3.1. Effective number of projections

The RH model describes the dependency of map resolution on the number of experimental images, and the logarithm trend of the reconstructed model resolution as a function of the number of measured projections is attributed to the *b*-factors (due to sample particle drifting, misalignment, numerical interpolation, *etc*.[26] Surprisingly, the logarithm trend was observed for the simulation data without explicitly applying the envelope function $e^{Bk^2/2}$ during the projection simulations (see Fig. 3(a)). Furthermore, the utilization of the known orientation as available information eliminated alignment errors that occurred during orientation recovery process. Therefore, the errors described with the *b*-factor envelope function in RH model are not the only cause of the logarithm trend. The numerical interpolation during structure reconstruction could not be avoided, and this might be one source of the logarithm trend. Another possibility is that the logarithm trend is due to uneven information embedding in single particle imaging datasets, where the low-resolution information is redundant, while the high-resolution information barely reaches the signal-to-noise threshold.

In the 2D cases, n measurements of the same image will boost the SNR (defined as ratio of signal variance and noise variance) n times if the noise types and levels are the same for all measurements.[30] In the 3D map reconstruction from 2D projection images, we observed the same relationship (see Fig. 3). To simplify the GN model, a new parameter, the 'effective number of projections', N_e, was defined as the product of the projection number and the average SNR of the individual projection image (see Eq. (4)). The noise term is absorbed into this new parameter N_e, and equation (2) becomes:

$$N_e(k) = A_{GN} k e^{Bk^2}, \quad N_e(k) = \frac{N(k)}{\sigma^2}. \tag{4}$$

This relation can be verified with the simulation results summarized in Fig. 3.

