Chin. Phys. Lett. **39**, 070501 (2022)

# Cryo-EM Data Statistics and Theoretical Analysis of KaiC Hexamer

Xu Han(韩旭)[1], Zhaolong Wu(吴赵龙)[1], Tian Yang(杨添)[1], and Qi Ouyang(欧阳颀)[1,2*]

[1]*Department of Physics, Peking University, Beijing 100871, China*
[2]*Center for Quantitative Biology and Peking-Tsinghua Center for Life Sciences, AAIC, Peking University, Beijing 100871, China*

(Received 11 May 2022; accepted manuscript online 7 June 2022)

Cryo-electron microscopy (cryo-EM) provides a powerful tool to resolve the structure of biological macromolecules in natural state. One advantage of cryo-EM technology is that different conformation states of a protein complex structure can be simultaneously built, and the distribution of different states can be measured. This provides a tool to push cryo-EM technology beyond just to resolve protein structures, but to obtain the thermodynamic properties of protein machines. Here, we used a deep manifold learning framework to get the conformational landscape of KaiC proteins, and further obtained the thermodynamic properties of this central oscillator component in the circadian clock by means of statistical physics.

DOI: 10.1088/0256-307X/39/7/070501

The robust circadian clock of the cyanobacterium, which is characterized by about 24 h rhythm in the phosphorylation level of KaiC, can be reconstituted *in vitro* by incubating KaiA, KaiB, KaiC and adenosine triphosphate (ATP).[1] Therefore, this simple system has become a powerful model to study circadian rhythms and timekeeping mechanisms.

Each KaiC hexamer has two domains (N-terminal CI domain and C-terminal CII domain), where CII domain possesses kinase and phosphotransferase activities.[2–4] In the CII domain, each KaiC subunit has two phosphorylation sites (S431 and T432). During the subjective day, KaiA combines with KaiC C-terminal tail within CII domain, thus results in T432 to capture a phosphate before S431.[2,5–9] When it comes to the objective night, KaiB sequesters KaiA and leads to dephosphorylation of KaiC also first at T432 and then at S431,[10–15] see Fig. 1 schematically for the molecular changes in the KaiCBA oscillator.

Previous studies proposed that the flexibility of CII ring and the dynamic equilibrium position of the A-loop are determinants of KaiC phosphorylation activity.[7] Many researchers demonstrated the dynamic structural properties within KaiC-CII domain. NMR methyl transverse relaxation optimized spectroscopy indicated that across a phosphorylation cycle, CII ring is most flexible at S/T state and most rigid at PS/T state, while the rigidity of CII ring enhances CII–CI interactions.[16] Time-resolved small-angle x-ray scattering and fluorescence studies showed various $R_g$ values (radius of gyration, a measure of the overall shape of the KaiC hexamer) among different phosphorylated states of KaiC.[17] The molecular dynamics simulation work also suggested that the nucleotide release in CII domain involves obvious conformational changes.[18] A recent NMR study revealed that ATP hydrolysis is coupled with conformational changes in the flexible KaiC C-terminal segments.[19]

In our previous study,[20] KaiC proteins were incubated at 30 ℃ for six hours and equilibrated to a functionally relevant state. A 3.5-μl drop of 0.35 μg/μl KaiC-AA (or KaiC-EE) solution was applied to the glow-discharged grids in an environmentally controlled chamber (FEI Vitrobot Mark Ⅳ) with 100% humidity and 4 ℃ temperature. After 1 blot force for 1 s blot time, the grid was plunged into liquid ethane and then was transferred to liquid nitrogen. The cryo-EM data was collected on a 300 kV Titan Krios G2 microscope, semi-automatically with SerialEM software[21] in a super-resolution counting mode. A total of 5125 movies of KaiC-AA and 3530 movies of KaiC-EE were collected. After a series of processing by RELION,[22,23] we reconstructed the cryo-EM density maps of KaiC-AA and KaiC-EE using raw single particle images prior to conformational classification. The overall resolutions of these two reconstructions are both 3.3 Å,[20] measured by gold-standard FSC.[24,25] Furthermore, based on these two high-resolution structures, we did conformational classification and found that the status of the A-loop has two distinct possible conformations (exposed and buried) for each KaiC subunit in a KaiC hexamer.[20]

