PaStEL: Generator of Pathways with Structural Change on Pseudo Free-Energy Landscape from CryoEM Images

Atsushi Tokuhisa‡,[#], Kimihiro Yamazaki~[#], Yuichiro Wada~†[#], Mutsuyo Wada~[#], Takashi Katoh~, Akira Nakagawa~, Yoshinobu Akinaga‡, Yoko Sasakura‡, Yasushi Okuno‡,⊥[*]*

‡RIKEN Center for Computational Science, 7-1-26, Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan

~Fujitsu Ltd.,4-1-1 ,Kamikodanaka , Nkahara-ku, Kawasaki, Kanagawa, 211-8588, Japan

† RIKEN AIP, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan

⊥Graduate School of Medicine, Kyoto University, Shogoin-Kawaharacho, Sakyo-ku, Kyoto 606-8507, Japan

[#] Equal contribution

[*] Corresponding author

ABSTRUCT Noisy cryoEM particle images reflect the conformational heterogeneity of biomolecules and have high potential for the study of biological process. As numerous simulation studies have shown, the study of biological process is attributed to the description of the free energy landscapes on the conformational pathways along with collective variable, which is usually difficult to define. In this study, we propose a methodology to automatically generate plausible conformational pathways via the theoretically isometric latent space trained by deep Auto-Encoder model using cryoEM experimental dataset directly. The proposed method of the PaStEL can speedily show structural change on the plausible conformational pathways along with free energy landscape. Solid theoretical guarantees and tests using synthetic cryoEM data have succeeded in obtaining qualitatively correct energy landscapes on the generated plausible pathways. Furthermore, benchmarking with real cryoEM experimental data of 50S Ribosome has successfully demonstrated that the conformational changes with energy landscapes consistent with existing studies without any manual labor. Finally, the PaStEL was applied to spike proteins of SARS-CoV-2 and successfully characterized the difference in the conformational changes between the wild type and the mutant (D614G) focusing on the Receptor Binding Domain

regions.

1. INTRODUCTION

Biomolecules are soft functional molecules that are responsible for biological activities by controlling important functions such as enzymatic reactions and signal transduction through diverse interactions within cells. In particular, proteins fold into relatively stable natural structures based on their amino acid sequence. Protein structure and function are closely related, and proteins are known to undergo large-scale conformational changes to metastable structures on the millisecond scale when they perform various functions in the cell.

In elucidating the molecular mechanisms of biological processes and in drug discovery to create substances that control biomolecular functions, it is important to elucidate not only native structures but also metastable structures and the continuous structural changes on the pathways connecting them.

Experimental measurements and simulations have been used to capture the continuous conformational changes of biomolecules. Physical quantities obtained from experimental measurements of biomolecules have lower dimensions and lower signal-to-noise ratios than the object being measured, making data interpretation difficult. Statistical analysis compensates for these characteristics, but the individuality reflected in the individual measurement data is lost due to averaging. On the other hand, molecular simulations of biomolecules can reproduce time-series changes in 3D structures down to μsec, level but it is difficult to capture large stochastic conformational changes that move back and forth between associate stable structures. Therefore, there is currently no universal method to elucidate metastable structures and continuous structural changes on the pathways connecting them. A new method that can capture the various structural changes of large biomolecules with more than several hundred residues on a time scale exceeding milliseconds with atomic resolution is desired.

In single-particle structure analysis, one of the representative analysis methods of cryoEM, individual particle images are picked up from micrographs, and 3D density maps are obtained through sifting of dominant images by 2D classification and 3D reconstruction. By performing such a multi-image analysis, the average image can be reconstructed as a 3D density map with a statistically improved signal-to-noise ratio. With the advancement of instruments and analysis methods, the structural resolution of 1.2 Å for biomolecules with little conformational change has almost achieved atomic resolution. On the other hand, multi-image analysis estimates the statistically predominant density position from a large number of measured images, and the flexible regions of biomolecules become statistically indefinite, making the existence of invisible regions problematic.

–In recent years, much effort has also been focused on elucidating the conformational diversity of biomolecules. There are two main types of methods: methods that obtain several discrete average 3D density maps, and methods that continuously obtain structural changes in 3D density maps via deep learning.

RELION, a typical example of former, achieves high resolution by multi-image processing using the maximum likelihood method, and obtains several different discretized average structures by classifying them into multiple classes. Additionally, cryoSPARC, has a user-friendly interface and outputs discrete multi conformations at high speed. Both have been applied to various biomolecules to elucidate their multi conformational states.

On the other hand, cryoDRGN[8], e2gmm[2], and 3Dflex[9] are representative examples of the latter. In these methods, the autoencoder is first trained using the cryoEM image data set. Then, a low-dimensional continuous sequence of latent variables is defined from the trained encoder, and the sequence is passed through the decoder to acquire continuous structural changes in the biomolecules, represented as a 3D density map. For example, the authors in [10] experimentally demonstrated the potential of cryoDRGN by applying it to a cryoEM image dataset of 50s ribosomes to computationally reproduce one of four plausible existing assembly pathways (see Figure 7 in [11]) of manually constructed ribosomes. However, as the author of cryoDRGN himself points out, the equivalence of the potential distribution of the three representative methods mentioned above and the energy distribution of biomolecules has not been fully clarified theoretically [8]. Therefore, the protocol proposed by [10] to construct a plausible conformational change pathway is essentially a heuristic method since it is not based on an energy distribution. And because of its heuristic, the protocol requires more than cryoEM images and analysis; it requires laboratory equipment and biological expertise, resulting in significant effort and time.

–In addition, molecular dynamics simulation (MD) has been used as another important tool to elucidate the multi conformational states of biomolecules. As computing power increases, MDs are becoming longer and larger in scale. However, large-scale structural changes, which are important in elucidating the molecular mechanisms of biological processes, are stochastic processes with a typical time scale of milliseconds, making it difficult to capture their structural transitions in brute-force microsecond MD simulations. This problem is called the sampling problem, and various sampling innovations have been used to capture stochastic and large conformational changes. In McMD, one of the advanced sampling methods, comprehensive sampling of a variety of structures at various temperatures has enabled us to draw detailed free energy landscape, including metastable states on Collective Variables (CV), for relatively small systems of ~100 residues. Once the free energy landscape can be drawn, the energy barriers can be discussed along with the conformational changes to understand the transition states, which is the key to understanding the molecular mechanism. The key to capturing structural transitions is to correctly estimate the most likely conformational path from a large number of possible paths. The biggest problem with this method is that there is a limit to the size of biomolecules to which it can be applied, and a computational explosion occurs as the molecular size increases.

The aim of this study was to develop a new method to capture continuous conformational changes of biomolecules, including those on the millisecond scale, exceeding several hundred residues, which are difficult to analyze experimentally or by simulation. The proposed method of PaStEL learns cryoEM particle images, which are snapshots of multi conformational states, to generate a continuous 3D density map structure change and free energy surface of plausible conformational pathways. The PaStEL uses the cryoTWIN model to obtain a multimodal distribution with GMM representation in a theoretically guaranteed latent space (Theoretically Guaranteed Isometric Latent Space) with the cryoEM particle image as input. This property and the Max-Flux algorithm allow for quick semi-automatic estimation of plausible conformational pathway. This allows us to capture the various comfomational changes of large biomolecules with more than several hundred residues on time scales exceeding milliseconds with high resolution, without requiring prior biological knowledge or expertise in setting appropriate CVs.