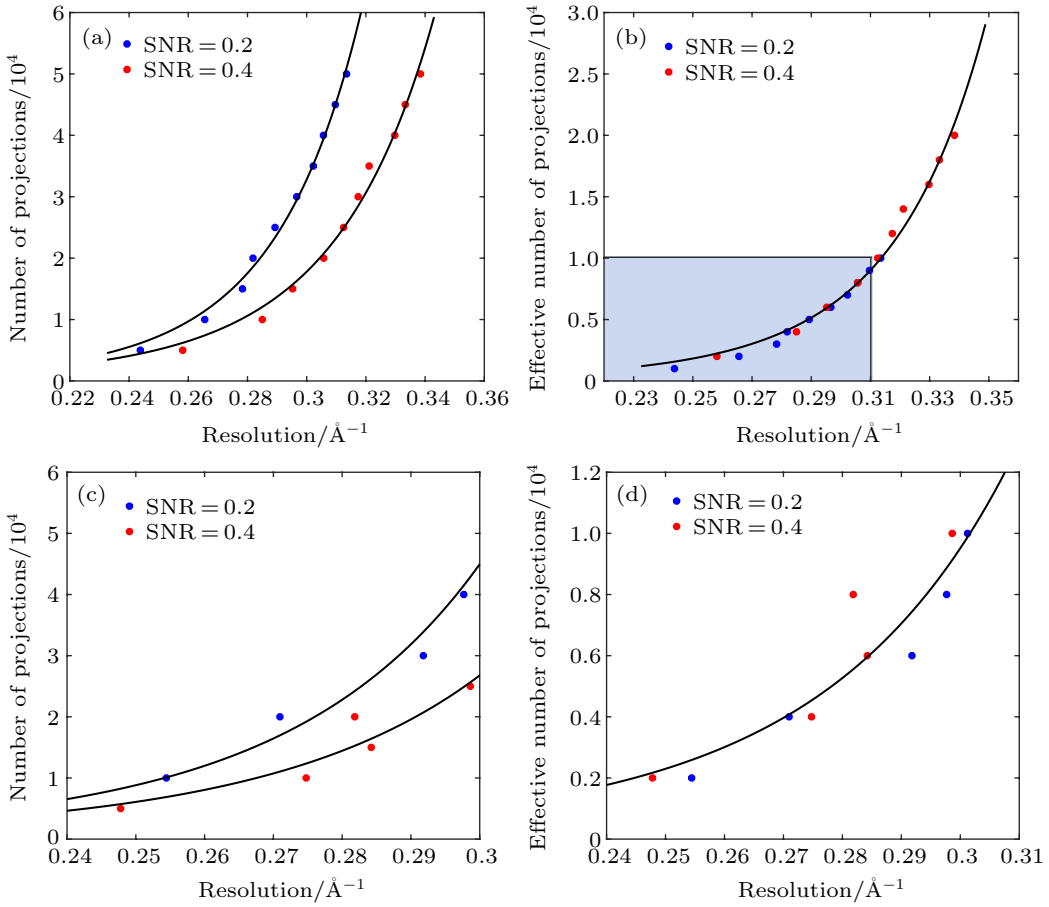


Fig. 3. (color online) Relation between model resolution and data quantity/quality. (a) The model resolution depends on the number of measurements. Two sets of data are plotted with signal to noise ratios (SNR) of 0.2 and 0.4. The lines were obtained by fitting the data points using the Gaussian noise model. *R*-square values were 0.9946 and 0.9865, respectively. (b) The noise term was combined with the number of projections. The two sets of data in Fig. 3(a) were fitted with a single set of parameters by defining the effective number of projections. *R*-square of the line is 0.9782. (c) and (d) The same plots as in panels (a) and (b), except that the maps were reconstructed using standard procedures in Relion without using known orientation information. Due to the heavy computing demands for standard reconstructions, a subset of data were investigated in the case of panels (c) and (d), and the sampled region in panel (d) corresponds to the blue shaded area in panel (b).

In Fig. 3(a), the SNR values were treated as a separate parameter, independently from N (the number of measurements) and the two sets of data with different SNR values were fitted to two equations (the two black curves). In Fig. 3(b), the data points were merged to a single line using the effective number of projections (N_e). Subsequently, a single equation is adequate to describe the relationship between the resolution and the number of 'effective' projections. Using the nonlinear curve fitting algorithm, the values of coefficient A_{GN} and B are 1298 and 46 (Å^2), respectively, for the GroEL simulation data.

The map reconstructions using back projection method are only possible if the image orientations were known, so it describes an ideal situation. In practice, the orientations needs to be assigned using iterative methods, which converges to the model that best satisfies the constraints of the whole dataset. Here, the auto-refine function implemented in Relion was applied for the model reconstruction. The gold standard rule was also used for resolution determination. To reduce the influence from randomness, the best resolution from five independent runs were reported as the map resolution for final statistics. The results are summarized in Figs. 3(c) and 3(d). The orientation assignment error indeed affects the quality of the reconstructed maps, indicated by the lower resolutions compared to that in the cases from back projection reconstructions. Nevertheless, after combining the number of projection images and the SNR, the overall trend is still reasonably described using the RH model, suggesting that the effective number of particles is valid in the simulated cases. However, the relation becomes invalid if the SNR is too low for accurate orientation assignment in extreme cases, which were not considered in this study.

3.2. Thermal fluctuation effect

Biological macromolecules exist in a thermal environment, and the structure fluctuates around the native states. In many cases, due to the functionality, molecules exist in several meta-stable conformations.[38–40] Using the normal mode perturbation approach, we attempted to simulate intrinsic motions and study their influence on the achievable structure resolutions at various conformation heterogeneity levels. The TF model described by Eq. (3) captures the relationship between map resolution and the number of projections in the presence of thermal fluctuations (*i.e.*, structure heterogeneity), using an analytical formula similar to the case of the temperature factor in x-ray crystallography.[41] Compared to the Gaussian noise model, the TF model predicts a different behavior in the resolution dependency on the number of measurements: the resolution gets worse faster for structure ensemble with larger thermal fluctuations.