*Corresponding author. Email: qi@pku.edu.cn
© 2022 Chinese Physical Society and IOP Publishing Ltd

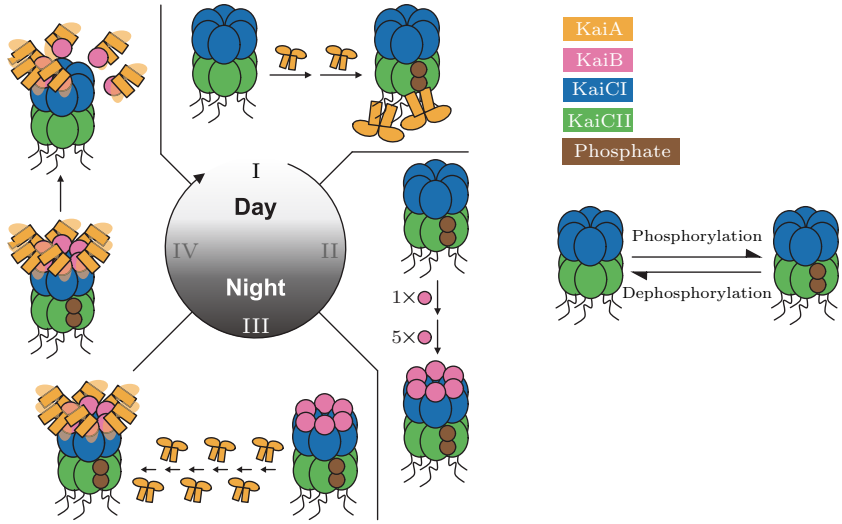


**Fig. 1.** Schematic diagram of the periodic assembly in the cyanobacterial circadian clock.[15]

In this Letter, we employed the recently developed AlphaCryo4D method[26] to analyze conformational states around the A-loop region. After performing the six-fold pseudo-symmetry operation, we used a bootstrapping approach by replicating each particle 3 times and dividing the resulting (3×) dataset randomly into 4 groups. Next, we used the A-loop region structure and the corresponding extracted feature of all the particles for clustering using RELION.[22,23] For each group, we took the structure of the A-loop region of all the particles and used them to train a three-dimensional (3D) residue neural network for feature extraction. Each resulting cluster was characterized by its 3D volume that was the averaged structure over all the particles belonging to the cluster. For classification, the 3D volume data for all the clusters obtained from all 4 groups of particles were then embedded in a low dimensional space based on the t-distributed stochastic neighbor embedding (t-SNE) method.[27,28]

Briefly, given a data set, t-SNE assumes that the joint probability $p_{ij}$ of original high dimensional data follows Gaussian distribution and a student t-distribution is employed to model the low dimensional distribution $q_{ij}$. After that, the Kullback–Leibler (KL) divergence[29] is used as an objective function to measure the distance between similarity distributions of $p_{ij}$ and $q_{ij}$, which is minimized by the gradient descent method to find the low dimensional latent variable of each data point. It is worth noting that the student t-distribution is a heavy-tailed distribution, which causes the closer (or farther) points in high dimensional space would closer (or farther) when mapping onto low dimensional space. Embedding into low dimensional space in this way can make similar points closer while dissimilar points farther away, thus alleviating the crowding problem. In high dimensional space, we manually identified the conformational classification results (focusing on the A-loop region), which can be roughly divided into two conformational states (exposed and buried). Therefore, we used two-dimensional manifold embedding for classification. By virtue of the bootstrapping method used in AlphaCryo4D,[26] each particle got three votes from clustering results based on different subsampling of the whole dataset, which allowed us to determine the conformational state for most of the particles with a higher statistical confidence.