Based on the solid theoretical guarantees shown in Result 2.1, the validity of the proposed method was verified by simulation data in Result 2.2. In Result 2.3, experimental data on 50s-ribosomes, for which there is a large amount of known information, show that the estimated conformational pathways are valid compared to the known information. In Result 2.4, PaStEL was applied to the SARS-CoV-2 spike protein, focusing on the Receptor Binding Domains (RBD) regions to characterize the differences in conformational changes between wild-type and mutant (D614G) with plausible pathway energy changes.

2. RESULT

2.1 The PaStEL method:

As briefly explained in the previous section, the State-Of-The-Art (SOTA) unsupervised methods based on an auto-encoder, such as cryoDRGN [8] and e2gmm [2], do not have sufficient analytical results for the relationship between latent and structural distributions, and therefore they do not guarantee the following equivalence theoretically: a protein conformational pathway computed via a low-dimensional latent distribution is equivalent to a plausible conformational pathway (e.g., optimal chemical reaction path a.k.a. MaxFlux pathway [3] and minimum free energy path) computed on a high-dimensional structural distribution. Note that the computational cost of the former pathway is low, whereas that of the latter pathway is usually high.

Our proposed method, PaStEL, achieves the above equivalence in the ideal condition:

(i) The structural distribution of the target protein is a low-dimensional manifold, i.e., the manifold assumption [1] holds for the structural distribution, and

(ii) For reconstructing each structure, the sufficient amount of cryoEM images and their accurate pose orientations are obtained.

PaStEL consists of our auto-encoder namely cryoTWIN and an algorithm that computes the conformational pathways. In PaStEL, first, cryoTWIN is trained using the cryoEM images and their pose orientations. After this training, a simple distributional model $p_\psi(z)$ fits the latent distribution, and the decoder outputs the corresponding 3D density map as input of a latent variable z; see Fig. 1b for the predicting procedure. Under the ideal condition, it is theoretically guaranteed that the latent space and the output space of the decoder (the space of the 3D density map) have an isometric relationship [4]. From this isometric property, we can immediately derive the proportionality between the latent distribution p_ψ and the distribution of the 3D density map, i.e., the structural distribution p; see the proportional relationship between the two distributions in Fig. 1c.

Thanks to the isometric property of cryoTWIN, only PaStEL can have the following two strong theoretical guarantees, compared to the SOTA methods:

(i) Suppose that the cryoEM images of target protein are collected in the equilibrium condition. Then, the free energy of the 3D density map is equivalent to the simplified low-dimensional formula $-\log p_\psi$ except for the constants; see equation (3).

(ii) The protein conformational pathway (as a sequence of the 3D density maps) computed via the low-dimensional latent distribution $p_\psi(z)$ is equivalent to the MaxFlux path computed directly on the high-dimensional structural distribution p; see equation (4).

Our computational algorithm is based on the theoretical guarantee (ii). In this algorithm, firstly, the ridgeline between the starting point z_0 and the end point z_1 given on the trained model p_ψ is obtained as a finite sequence of latent variables; see the orange dash line on the left in Fig. 1c for the ridgeline. Then secondly, the sequence is transformed into a sequence of the 3D density maps by the decoder. At last, the algorithm outputs the 3D density map's sequence.

In practice, it is difficult to obtain a large number of cryoEM images and their accurate pose orientations. In this case, for a limited amount of cryoEM images, we first estimate the pose orientations by an existing technique such as cryoSPARC [5], and then approximate the expected loss of cryoTWIN using the cryoEM images with their estimated orientations. See the definitions of the expected loss and the approximated loss in equation (1) and (5), respectively. In addition, see Fig. 1a for the diagram of how to compute the approximated loss. Here, we remark that the weight to define the weighted squared L2 loss in Fig. 1a is related to the theoretical guarantee of PaStEL.

PaStEL and molecular dynamics can be complementary technologies to each other, because of the above theoretical guarantees (i) and (ii). Further, since PaStEL requires only cryoEM images and reasonable computational resources for the implementation, the method has a potential to revolutionize the process of drug discovery in the future. In the following sections, we evaluate the performance of PaStEL using a limited amount of cryoEM images.

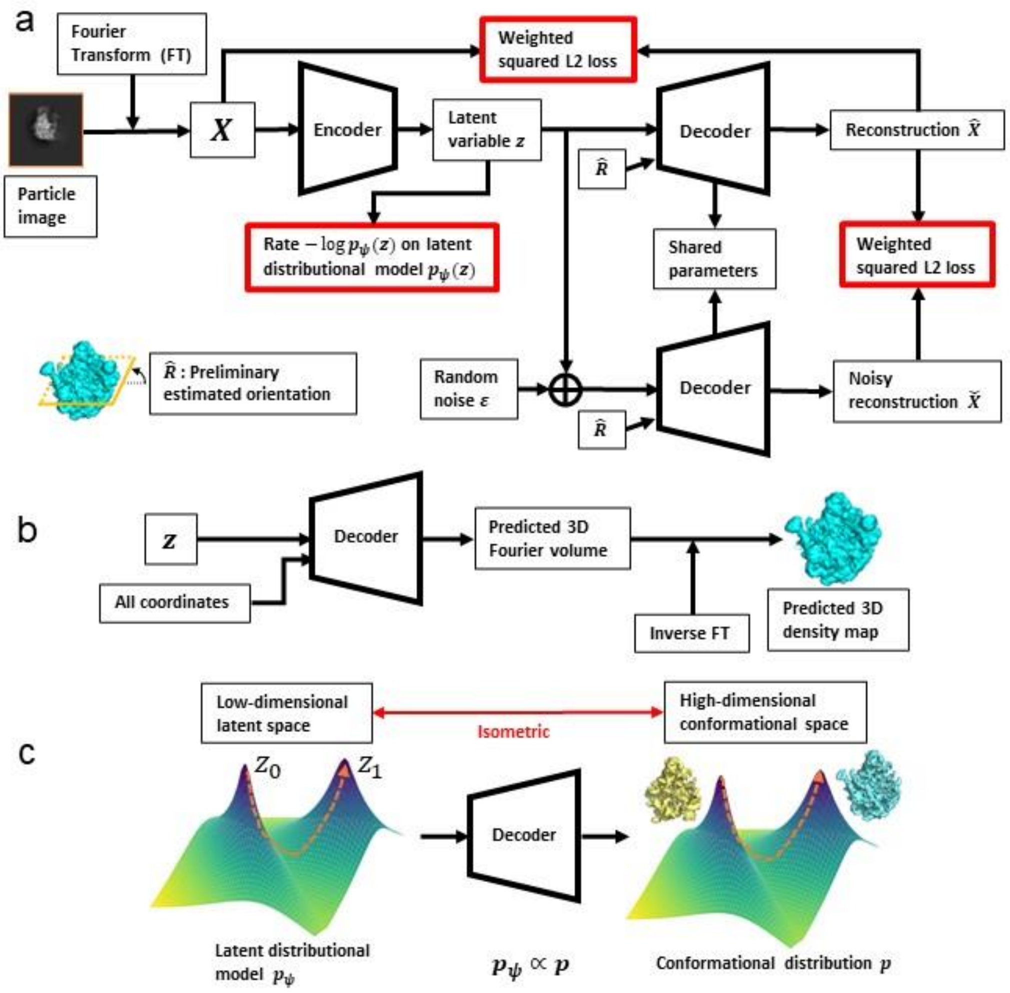

a
Fourier Transform (FT)
Particle image
X
Encoder
Latent variable z
Weighted squared L2 loss
Decoder
Reconstruction $\hat{X}$
$\hat{R}$
Rate $-\log p_\psi(z)$ on latent distributional model $p_\psi(z)$
Shared parameters
Weighted squared L2 loss
$\hat{R}$: Preliminary estimated orientation
Random noise ε
Decoder
Noisy reconstruction $\breve{X}$
$\hat{R}$
b
z
All coordinates
Decoder
Predicted 3D Fourier volume
Inverse FT
Predicted 3D density map
c
Low-dimensional latent space
Isometric
High-dimensional conformational space
z_0
z_1
Decoder
Latent distributional model p_ψ
$p_\psi \propto p$
Conformational distribution p