In practice, both experimental noise and molecular thermal fluctuation have impacts on the CryoEM single particle experimental data. Therefore, it is necessary to derive a model that combines the GN and TF models. Intuitively, the following formulation is devised by treating the Gaussian noise and thermal fluctuation as independent factors that affect the structure resolutions:

$$N_e(k) = A_{TF} k\, e^{\left(Bk^2 + C\langle RMSD^2\rangle k^2\right)}, \quad N_e(k) = \frac{N(k)}{\sigma^2}. \tag{5}$$

Note that the effective number of particles is used in Eq. (5). To get a sensible fitting to the simulation data, we divided the TF model to two regimes using $\langle RMSD^2\rangle = 10$ Å^2 as a threshold. As a result, two sets of parameters were obtained: for $\langle RMSD^2\rangle > 10$ Å^2, $A_{TF} = 1702.0$, $B_{TF} = 58.33$ Å^2, and $C_{TF} = 0.75$; for $\langle RMSD^2\rangle \leqslant 10$ Å^2, $A_{TF} = 2465$, $B_{TF} = 41.06$ Å^2, and $C_{TF} = 2.88$. Without the thermal fluctuation term $C\langle RMSD^2\rangle$ in Eq. (5), we will need five sets of parameters to fit five datasets in Fig. 4.

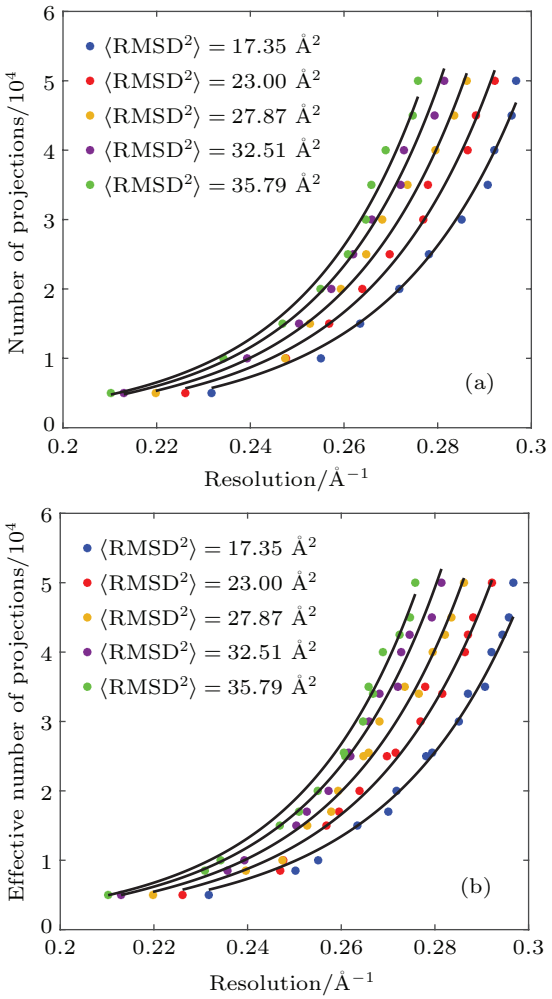


Fig. 4. (color online) The influence of thermal fluctuation to the structure resolution. (a) The relationship between the resolution and number of projections at various thermal fluctuation levels. The thermal fluctuation is quantified using the average square root-mean-square-derivation $\langle RMSD^2\rangle$ of the structure ensemble. All the projections were simulated with Gaussian noise with an SNR of 1.00. The R-square values of the fittings to the GN model are approximately 0.96. (b) The five datasets were fitted using a single set of parameters with an R-square of 0.98 using Eq. (5).

3.3. Statistics from EMDB database

To compare our theory and simulation results with experimental data and cross-validate the conclusions, we conducted a systematic survey on the structures determined using CryoEM single particle imaging technology, which are deposited in the EMDB database. It is worth noting that only the subset of structures that met the criteria described in the Methods section were used for the statistics. The resolution distribution nicely resembles the relationship described using the noise model and the thermal fluctuation model. Using these parameters (A_{GN}, A_{TF}, B, C in Eqs. (4) and (5)), we can draw theoretical guidelines to estimate the required number of projections and obtain the desired resolutions; see Fig. 5. Because the original model 1XCK weighed approximately 0.81 MDa, the data for molecule weight ranged from 0.5 MDa to 1 MDa were compared with the analytical curves of the two models.