From the conformational landscape based on the A-loop area for KaiC-AA in Fig. 2(a) and KaiC-EE in Fig. 2(c), there is a well-defined cluster of the buried state. From the center of the buried state cluster, we could draw circles of increasing radius and define the boundary of the buried state cluster as the circle on which the fraction of the exposed states reaches 20% threshold (the boundary does not depend sensitively on the threshold). In principle, all 3D volumes outside of this buried-state boundary could be candidates for the exposed-state. Here, we took the point that is farthest away from the center of the buried-state cluster as the center of the exposed-state cluster and defined the boundary of the exposed-state cluster as the circle tangent to the buried-state circle. Particles within the buried and exposed circles were voted as buried (Bu) or exposed (Ex) state, respectively; while particles outside of both circles were voted as undefined (Un). See Fig. 2(b) of KaiC-AA and Fig. 2(d) of KaiC-EE for the typical conformational states in the buried cluster, exposed cluster and on the buried boundary, respectively.

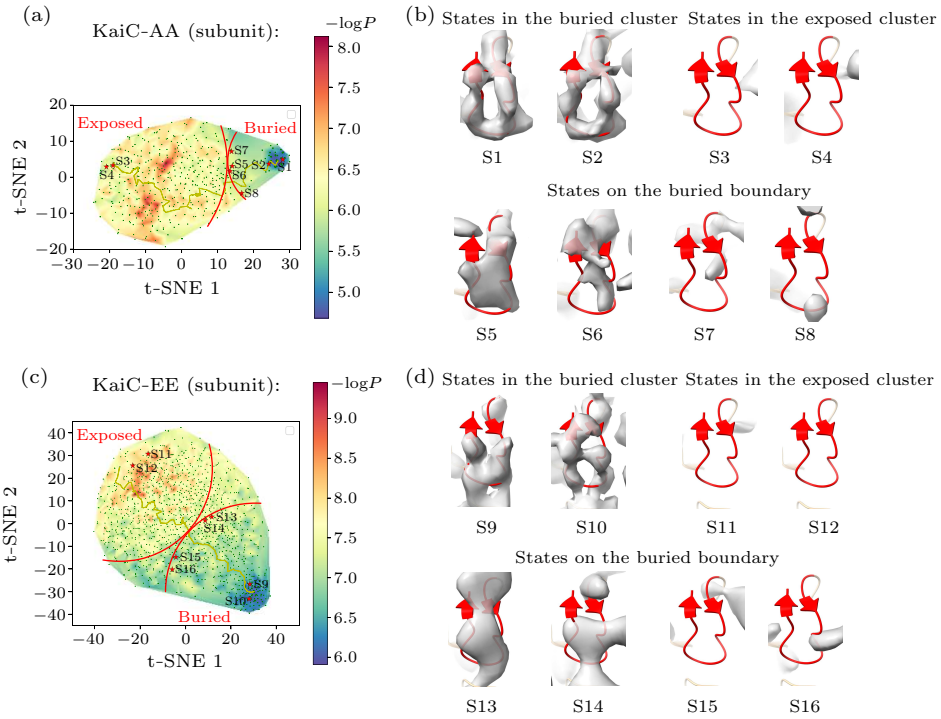


**Fig. 2.** Division of buried state and exposed state on the conformational landscape. Conformational landscape (a) of KaiC-AA and (c) of KaiC-EE. Typical states in the buried cluster, exposed cluster and on the buried boundary (b) of KaiC-AA and (d) of KaiC-EE. Cryo-EM densities focused on the A-loop area are rendered as transparent surfaces colored in gray, superimposed with the cartoon representation of the atomic model of the A-loop (residue 488–497, colored in red).