Figure1. Overview of our method, PaStEL. a, Diagram of how to define loss function in practice for training our auto-encoder (cryoTWIN). The loss is defined by three terms in red rectangles, where $\boldsymbol{\psi}$ is a set of the trainable parameters. b, Diagram of how to predict the corresponding 3D density map to a latent variable $\boldsymbol{z}$ by trained decoder of cryoTWIN. c, Illustration of how our algorithm computes conformational pathway as inputs of start point $\boldsymbol{z_0}$ and end point $\boldsymbol{z_1}$ based on a cryoTWIN with isometric property. First, the orange ridgeline from $\boldsymbol{z_0}$ to $\boldsymbol{z_1}$ is computed, and then it is transformed to conformational pathway by the decoder.

2.2. The validity of the PaStEL by simulation data of chignolin

In this section, the simulation data will be used to validate the pathways presented by PaStEL and the generated density maps on the pathways. The cryoEM experimental data, in which samples were flash-frozen, contain a variety of conformational states, and some correlation between observation frequency and structural stability can be expected. In other words, more stable conformational states are observed more frequently and rare conformational states are observed less frequently. The correlation between the free energy surface, a measure of protein conformational stability, and the distribution in the latent space obtained by learning the observed image set was investigated using the McMD calculation. Specifically, for the Chignolin protein, we employed a McMD-assisted method (see Method 2: Synthetic cryoEM particle images) to prepare a variety of structures and proceeded with verification experiments. 1.5 million artificial 2D images were prepared from free energy-based weighted structural sampling and random orientation projection, which were used as a pseudo cryoEM image set to train the autoencoder cryoTWIN.

Figure 2-a shows the pseudo-free energy surface acquired by training a set of simulated cryoEM images. On the other hand, Figure 2-b shows the free energy surface obtained by McMD, which is the baseline data. In the free energy plane, in addition to the most stable structure observed in the X-ray crystal structure (stable), there are several meta-stable structural, and there is an energy barrier that must be crossed between them. In Figure 2-a, the most stable structure, metastable structure, and energy barrier can be seen at positions similar to the free energy surface obtained by McMD. The correlation value between the free energy surface and the pseudo-free energy surface acquired from the image distribution was sufficiently high at 0.84 (Figure 3-c).

Figure 2-a shows dotted plausible paths transitioning between stable and metastable structures, as estimated by the maximum flux algorithm [7] in latent space. Figure 2-d shows the normalized free energy on the conformational path. The pseudo-free energy values of the plausible pathways estimated by PaStEL are in good agreement with those of McMD. The MD structure and generated volumes on the corresponding pathways are shown in Figure 2-e. In addition to the unevenness of the energy surface, the high similarity between the MD structure on the pathway and the generated volume can be confirmed. On the other hand, it has been confirmed that when the number of images is less than about 500, the reconstructed structure is not reproduced well. The fourth reconstructed volume in Figure 2-e is an example of this problem. The above results show that there is a high correlation between the pseudo-free energy surface obtained by learning the cryoEM images and the free energy surface that is the baseline, and that the proposed method can estimate the pseudo-free energy from a finite set of cryoEM images with high accuracy. In addition, we showed that there is a high similarity between the generated volume on the pathway and the MD based structure that is the baseline. These results indicate that it is possible to analogize the pseudo free energy surface of a protein from cryoEM images alone, if sufficient observational data are available. Furthermore, a plausible path can be selected semi-automatically from a pseudo free energy surface, and the transition of energy and structure on that conformational path can be visualized.

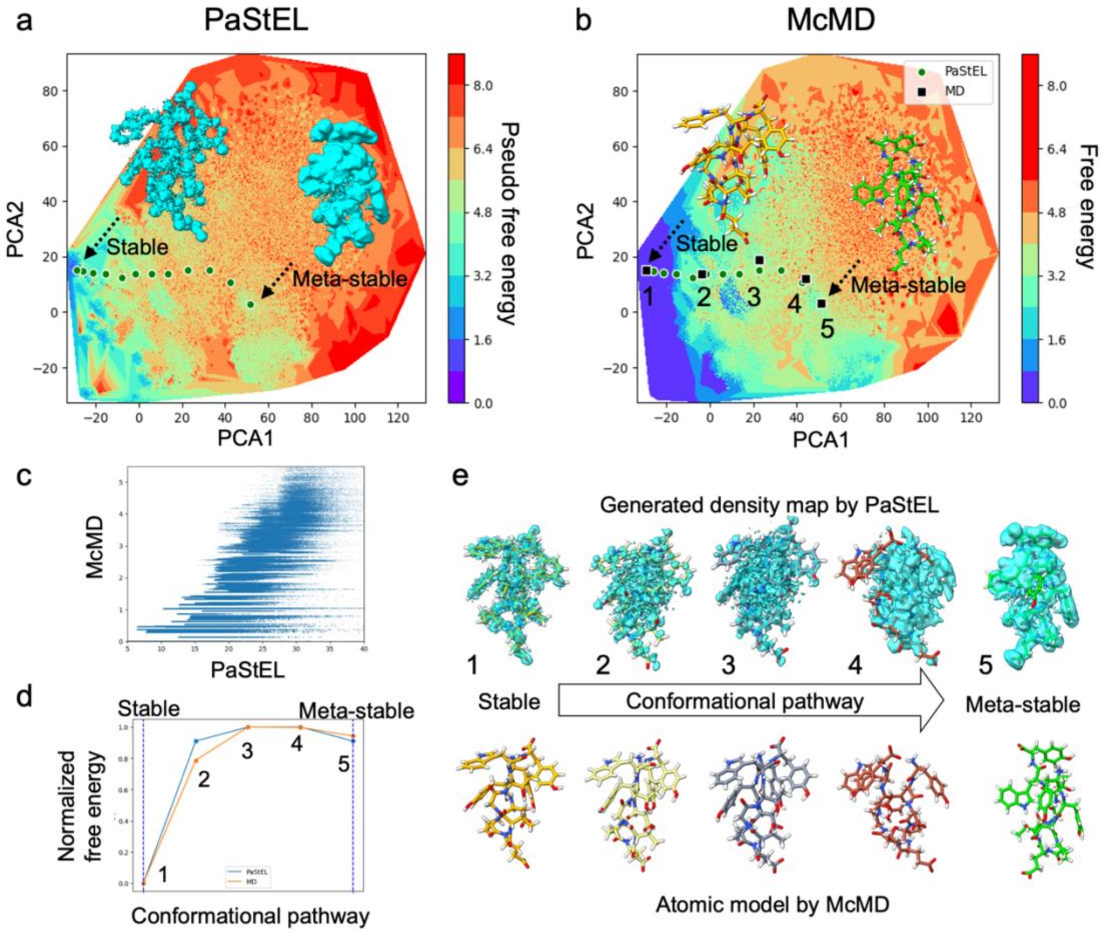


Figure2. (a). Pseudo free energy surfaces estimated by PaStEL. plausible conformational pathways between stable and meta stable states, estimated by PaStEL, are indicated by green dots. (b). The free energy surfaces calculated from the baseline of McMD. MD sampling points that are closest to the estimated conformational pathway are indicated by black dots (ID1-5). (c) Energy scatter diagram. Horizontal axis is pseudo free energy of paStEL. Vertical axis is McMD free energy. (d). Normalized free energy in the estimated path. Blue line: PaStEL. orange line: McMD. (e). Structure transition in the estimated pathway. Upper panel: 3D density map generated by PaStEL. Bottom: Atomic model of MD sampling points.