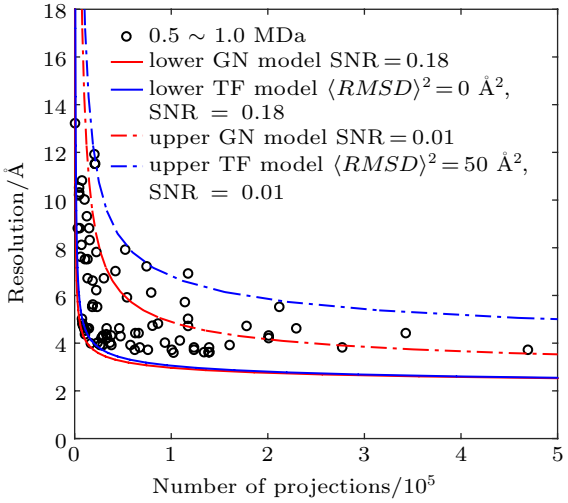


Fig. 5. (color online) The distribution of model resolutions as a function of particle numbers for experimentally determined structures. The theoretical lines indicate the resolution limits for molecules with molecular weights between 0.5 MDa and 1.0 MDa. The curves were drawn with SNR = 0.01 and 0.18 for upper and lower bounds predicted using GN model. The upper bound were predicted using TF model with $\langle RMSD^2\rangle = 50$ Å^2 and SNR = 0.01, the corresponding lower bound were predicted with $\langle RMSD^2\rangle = 0$ Å^2 and SNR = 0.18.

As shown in Fig. 5, the experimental model resolutions were bounded by the theoretical curves. In the bounds predicted by GN and TF models, the parameters obtained from GroEL data fitting were used. For GN model, the SNR levels were estimated to be 0.01 and 0.18 for upper and lower bounds. For the TF model, the same SNR levels were used, except that the upper bound has a thermal fluctuation term with $\langle RMSD^2\rangle = 50$ Å^2. The lower bounds for both GN and TF models are very similar; however, it is clear that the TF model provide a better estimate of the upper bound of resolution for experimental data. Another interesting observation is that the resolution hardly broke the barrier of 3 Å, although the number of projections covers a broad range from 2×10^4 to 2×10^5. This is likely due to the thermal fluctuations or the coexistence of multiple conformations of the molecule.

4. Discussion and conclusion

Based on theoretical frameworks, we systematically investigated the map resolutions and factors using numerical simulation methods, including the number of measured projections, signal-to-noise ratio, and the heterogeneity of the molecules. Two theoretical frameworks were proposed to describe the relationship between these factors and the resolutions of reconstructed maps. The Gaussian noise model is essentially the same as the formula proposed by Rosenthal and Henderson, and we found that the noise term could be combined with the number of projections by defining an 'effective number of projections'. In the thermal fluctuation model, the intrinsic dynamic characters of the sample were considered, and the final resolution could be affected by the fluctuation level of molecules. The noise and thermal fluctuation frameworks can be used to provide guidelines to estimate the required number of projections to reach the desired resolutions, which was validated using the statistics from the structures that were experimentally resolved using CryoEM single particle imaging method.

Simulation studies were carried out in a controlled manner so that the influence of individual factor can be decoupled from that of other factors. We noticed that the uniform orientation sampling is an ideal situation, because orientation bias often exists in experimental datasets. In extreme cases, the missing cone problem can result in strong artifacts in the reconstructed models. Therefore, the FSC criteria measured the model consistency and not necessarily the correctness. Because of the scope of this study, we used the uniform distribution of orientations and the back-projection method to ensure that the model reconstructions were carried out properly. These operations are useful to secure the validity of the FSC criteria in the resolution cutoff. Nonetheless, the correctness of the model should be checked using complementary methods, such as visual inspection of the density maps, local resolution estimation, or validation using biochemistry assays.

In this work, the noises were simulated from a Gaussian distribution to study their influence on the model resolutions. This largely simplifies the noise sources, where the major sources include background scattering from vitreous ice, sample drifting during measurement, misalignment in the orientations and errors introduced in the orientation discretization. Some of these errors could be mimicked in the simulation framework, such as using an envelope function with b-factors. However, this is beyond our focus in this study, although it may be a subject for future research.

Despite the simplicity of the formulation, the proposed models can be useful in structure determination with CryoEM single particle imaging methods. One application is to design a data collection strategy to reach the desired resolutions. SNR can be estimated based on sample screening data, and then equations (4) and (5) can be used to estimate the number of required measurements. Although the current theoretical formulation is based on uniform distributed orientation, and the parameters may need to be refined for other molecules (because the parameters obtained in this work are for GroEL molecules), the results presented in this paper can provide hints about the obtainable structure resolution for a given number of measured projections. The data survey on the experimental structures provides evidence that the Gaussian noise and thermal fluctuation models are applicable for estimating the resolution limits in molecular structure determination.