**Table 1.** The error probability ($\ll 1$) based on its three identifications, and the voting error for exposed (or buried) state that was evaluated from KaiC-AA, KaiC-EE and mixed sample data set, respectively.

| Exposed(Ex) | | | Buried(Bu) | | | Voting error | | |
|---|---|---|---|---|---|---|---|---|
| Votes | Number | Error probability | Votes | Number | Error probability | | $\varepsilon_{\rm Ex}$ | $\varepsilon_{\rm Bu}$ |
| 3Ex | $N_{\rm 3Ex}$ | 0 | 3Bu | $N_{\rm 3Bu}$ | 0 | KaiC-AA | 0.14 | 0.18 |
| 2Ex 1Bu | $N_{\rm 2Ex,1Bu}$ | $\varepsilon$ | 2Bu 1Ex | $N_{\rm 2Bu,1Ex}$ | $\varepsilon$ | KaiC-EE | 0.24 | 0.18 |
| 2Ex 1Un | $N_{\rm 2Ex,1Un}$ | $\varepsilon/2$ | 2Bu 1Un | $N_{\rm 2Bu,1Un}$ | $\varepsilon/2$ | Mixed | 0.23 | 0.21 |

For each subunit with a given assignment [exposed (Ex) or buried (Bu)], we could define an error probability ($\ll 1$) based on its three identifications (Table 1), where $\varepsilon$ has a finite but small value, it measures the error probability for voting '2Ex 1Bu' as an exposed (Ex) state. The voting error for exposed (or buried) state can be evaluated for a given data set (Table 1) with

$$\varepsilon_{\rm Ex} = \frac{N_{\rm 2Ex,1Bu} + N_{\rm 2Ex,1Un}}{3 \times (N_{\rm 3Ex} + N_{\rm 2Ex,1Bu} + N_{\rm 2Ex,1Un})},$$

$$\varepsilon_{\rm Bu} = \frac{N_{\rm 2Bu,1Ex} + N_{\rm 2Bu,1Un}}{3 \times (N_{\rm 3Bu} + N_{\rm 2Bu,1Ex} + N_{\rm 2Bu,1Un})},$$

and the larger one was chosen as the voting error $\varepsilon$ for all monomers for further calculation.

With the voting error $\varepsilon$, we used the Monte Carlo method to obtain the distribution and the error of the distribution. We repeated the simulation 100 times and obtain 100 distributions for a given data set, then the mean and the standard deviation of these distributions were calculated as the final results for each data set. Figures 3(a) and 3(b) show the case with $\varepsilon = 0$ and $\varepsilon$ with larger one between $\varepsilon_{\rm Ex}$ and $\varepsilon_{\rm Bu}$ ($\varepsilon = \max(\varepsilon_{\rm Ex}, \varepsilon_{\rm Bu})$) for KaiC-AA and KaiC-EE data sets, respectively.

Figure 3(c) demonstrates the best fitting of the experimental data with our previous model.[20] $R^2$ between model prediction and actual experimental data is 0.97 (consider both KaiC-AA and KaiC-EE together). The best fitting parameters are $J = 0.19 \pm 0.03$, $B_{\rm AA} = 0.24 \pm 0.03$, $B_{\rm EE} = -0.22 \pm 0.03$ (error bars were computed with $R^2 \geqslant 0.95$), roughly in accord with that in our previous study.[20]

We also considered an extended model with nonlocal interactions, i.e., the Ising model with interactions for all possible subunit pairs. In this case, Hamiltonian of each configuration can be expressed as

$$H(\boldsymbol{S}) = -J_1 \sum_{\langle ij \rangle} S_i S_j - J_2 \sum_{(ij)} S_i S_j - J_3 \sum_{\{ij\}} S_i S_j - B_{\rm AA(EE)} \sum_i S_i. \quad (1)$$

The symbols $\langle ij\rangle, (ij), \{ij\}$ represent that $i$ and $j$ are nearest, next-nearest and opposite subunits, respectively. The best fitting parameters are $J_1 = 0.18 \pm 0.02$, $J_2 = 0.03 \pm 0.02$, $J_3 = 0.02 \pm 0.01$, $B_{\rm AA} = 0.21 \pm 0.02$, $B_{\rm EE} = -0.19 \pm 0.02$ (error bars were computed with $R^2 \geqslant 0.95$), and $R^2$ is 0.99 (Fig. 4 considers both KaiC-AA and KaiC-EE together). Nevertheless, the estimated interaction strength between non-adjacent subunits is much smaller than $J$.