2.3 Application of PaStEL to experimental data of 50s-Ribosomes

In the previous section, we showed that PaStEL can visualize the transition of energy and structural changes on the pathway from cryoEM images with high accuracy by using simulated data. In this section, we verify the validity of the plausible pathways estimated by PaStEL using actual cryoEM experimental data. The assembly pathway of the large subunits of the Ribosome, a protein synthesis system, was investigated as an example. The Ribosome has been intensively studied because of its biological importance (see references).

About 130,000 particle images of ribosomal cryoEM data (EMPIAR-10076: complex) were used to train of cryoTWIN. The orientation information obtained from the preliminary analysis by CryoSPARC was given in train process. The pseudo-energy surface obtained from the trained cryoTWIN is shown (Fig 3-i). There are several metastable structures in the free energy surface, and important information for analyzing the structural transition is presented here. Based on this energy landscape map, the conformational pathways of conformational transition were calculated by PaStEL, along with their likelihood of occurrence, and the results are shown in Figure 3-a as a non-directed graph. Each node of the graph corresponds to the density map (volume) of the four states of interest (B, C, D, E) of the Ribosome. State B is the structure lacking the most overall density, indicating that it is an immature 50S-Ribosome. The state C is one with increased basal density, while the state D has CP present but lacks basal and intersubunit density. The state E has basal and CP densities present, with changes in density around uL1 and uL10/11 stalk. Each of these conformational states was labeled by FSC with the existing density map (details in Method). The distance between edges also represents the likelihood of structural transitions occurring, obtained by solving the optimal reaction path equation [7]. In other words, the shorter the distance between edges, the more likely a structural transition is to occur. The PaStEL also defines a single conformational pathway by specifying the number of starting, ending, and transit points. From the MaxFluxScore, which is the sum of edge distances, the likelihood of each pathway can be estimated.

Using the assembly process of the 50S-Ribosome as an example, the pathways likely to be caused by PaStEL were investigated. An exhaustive pathway analysis was performed using state B, the initial 50S-Ribosome assembly state, as the starting point and state E, the final 50S-Ribosome assembly state, as the end point. The results are shown in Figures 3-b (4 transit points) and 3-c (5 transit points). Both figures are histogram diagrams of MaxFluxScore obtained by exhaustive analysis, showing that pathways located on the left side are more likely to occur. Existing studies by Davis et al. show that there are four major reaction pathways (p1, p2, p3, and p4) that connect B-E. All of these major reaction system pathways appeared at the top of the MaxFluxScore in our analysis. The p5 pathway, which appeared at the top of the list, was found to contain the C4 structure suggested by the cryoDRGN analysis using same data. In summary, PaStEL can output plausible structural pathways without prior biological knowledge, as long as one has prior knowledge of the 3D structures of the start and end points, a cryoEM image of the target protein, and appropriate computational resources. We further emphasize that the runtime of our method is about 10 times shorter than that of cryoDRGN.

In addition, our method allows us to obtain transitions in energy and conformational changes along the reaction pathway that would otherwise be difficult to obtain. The energy transition diagrams on the main pathway transitioning from B-E are shown in Fig. 3 d,e,f,g,h. Looking at the energy surface in Figure 3d-h, one can see the transition from the relatively high energy B state to the lower energy state E5. This suggests that the change of state from B to E5 is energetically reasonable. There are valleys and peaks of energy between each state, showing different trends in each pathway. Of the known pathways, p1-4, p2 and p3 were also ranked higher in MaxFluxScore, indicating that they are more likely to proceed, while p1 and p4 are less likely to proceed in comparison. Energy peaks are seen between C1 and E2 in p1 of Figure e and between D4 and E3 in p4 of Figure f, indicating that a high energy barrier must be exceeded to transition between them. This process is likely due to the large conformational changes from C1 to E2 associated with CP (the central protuberance) reorganization and the divergence and recombination of uL16 between D4 and E3. The novel pathway p5 including C4 was also suggested to be an easier pathway to advance compared to p1 and p4 in MaxFluxScore.