The thermal fluctuation model can be used to assess the structure heterogeneity by reconstructing structures with subsets of data under the homogenous conformation assumption. With the obtained structure, the same simulation study can be carried out to generate a series of curves that are associated with different structural heterogeneity (see Fig. 4). Then, the resolution of the experimentally determined structure as a function of the projection number can be compared with the curves to infer the level of structure heterogeneity. If the $\langle RMSD^2 \rangle$ values are large, then multiple structures should be reconstructed using the 3D classification approach or a similar method.[38,42,43] For instance, as shown in Fig. 5, there are six data points near the upper bound of TF model. After a close examination, we found that five data points (EMD-6739, EMD-6740, EMD-6835, EMD-6836, and EMD-6837) are the same molecular complex (di-nucleosome with linker DNA and HP1 protein). This suggests that significant conformation heterogeneity still exist in each dataset that was used for map reconstruction.

It should be noted that the conclusion learned from the numerical simulations is conditional thanks to the simplified the theoretical framework and the idealized treatment of simulation data. The effective number of projections, N_e, is valid under the assumption that the SNR is high enough to allow most of projections to be assigned to their correct orientations. Otherwise, the N_e will not be a simple product of SNR and the number of projections. One would expect a much smaller N_e than SNR*N_p, if the low SNR leads to significant alignment errors. Including more projections images should improve the model resolution in practice, so it is not suggested to exclude a particular set of images that are at high defocus levels, because they can help improve the orientation assignment, especially when dataset is not extremely large.

In summary, the resolution limiting factors in the CryoEM single particle imaging method were investigated under theoretical frameworks with a numerical simulation approach. The results suggest that the resolution of the reconstructed structure strongly depends on the number of measurements, image quality, and molecular flexibility. Our results can provide a guidance to design appropriate experimental strategies in data collection in general and model reconstruction in the case of molecules with heterogeneous conformations.

Acknowledgment

We thank Mao Y D, Wu J Y, and Zhu Y N for helpful discussions. The computation was performed in part using the high-performance computational platform at the Peking–Tsinghua Center for Life Science at Peking University, Beijing, China.