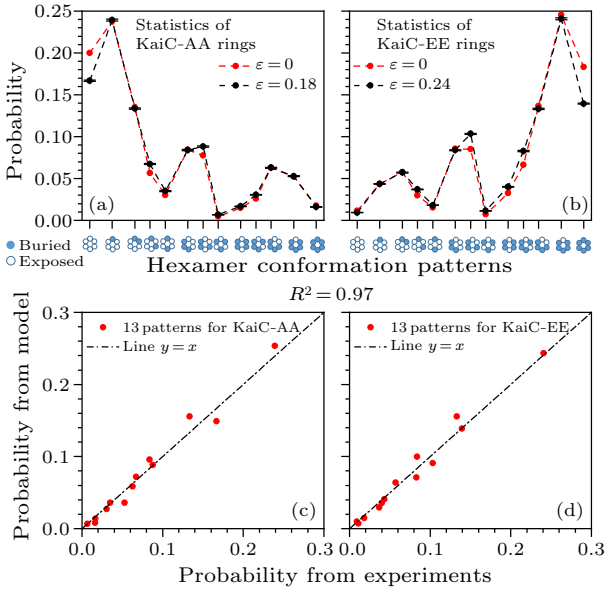


**Fig. 3.** Statistical results of pure KaiC-AA and KaiC-EE hexamers. Probabilities of different conformational patterns of (a) KaiC-AA data set (81243 rings) and (b) KaiC-EE data set (129488 rings) with error bar analysis. (c) Comparison between experiment and model for KaiC-AA (left) and KaiC-EE (right) data set, respectively.

In our previous study, we constructed a mixed sample.[20] Briefly, KaiC mutants (KaiC-AA, KaiC-EE, 1 : 1) were buffer exchanged into the running buffer with 0.5 mmol ADP, and incubated at 4 ℃ for about 24 hours to disrupt hexamer structures.[4] Then monomerized KaiC-AA and KaiC-EE were mixed before re-hexamerization via the addition of ATP.[30] Following the same data collection procedures, we collected cryo-EM data of this mixed sample with a 300 kV Titan Krios G2 microscope device. After a series of processing by RELION,[22,23] we obtained the cryo-EM density map with an overall resolution of 3.8 Å,[20] measured by gold-standard FSC.[24,25] By following the same procedure as the pure samples, we obtained the conformational landscape [Fig. 5(a)] and the probabilities of the 13 hexamer conformational patterns [Fig. 5(b)] for the mixed sample. As described in our previous work,[20] each hexamer can have 14 distinct KaiC-AA and KaiC-EE subunit arrangements, whose probability is denoted as $q_l$ ($l = 1, 2, \ldots, 14$). Considering the probability of each hexamer conformational pattern $p_l(k)$, the theoretical predicted overall distribution of the hexamer conformational patterns is $P_k = \sum_l q_l p_l(k)$.[20] $R^2$ between model prediction and actual experimental data is 0.88 [Fig. 5(c)].

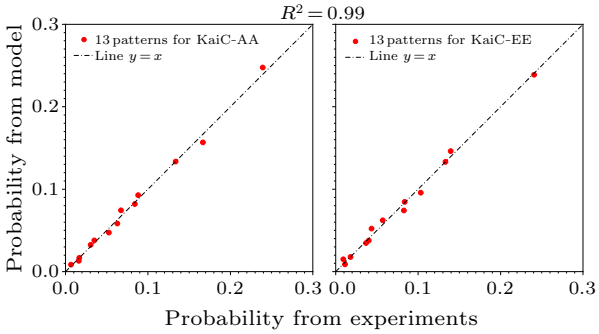


**Fig. 4.** Comparison between experiment and models with the assumption that nearest subunits, interval subunits and opposite subunits all interact with each other.

In our minimal model, we assumed the interaction strength between neighboring KaiC monomers is independent of their phosphorylation levels. However, this would not be the case. Specifically, we modified our original Ising model[20] by introducing the composition-dependent coupling constant $\Delta J$:

$$J_{ij} = J + \Delta J[(2\sigma_i - 1)(2\sigma_j - 1) - 1], \qquad (2)$$

where $\Delta J$ is the difference in coupling constant between the same type of monomers (EE-EE or AA-AA) and different types of monomers (AA-EE), $\sigma_i = 0$ or 1 if site-$i$ is AA or EE.