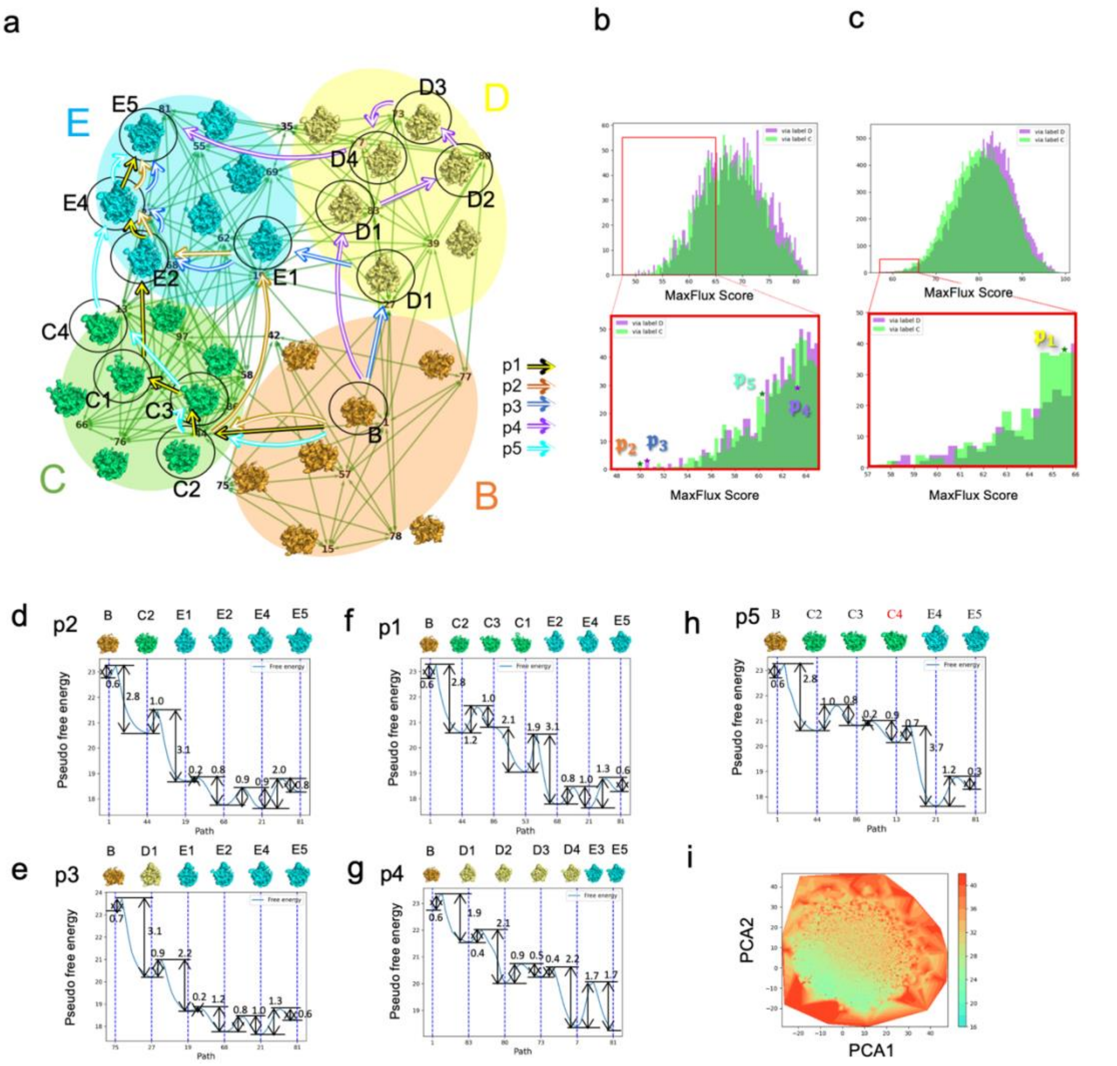


Figure3. (a) Pathway graph estimated from the learned PaStEL. Pathways p1-p4 of the existing studies and the proposed new pathway p5 are shown. (b) Histogram of calculated MaxFluxScore for each pathway with the number of nodes from B to E5 structure set to 4. Pathways of existing studies appear at the top of the histogram. The number of nodes via c and b is set to 4. Pathways of existing studies appear at the top of the histogram. (d) Pseudo free energy transition for the pathway of p2. (e) Pseudo free energy transition diagram for the pathway of p3. (f) Pseudo free energy transition diagram for the pathway of p1. (g) Pseudo free energy transition diagram for the pathway of p4. (h) Pseudo free energy transition diagram for the pathway of p5.(i) Pseudo free energy surfaces estimated by PaStEL.

2.4 Application of PaStEL to experimental data of Spike Protein from SARS-CoV-2

In this section, we applied PaStEL to cryoEM experimental data of two types of Spike Protein (wild type and D614G mutant) in order to clarify the differences in their conformational states. The spike protein from SARS-CoV-2 provides a critical trigger for the virus to enter our bodies by binding to ACE2 in the cell membrane. In particular, receptor binding domain (RBD) regions are the primary targets of neutralizing antibodies, which characterize the binding affinity of SARS-CoV-2. We focused on the three RBDs present in the spike protein and investigated by PaStEL the differences in conformational state between the wild type and the D614G mutant, in which the 614th aspartic acid is mutated to glycine.

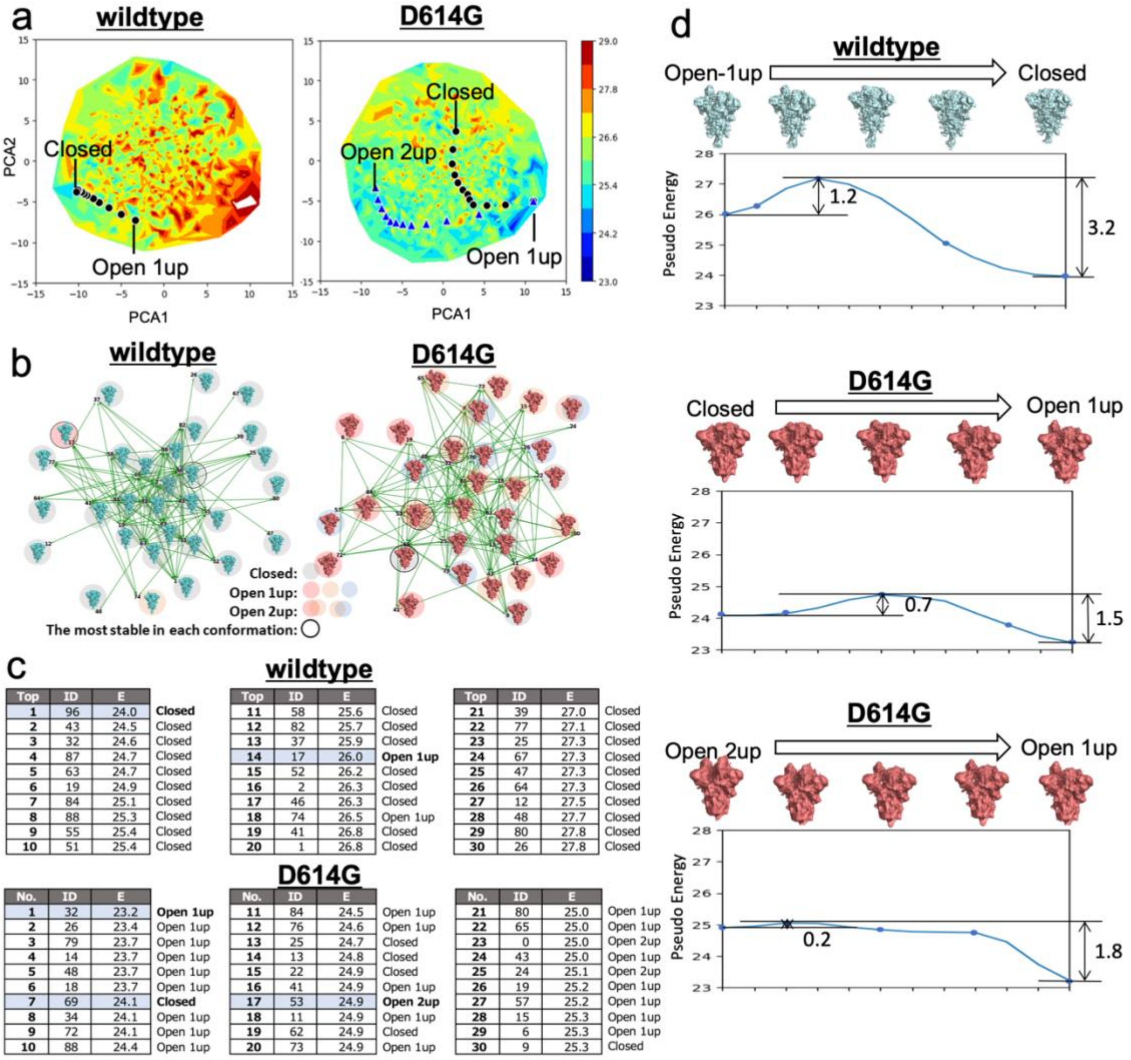
Figure 4-a shows the pseudo-free energy surface (PC1-PC2 surface) estimated by PaStEL. The energy surface of the wild type is more rugged than that of D614G, indicating the presence of many relatively high energy barriers. This suggests that D614G is able to transition more smoothly to another conformational state. Pathway analysis was performed to capture the diversity of conformational states of both species.