References

[1] Bai X C, McMullan G and Scheres. S H W 2015 *Trends BioChem. Sci.* **40** 49
[2] Faruqi A R and McMullan G 2011 *Quarterly Rev. Biophys.* **44** 357
[3] Tang G, Peng L, Baldwin P R, Mann D S, Jiang W, Rees I and Ludtke S J 2007 *J. Struct. Biol.* **157** 38
[4] Scheres S H 2012 *J. Struct. Biol.* **180** 519
[5] Grigorieff N 2016 *Methods Enzymol.* **579** 191
[6] da Fonseca P C A and Morris E P 2015 *Nat. Commun.* **6** 7573
[7] Passmore L A and Russo C J 2016 *Methods in Enzymology* **579** 51
[8] Bernecky C, Herzog F, Baumeister W, Plitzko J M and Cramer P 2016 *Nature* **529** 551
[9] Li X, Mooney P, Zheng S, Booth C R, Braunfeld M B, Gubbens S, Agard D A and Cheng Y 2013 *Nat. Methods* **10** 584
[10] Zheng S Q, Palovcak E, Armache J P, Verba K A, Cheng Y and Agard D A 2017 *Nat. Methods* **14** 331
[11] Yu X, Jin L and Zhou Z H 2008 *Nature* **453** 415
[12] Chen J Z, Settembre E C, Aoki S T, Zhang X, Bellamy A R, Dormitzer P R, Harrison S C and Grigorieff N 2009 *Proc. Natl. Acad. Sci. USA* **106** 10644
[13] Zhang X, Jin L, Fang Q, Hui W H and Zhou Z H 2010 *Cell* **141** 472
[14] Cao E, Liao M, Cheng Y and Julius D 2013 *Nature* **504** 113
[15] Bernstein F C, Koetzle T F, Williams G J, Meyer E F Jr, Brice M D, Rodgers J R, Kennard O, Shimanouchi T and Tasumi M 1977 *J. Mol. Biol.* **112** 535
[16] Berman H M, Westbrook J, Feng Z, Gillil, G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Res.* **28** 235
[17] Newman R, Chagoyen M, Carazo J M and Henrick K 2002 *Trends BioChem. Sci.* **27** 11
[18] Henrick K, Newman R, Tagari M and Chagoyen M 2003 *J. Struct. Biol.* **144** 228
[19] Rose P W, Prlic A, Bi C, Bluhm W F, Christie C H, Dutta S, Green R K, Goodsell D S, Westbrook J D, Woo J, Young J, Zardecki C, Berman H M, Bourne P E and Burley S K 2015 *Nucleic Acids Res.* **43** D345
[20] Nogales E 2016 *Nat. Methods* **13** 24
[21] Lawson C L, Patwardhan A, Baker M L, Hryc C, Garcia E S, Hudson B P, Lagerstedt I, Ludtke S J, Pintilie G, Sala R, Westbrook J D, Berman H M, Kleywegt G J and Chiu W 2016 *Nucleic Acids Res.* **44** D396
[22] Morris A L, MacArthur M W, Hutchinson E G and Thornton J M 1992 *Proteins: Structure, Function, and Genetics* **12** 345
[23] Karplus P A and Diederichs K 2012 *Science* **336** 1030
[24] Heel M v and Harauz G 1986 *Optik* **73** 146
[25] Böttcher B, Wynne S A and Crowther R A 1997 *Nature* **386** 88
[26] Rosenthal P B and Henderson R 2003 *J. Mol. Biol.* **333** 721
[27] Scheres S H and Chen S 2012 *Nat. Methods* **9** 853
[28] Kucukelbir A, Sigworth F J and Tagare H D 2014 *Nat. Methods* **11** 63
[29] Henderson R 1995 *Quarterly Rev. BioPhys.* **28** 171
[30] Penczek P A 2010 *Methods Enzymol.* **482** 73

[31] Liao H Y and Frank J 2010 *Structure* **18** 768
[32] Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X W, Milne J L S and Subramaniam S 2015 *Science* **348** 1147
[33] Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis M I, Pragani R, Boxer M B, Earl L A, Milne J L and Subramaniam S 2016 *Cell* **165** 1698
[34] Bartolucci C, Lamba D, Grazulis S, Manakova E and Heumann H 2005 *J. Mol. Biol.* **354** 940
[35] Shaikh T R, Gao H, Baxter W T, Asturias F J, Boisset N, Leith A and Frank J 2008 *Nat. Protocols* **3** 1941
[36] Yershova A, Jain S, Lavalle S M and Mitchell J C 2010 *Int. J. Rob. Res.* **29** 801
[37] Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon T R and Bahar I 2014 *Bioinformatics* **30** 2681
[38] Dashti A, Schwander P, Langlois R, Fung R, Li W, Hosseinizadeh A, Liao H Y, Pallesen J, Sharma G, Stupina V A, Simon A E, Dinman J D, Frank J and Ourmazd A 2014 *Proc. Natl. Acad. Sci. USA* **111** 17492
[39] Chen S, Wu J, Lu Y, Ma Y B, Lee B H, Yu Z, Ouyang Q, Finley D J, Kirschner M W and Mao Y 2016 *Proc. Natl. Acad. Sci. USA* **113** 12991
[40] Dashti A, Ben Hail D, Mashayekhi G, Schwander P, des Georges A, Frank J and Ourmazd A 2017 *BioRxiv* 167080
[41] Trueblood K N, Bürgi H B, Burzlaff H, Dunitz J D, Gramaccioli C M, Schulz H H, Shmueli U and Abrahams S C 1996 *Acta Crystallographica* **A52** 770
[42] Frank J and Ourmazd A 2016 *Methods* **100** 61
[43] Hosseinizadeh A, Mashayekhi G, Copperman J, Schwander P, Dashti A, Sepehr R, Fung R, Schmidt M, Yoon C H, Hogue B G, Williams G J, Aquila A and Ourmazd A 2017 *Nat. Methods* **14** 877