We fitted the theoretical predicted $P_k$ to experimental observation ($P_{k(\rm Ex)}$) by

$$\min_q \Bigg[\sum_{k=1}^{13}\Big(\sum_{l=1}^{14} q_l p_l(k) - P_{k(\rm Ex)}\Big)^2 + \lambda\Big(\sum_{l=1}^{14} q_l(1-E_l) - \sum_{l=1}^{14} q_l E_l\Big)^2\Bigg], \qquad (3)$$

subject to the constraints: $\sum_{l=1}^{14} q_l = 1$, $0 \leqslant q_l \leqslant 1$.[20] We chose $\lambda = 0.1$ to enforce the difference between KaiC-AA and KaiC-EE percentages to be less than 10%.

As shown in Fig. 5(d), this extended Ising model can fit the mixed hexamer data better for a range of $J' \neq J$ (or equivalently $\Delta J \neq 0$), which suggests that the subunit–subunit interaction strength could depend on their phosphorylation levels. Our analysis of the mixed sample indicates that the mixing of monomers with different phosphorylation levels may not be random, e.g., values of the optimized $q_l$ [Fig. 5(e), cyan] indicate that there is a higher probability for two neighboring subunits to have different phosphorylation levels.

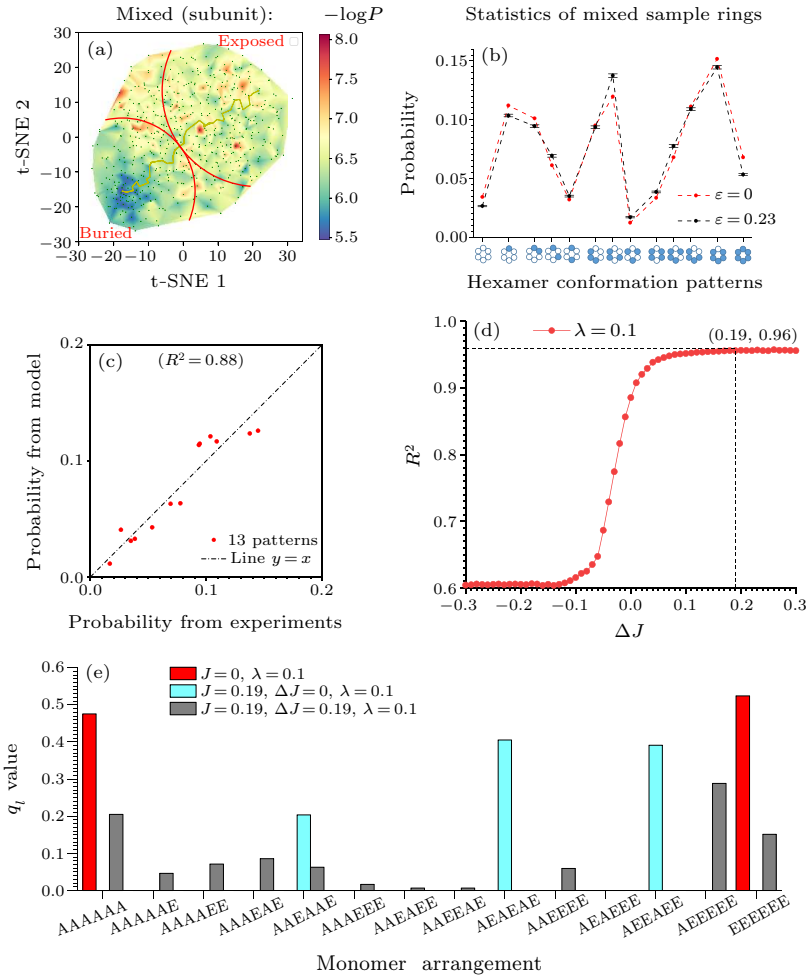


**Fig. 5.** Statistical results of the mixed sample. (a) Conformational landscape of the mixed sample data set. (b) Probabilities of different conformational patterns of the mixed sample data set (61014 rings). (c) Comparison between the mixed experimental data and model results with $q_l$ treated as free parameters and $\lambda = 0.1$, $J = 0.19$. (d) $R^2$ versus $\Delta J$ with the choice of $\lambda = 0.1$. (e) According to Eq. (3) with $\lambda = 0.1$, the $q_l$ values when $J = 0$ (red), $J = 0.19$ (cyan) and $J = 0.19$, $\Delta J = 0.19$ (gray).

Due to the low signal-to-noise ratio (SNR) of cryo-EM raw data, it is difficult to identify the conformational information of each raw single particle. Many methods have been developed to solve this problem. In this work, the AlphaCryo4D method allowed us to determine the conformational state for most of the particles with a