Here, an exhaustive pathway analysis was performed for structural changes with the most energetically stable structure as the endpoint and other structures indicated by PaStEL as the starting point. The energy surfaces and structure types for each structure are shown in Figure 4-b. Unlike the Ribosome case, which searches for assembly pathways, the point of transit was set to 0 as in the case of chignolin, because the protein conformational changes are comprehensively searched for. The top 30 densest nodes were selected from the average vector of GMMs approximated by 100 Gaussians. The MaxFluxScore was calculated for all edges between the 30 nodes, and to improve readability, the edges below the top 1/3 were cut off and an undirected graph was plotted (Figure 4-b). Figure 4-c shows the node IDs and energy values of the top 30 nodes for the wild type and mutant, respectively. In the wild type, 28 states were closed states and 2 states (IDs 17 and 74) were Open 1up states with one RBD up. In contrast, the D614G mutant had 21 Open 1up states, 6 Close states (ID 7, 25, 13, 22, 62and 9), and 3 Open 2up states (ID 17, 23 and 25) with two RBDs up. There is a clear difference in conformational state between the them, with an overwhelming probability of being present in the Open 1up state in the D614G mutant.

Next, the details of the structural changes between the different conformational states were examined. The topmost Close state of the wild type (ID 96) and the conformational change of Open 1up (ID 14), which is the highest (14th) of the different conformational states, and the pseudo-free energy change on the conformational pathway are shown in the top row of Figure 4-d. The Closed state is about 2 units more stable than the Open 1up state. A relatively high energy barrier of 1.2 was found to exist in the transition state. On the other hand, the transition state from the Closed to Open 1up states of the D614G mutant is shown in the middle panel of Figure 4-d. In the mutant, the Open 1up state is 0.8 more stable than the Closed state. The transition state of the mutant has a small barrier of 0.7 compared to the wild type, suggesting that it can easily transition between the Closed and Open 1up states. The transition state of the mutant has a relatively small barrier of 0.7 compared to the wild type, suggesting that it can easily transition between the Closed and Open 1up states. The transition state of the mutant has only a small barrier of 0.7 compared to the wild type, suggesting that it can easily transition between the Closed and Open 1up states.

The results also suggest that Open 1up and Open 2up conformational transitions occur with some frequency in the mutant (Figure 4, lower panel). The barrier between the transition energies of Open 1up and Open 2up is extremely small at 0.2, indicating that both conformational transitions occur easily.

In addition, these three conformational change pathways are shown on the PCA1-PCA2 plane in Figure 4-a. In the mutant, there are multiple conformational state change pathways, centered on the restable Open 1up and changing to Closed and Open 2up. In summary, the transition structures of intermediate states with high energy barriers can be obtained by this method when transitioning between each conformational state.



wildtype

Top	ID	E	
1	96	24.0	**Closed**
2	43	24.5	Closed
3	32	24.6	Closed
4	87	24.7	Closed
5	63	24.7	Closed
6	19	24.9	Closed
7	84	25.1	Closed
8	88	25.3	Closed
9	55	25.4	Closed
10	51	25.4	Closed
11	58	25.6	Closed
12	82	25.7	Closed
13	37	25.9	Closed
14	17	26.0	**Open 1up**
15	52	26.2	Closed
16	2	26.3	Closed
17	46	26.3	Closed
18	74	26.5	Open 1up
19	41	26.8	Closed
20	1	26.8	Closed
21	39	27.0	Closed
22	77	27.1	Closed
23	25	27.3	Closed
24	67	27.3	Closed
25	47	27.3	Closed
26	64	27.3	Closed
27	12	27.5	Closed
28	48	27.7	Closed
29	80	27.8	Closed
30	26	27.8	Closed

D614G

No.	ID	E	
1	32	23.2	**Open 1up**
2	26	23.4	Open 1up
3	79	23.7	Open 1up
4	14	23.7	Open 1up
5	48	23.7	Open 1up
6	18	23.7	Open 1up
7	69	24.1	**Closed**
8	34	24.1	Open 1up
9	72	24.1	Open 1up
10	88	24.4	Open 1up
11	84	24.5	Open 1up
12	76	24.6	Open 1up
13	25	24.7	Closed
14	13	24.8	Closed
15	22	24.9	Closed
16	41	24.9	Open 1up
17	53	24.9	**Open 2up**
18	11	24.9	Open 1up
19	62	24.9	Closed
20	73	24.9	Open 1up
21	80	25.0	Open 1up
22	65	25.0	Open 1up
23	0	25.0	Open 2up
24	43	25.0	Open 1up
25	24	25.1	Open 2up
26	19	25.2	Open 1up
27	57	25.2	Open 1up
28	15	25.3	Open 1up
29	6	25.3	Open 1up
30	9	25.3	Closed

Figure4. Differences in the structural states of two types of Spike Protein (wild type and D614G mutant type) a, PCA projection of latent space of the wild type (left) and D614G mutant (right). Structural transition from Open 1up to Closed in the wild type (black circle), Structural transition from Closed to Open 1up in D614G mutant (black circle), Structural transition from Open 2up to Open 1up (blue triangle). b, Structural transition generated from the latent space of a of the wild type and D614G mutant. c, Energy transition of the structures corresponding to b of the wild type and D614G mutant.

3. DISCUSSION

The advantages of the proposed method are described. Higher dimensional spaces are usually required to describe conformational changes in flexible biomolecules. Until now, methods such as projection onto low-dimensional space, as typified by PCA, have been used to capture the characteristics of conformational changes in biomolecules. The proposed method, PaStEL, can acquire a low-dimensional space that captures the features of structural deformation of biomolecules through deep learning using oriented cryoEM images as input. By calculating the MaxFluxScore in the acquired low-dimensional space described by GMM, a plausible structural transformation path can be extracted semi-automatically based on the distribution of experimentally observed images. This allows visualization of the continuous conformational change of the biomolecule, along with the pseudo-energy surface on the conformational change pathway, by means of a 3D density map representation.

This result strongly indicates that the manifold hypothesis holds for the conformational space of biomolecules. At the same time, cryoEM images are able to capture snapshots of multi conformational states of biomolecules, indicating that they have high potential for analyzing the multi conformation of biomolecules. Non-deep learning cryoEM analysis has mainly analyzed the average structure or a few classified structures, which requires specialized knowledge. We emphasize that the PaStEL can perform the same analysis semi-automatically as an expert would have done, by giving the number of transit points at the beginning and end of the pathway.

We also emphasize that the real time required for this has been reduced by about 10-folds of the existing methods when 50S-Ribosome is used, and we have achieved a fast process.

On the other hand, we believe that the proposed method has the following limitations at this time. First, it estimates a pseudo-energy landscape that depends on the single-particle image set used in the analysis. In order to obtain the more true energy landscape of biomolecules, a data set that exceeds the number of images required for 3D reconstruction analysis is presumably necessary. This suggestion may change the way experimental data is taken, and it is hoped that more images will be acquired in the near future under experimental conditions that include a wider range of structural states. In the current implementation, it is also necessary to provide orientation estimations to the images as prior information; the estimation of image orientation for cryoEM images relies on the implementation of RELION and cryoSPARC, which are preanalysis, and requires specialized knowledge.

As another problem, the proposed method is affected by multi-image processing, and the problem remains that the flexible regions as the average density map are indefinite on the 3D density map. In addition, rare structures have a small number of observed images, and their effects appear in the 3D reconstructed structure, such as the loss of volume. At present, it is known that the accuracy of 3D entertainment manufacturing declines when the number of images strongly associated with a particular structure falls below 500. We are currently working on improving this by incorporating devices that artificially bring smoothness to the latent space. e.g., Virtual Adversarial Training [IEEE PAMI 2018].

The potential to utilize the acquired latent space is high, and it may be possible to efficiently eliminate garbage data in the cryoEM experimental data by removing outliers in the distribution of the low-dimensional space.

The most promising field for social implementation of the proposed method is drug discovery. In drug discovery, it is important to consider various structures of target proteins for rational molecular design. The proposed method, which semi-automatically provides continuous conformational change pathways, can be a powerful tool in drug discovery. On the other hand, it is not enough to capture the continuous or structural changes as a 3D density map to link to drug discovery. In the future, it will be necessary to construct an all-atom structure model from the density map, and for this purpose, it is important to actively utilize not only the structure obtained by molecular dynamics simulations such as FF, but also inferential atomic structure models such as AF2.

4. CONCLUSION

In this paper, we propose PaStEL, which efficiently estimates plausible conformational pathways using a set of oriented cryoEM images as input. Normally, the optimal reaction path equation must be solved in a high-dimensional space, but our method can solve a completely equivalent equation in a low-dimensional space via a theoretically guaranteed isometric potential space. This allows us to speedily show conformational changes on plausible pathways along with pseudo-free energy landscapes.

The performance of PaStEL was verified using synthetic data with the MD simulations and cryoEM experimental data for 50s-Ribosomes. As a result, we succeeded in obtaining a high correlation of 0.84 between the free energy surface calculated by the simulation and the pseudo free energy surface obtained by PaStEL. In the assembly pathway analysis of the 50s-Ribosome, we succeeded in semi-automatically obtaining pathways consistent with those shown by existing studies.

In addition, by applying the proposed method to two spike proteins from SARS-CoV-2, we elucidated the difference in conformational stability states between wild type (D614) and mutant (D614G), along with pseudo-free energy values on the conformational pathways.

The above means that we can provide a new method to semi-automatically and speedily analyze the continuous and rapid structural changes of biomolecules, which have been difficult to analyze because of the high-dimensional space, through an equivalent low-dimensional space obtained by deep learning.

5. METHODS

5.1 Theoretical detail of PaStEL

Let f_θ, g_ϕ, p_ψ denote an encoder, decoder, and latent distributional model in cryoTWIN, respectively. The symbols θ, ϕ, ψ denote a set of trainable parameters. In addition, let $X_R \in \mathbb{R}^{s \times s}$ denote the Fourier transformation of an cryo-EM image, whose pose orientation is R in a 3D density map V. The decoder g_ϕ outputs $\hat{X}_R := g_\phi(z, R)$, i.e., the reconstruction of X_R, as inputs of $z \in \mathbb{R}^d$ and R, where z is a latent variable, and this latent variable is an output of the encoder f_θ. Moreover, let $\check{X}_R := g_\phi(z + \varepsilon, R)$ denote the noisy reconstruction with X_R. The symbol ε is a random noise vector satisfying the following condition: $\varepsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_d) \in \mathbb{R}^d, \forall i; \mathbb{E}[\epsilon_i] = 0, \forall (i, j); \mathbb{E}\left[\epsilon_i \epsilon_j\right] = \sigma^2 \mathbf{1}_{[i=j]}$. The function $\mathbf{1}_{[i=j]}$ is the indicator function: $\mathbf{1}_{[i=j]} = 1$ if $i = j$, otherwise 0. Furthermore, let W denote a $s \times s$ matrix, whose (i,j)-th $(-s/2 \leq i,j \leq s/2, \text{and } i,j \in \mathbb{Z})$ factor is $\sqrt[4]{i^2 + j^2}$. The training objective of cryoTWIN is given as follows:

$$\theta^*, \phi^*, \psi^* = \arg\min_{\theta,\phi,\psi} \mathbb{E}_{V\sim p,\varepsilon}[L], \tag{1}$$

where $L := -\log p_\psi(z) + \lambda_1 \mathbb{E}_R[\| W \odot (X_R - \hat{X}_R) \|^2 \, |V] + \lambda_2 \mathbb{E}_R[\| W \odot (\hat{X}_R - \check{X}_R) \|^2 \, |V]$. The symbols $*$, $\odot$, and λ_1, λ_2 mean the optimality, Hadamard product, positive hyper-parameters, respectively. From Appendix A of our previous work [7], L in equation (1) is equivalent to $-\log p_\psi(z) + \lambda_1 \pi \parallel V - \hat{V}_z \parallel^2 + \lambda_2 \pi \parallel V - \check{V}_z \parallel^2$, where $\hat{V}_z$ (resp. $\check{V}_z$) denotes the predicted 3D density map (resp. noisy 3D density map) with the latent variable z (resp. noisy latent variable $z + \varepsilon$) via the decoder. Thus, from equation (9) in our previous work [4], cryoTWIN trained by equation (1) guarantees the isometric relationship between the latent space and the output space by the decoder (the space of the 3D density map); see the mathematical expression of the isometric property in equation (14) of [4]. By this isometry, the following properties immediately hold:

$$\|z - z'\| \approx 0 \Rightarrow \|z - z'\| \propto \|V_z - V_{z'}\|, \text{and } p_{\psi^*}(z) \propto p(V_z), \tag{2}$$

where V_z expresses the true 3D density map with respect to z.

Assume that the cryo-EM images are collected in the equilibrium condition. We here consider the mathematical expression of the free energy with V_z using the trained model $p_{\psi^*}(z)$. Firstly, from the definition of Boltzmann distribution, $E(V_z) = -k_B T \log p(V_z) + \text{const.}$, where k_B and T are the Boltzmann constant and the temperature, respectively. Then secondly, from the second proportionality in equation (2), $\log p(V_z) = \log p_{\psi^*}(z) + \text{const.}$ holds. Therefore,

$$E(V_z) = -k_B T \log p_{\psi^*}(z) + \text{const.}. \tag{3}$$

Next, we prove that a conformational pathway computed via $p_{\psi^*}(z)$ can be equivalent to a MaxFlux path computed directly on the structural distribution $p(V_z)$. To do so, we first consider a sequence of the latent variables with the start point z_0 and end point z_1: $z^{(1:m)} := (z^{(1)}, z^{(2)}, \ldots, z^{(m)})$, which satisfies $m \gg 1$, $\forall j \in \{1, \ldots, m+1\}; \parallel z^{(j-1)} - z^{(j)} \parallel \approx 0$, and both $z^{(0)} := z_0$ and $z^{(m+1)} := z_1$ are fixed points. Suppose that $p(V_z)$ is a connected manifold, and consider the following function for the sequence $z^{(1:m)}$: $\Delta := \sum_{j=1}^{m+1} \frac{1}{p(V_{z^{(j)}})} \parallel V_{z^{(j-1)}} - V_{z^{(j)}} \parallel$.

The MaxFlux path between V_{z_0} and V_{z_1} is defined by a minimizer of Δ, and the minimizer is expressed as a sequence of the 3D density maps: $(V_{z_0}, V_{z^{(1)}}, \ldots, V_{z^{(j)}}, \ldots, V_{z^{(m)}}, V_{z_1})$. From equation (2), the aforementioned minimization problem with Δ is equivalent to

$$\arg\min_{z^{(1:m)}} \Delta = \arg\min_{z^{(1:m)}} \widetilde{\Delta}, \tag{4}$$

where $\widetilde{\Delta} \coloneqq \sum_{j=1}^{m+1} \frac{1}{p_{\psi^*}(z^{(j)})} \parallel z^{(j-1)} - z^{(j)} \parallel$. This equation (4) implies that a sequence of the 3D density maps obtained by decoding the MaxFlux path on the low-dimensional latent distribution $p_{\psi^*}(z)$ is equivalent to the MaxFlux path computed on the high-dimensional structural distribution $p(V_z)$.