higher statistical confidence by virtue of the bootstrapping method followed by particle voting step. These new designs in AlphaCryo4D can improve the robustness of classification results to overcome the overfitting problem to some extent. In addition, we also provide a method to measure the conformational classification errors with an error probability $\varepsilon$. The error bars estimate the extent of conformational classification/recognition errors, which seems to have a little impact on the final statistical results. Indeed, the accurate conformational classification/recognition in cryo-EM data processing is still an important open question for future research.

In summary, based on the conformational landscape of the KaiC subunit, we have characterized the distribution of thirteen KaiC hexamer conformational patterns. The results support the robust statistical analysis using pure KaiC-AA and KaiC-EE cryo-EM data in comparison with our previous study,[20] which provides a typical example of using cryo-EM to study the thermodynamic properties of a protein machine. Moreover, the analysis of the mixed sample suggests some preferential mixing phenomenon. The relevant possible underlying mechanism is worth further investigation given that KaiC hexamers constantly disassemble and reassemble during circadian oscillation through monomer shuffling,[6,31,32] which is thought to be involved in synchronization.[33–35]

*Acknowledgments.* We are grateful to Y. H. Tu, M. J. Rust, Y. D. Mao, D. L. Zhang, D. Q. Yu, L. Hong, Y. N. Zhu, B. Liu, S. W. Zhang, S. T. Zou, H. Liang, Y. H. Wang, K. Y. Wang, D. Y. Yin, and W. L. Wang for constructive discussions, Y. Ma, X. Li, and X. Pei for technical supports. The data pro-

cessing was performed on the Weiming No. 1 and Life Science No. 1 High Performance Computing Platform in Peking University. This work was supported by the National Natural Science Foundation of China (Grant No. 12090054).

## References


[1] Nakajima M, Imai K, Ito H, Nishiwaki T, Murayama Y, Iwasaki H, Oyama T and Kondo T 2005 *Science* **308** 414
[2] Iwasaki H, Nishiwaki T, Kitayama Y, Nakajima M and Kondo T 2002 *Proc. Natl. Acad. Sci. USA* **99** 15788
[3] Egli M, Mori T, Pattanayek R, Xu Y, Qin X and Johnson C H 2012 *Biochemistry* **51** 1547
[4] Nishiwaki T and Kondo T 2012 *J. Biol. Chem.* **287** 18030
[5] Vakonakis I and Liwang A C 2004 *Proc. Natl. Acad. Sci. USA* **101** 10925
[6] Kageyama H, Nishiwaki T, Nakajima M, Iwasaki H, Oyama T and Kondo T 2006 *Mol. Cell* **23** 161
[7] Kim Y I, Dong G, Carruthers C W, Golden Susan S and Liwang A 2008 *Proc. Natl. Acad. Sci. USA* **105** 12825
[8] Pattanayek R and Egli M 2015 *Biochemistry* **54** 4575
[9] Pattanayek R, Williams D R, Pattanayek S, Xu Y, Mori T, Johnson C H, Stewart P L and Egli M 2006 *EMBO J.* **25** 2017
[10] Brettschneider C, Rose R J, Hertel S, Axmann I M, Heck A J R and Kollmann M 2010 *Mol. Syst. Biol.* **6** 389
[11] Qin X, Byrne M, Mori T, Zou P, Williams D R, Mchaourab H and Johnson C H 2010 *Proc. Natl. Acad. Sci. USA* **107** 14805
[12] Phong C, Markson J S, Wilhoite C M and Rust M J 2013 *Proc. Natl. Acad. Sci. USA* **110** 1124
[13] Chang Y G, Cohen S E, Phong C, Myers W K, Kim Y I, Tseng R, Lin J, Zhang L, Boyd J S, Lee Y, Kang S, Lee D, Li S, Britt R D, Rust M J, Golden S S and Liwang A 2015 *Science* **349** 324
[14] Tseng R, Goularte N F, Chavan A, Luu J, Cohen S E, Chang Y G, Heisler J, Li S, Michael A K, Tripathi S, Golden S S, Liwang A and Partch C L 2017 *Science* **355** 1174
[15] Snijder J, Schuller J M, Wiegard A, Lossl P, Schmelling N, Axmann I M, Plitzko J M, Forster F and Heck A J 2017 *Science* **355** 1181
[16] Chang Y G, Kuo N W, Tseng R and Liwang A 2011 *Proc. Natl. Acad. Sci. USA* **108** 14431
[17] Murayama Y, Mukaiyama A, Imai K, Onoue Y, Tsunoda A, Nohara A, Ishida T, Maeda Y, Terauchi K, Kondo T and Akiyama S 2011 *EMBO J.* **30** 68
[18] Hong L, Vani B P, Thiede E H, Rust M J and Dinner A R 2018 *Proc. Natl. Acad. Sci. USA* **115** E11475
[19] Yunoki Y, Ishii K, Yagi-Utsumi M, Murakami R, Uchiyama S, Yagi H and Kato K 2019 *Life Sci. Alliance* **2** e201900368
[20] Han X, Zhang D L, Hong L, Yu D Q, Wu Z L, Yang T, Rust M, Tu Y H and Ouyang Q 2022 *bioRxiv preprint* doi:10.1101/2022.02.27.481910
[21] Mastronarde D N 2005 *J. Struct. Biol.* **152** 36
[22] Scheres S H 2016 *Methods Enzymol.* **579** 125
[23] Zivanov J, Nakane T, Forsberg B O, Kimanius D, Hagen W J, Lindahl E and Scheres S H 2018 *Elife.* **7** e42166
[24] Rosenthal P B and Henderson R 2003 *J. Mol. Biol.* **333** 721
[25] Penczek P A 2010 *Methods Enzymol.* **482** 73
[26] Wu Z L, Chen E B, Zhang S W, Ma Y P, Liu C C, Yin C C and Mao Y D 2021 *bioRxiv preprint* doi:10.1101/2021.08.09.455739
[27] Hinton G E and Roweis S T 2002 *Proceedings of the 15th International Conference on Neural Information Processing Systems* edited by Becker S, Thrun S and Obermayer K (Cambridge: MIT Press)
[28] van der Maaten L and Hinton G 2008 *J. Machine Learning Res.* **9** 2579
[29] Kullback S and Leibler R A 1951 *Ann. Math. Stat.* **22** 79
[30] Lin J, Chew J, Chockanathan U and Rust M J 2014 *Proc. Natl. Acad. Sci. USA* **111** E3937
[31] Mori T, Williams D R, Byrne M O, Qin X, Egli M, Mchaourab H S, Stewart P L and Johnson C H 2007 *PLoS Biol.* **5** e93
[32] Ito H, Kageyama H, Mutsuda M, Nakajima M, Oyama T and Kondo T 2007 *Nat. Struct. Mol. Biol.* **14** 1084
[33] Yoda M, Eguchi K, Terada T P and Sasai M 2007 *PLoS One* **2** e408
[34] Nagai T, Terada T P and Sasai M 2010 *Biophys. J.* **98** 2469
[35] Zhang D, Cao Y, Ouyang Q and Tu Y 2020 *Nat. Phys.* **16** 95