5.2 Implementation of cryoTWIN

Let us consider the following practical condition: the number of cryo-EM images is not large, and their accurate pose orientations are not provided. Let X and $\hat{R}$ denote the Fourier transformation of a cryo-EM image and the estimated pose orientation, respectively. In this case, equation (1) is approximated based on Monte Carlo method, using the set $\{(X_i, \hat{R}_i)\}_{i=1}^{N}$:

$$\arg\min_{\theta,\phi,\psi} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\varepsilon}\left[-\log p_{\psi}(z_i) + \lambda_1 \parallel W \odot \left(X_i - \hat{X}_{\hat{R}_i}\right) \parallel^2 + \lambda_2 \parallel W \odot \left(\hat{X}_{\hat{R}_i} - \check{X}_{\hat{R}_i}\right) \parallel^2\right], \qquad (5)$$

where $\hat{X}_{\hat{R}_i} = g_{\phi}(z_i, \hat{R}_i), \check{X}_{\hat{R}_i} = g_{\phi}(z_i + \varepsilon, \hat{R}_i)$, and $z_i = f_{\theta}(X_i)$. In this paper, each element in the noise vector $\varepsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_d) \in \mathbb{R}^d$ follows identical and independent uniform distribution, whose mean is zero. The latent distribution model is defined by Gaussian Mixture Model (GMM) $p_{\psi}, \psi = \{(\pi_c, \mu_c, \Sigma_c)\}_{c=1}^{C}$, where π_c, μ_c, Σ_c are the weight, mean, and variance of the c-th Gaussian, respectively. Furthermore, the computation of the decoder g_{ϕ} in equation (5) is based on an MLP-type neural network like cryoDRGN [8]: the network takes the latent variable z and the 3D coordinates as inputs and outputs the weights of the corresponding 3D coordinates in the 3D Fourier volume. The computation with the encoder f_{θ} in equation (5) is by an MLP-type neural network that transforms the Fourier image X into the latent variable z. Throughout Result 2 to Result 4, the MLP architecture is 1024 x 3 (d x l, d: number of nodes per layer; l: number of layers) for both encoder and decoder, and the dimension of latent variable is fixed to 8. In addition, equation (5) is defined with $\lambda_1 = 1/8, \lambda_2 = 1/8$, and the number of GMM components C is set to 100. The trainable parameters in the auto-encoder are randomly initialized, and they are tuned based on equation (5) using the RAdam optimizer [12]. The learning rate of the RAdam is 0.0001, the number of epochs is 100, and the mini-batch size is 128 (only Result 2 employs a size of 256). Finally, the computing environment is NVIDIA V100 GPU with 2 Intel Xeon Gold 6148 processors.

5.3 Pathway computing algorithm

The algorithm requires the trained GMM $p_{\psi}, \psi = \{(\pi_c, \mu_c, \Sigma_c)\}_{c=1}^{C}$ of equation (5) and the two influential mean vectors μ_i and μ_j as inputs. Then, it approximately solves the right hand minimization problem of equation (4), whose minimizer is the MaxFlux path between μ_i and μ_j on p_ψ, and the minimizer is expressed as a sequence of m latent variables: $z^{(1:m)} = (z^{(1)}, z^{(2)}, \dots, z^{(m)})$. Thereafter, the algorithm decodes $z^{(1:m)}$ into a sequence of the 3D density maps, using the trained decoder g_ϕ. At last, this 3D density maps' sequence is defined as the output, i.e., the plausible conformational pathway. Here, the influential mean is defined by μ with large density value $p_\psi(\mu)$. Additionally, the approximated solution $z^{(1:m)}$ is obtained by conducting m-times greedy optimizations. Suppose that we finish $t-1$-th $(1 \leq t \leq m)$ greedy optimization, and we obtain $z^{(t-1)}$. Then, at t-th $(1 \leq t \leq m)$ greedy optimization, firstly, we generate sufficient amount of C-dimensional random weight vectors, $(\alpha_1, \dots, \alpha_C)^\top$, satisfying $\sum_{c=1}^{C} \alpha_c = 1, \forall c; \alpha_c \geq 0, (\alpha_i, \alpha_j) = ((1 - t/m)^\omega, t/m)$, where only α_i and α_j are constants, and the others are random variables. Thereafter, we solve $z^{(t)} := \underset{\{\mathrm{z}\}}{\mathrm{argmin}} \frac{1}{p_\psi(\mathrm{z})} \left\| z^{(t-1)} - \mathrm{z} \right\|$, where $\mathrm{z} := \left(\sum_{c=1}^{C} \alpha_c \Sigma_c^{-1}\right)^{-1} \left(\sum_{c=1}^{C} \alpha_c \Sigma_c^{-1} \mu_c\right)$, and it is defined by the generated weight vector. By its definition, $z^{(1:m)}$ is a sequence starting at μ_i and converging to μ_j; from the previous study [5], $z^{(t)}$ $(t = 1, \dots, m)$ belongs to a set of candidate points, which locally maximize p_ψ. Throughout Result 2 to Result 4, we set $(m,\ \omega) = (11, 1.2)$.

5.4 Experimental cryoEM data preparation

50S Ribosome system

Since we employ cryoDRGN [28] as the baseline method, the ribosome tutorial dataset provided by cryoDRGN was used as our benchmark system. This data set consists of 131,899 images with an image size of 128x128. In the cryoDRGN tutorial, 3D reconstruction has already been performed to estimate the orientation information of each particle image(https://github.com/zhonge/cryodrgn_empiar/tree/main/empiar10076/inputs). We used those values as well.

SARS-CoV-2 Spike protein systems, wild type:D614 and mutant:D614G

The cryoEM experimental images of two SARS-CoV-2 spike proteins were processed using the stored data in EMPIAR data base. EMPIAR-10469 was used for the wild-type of D614 cryoEM experimental data. EMPIAR-10725 was used for the mutant of D614G cryoEM data. Image processing was performed using RELION software according to normally procedures. After conducting motion correction of measured movies and CTF estimations, particles images was picked up by manually and automatically using template-based particle picking algorithm. All particles images were classified into 5 classes by 3D classification algorithm and the class with clear structural density map was selected. Finally, the number of particle images of the wild type was 84,945 with a resolution of 4.1

Å. The number of images of the mutant of D614G was 40,954, with a resolution of 6.6 Å. The orientation information and particle images prepared in this way were used as input data for training of cryoTWIN.

5.5 Preparation of synthetic cryoEM particle images

Preparation of various structure sets by McMD

To prepare datasets with a variety of structures, we performed multicanonical (Mc)MD simulations using Gromacs (version 2018.2) software. McMD simulations consist of three steps: preparation of the system, a prerun to iteratively estimate the density of states, and a broad structural production run to sample the ensemble.First, as preparation, the target system, Chignolin (PDB-ID: 5awl), was placed in a dodecahedral solvated water box containing Na and Cl ions at 0.1 M concentration and subjected to 100 ps of energy minimization, NVT- and NPT-MD simulations. The AMBER99SB-ildn-ions force field23 and TIP3P water24 were used for molecular parameterization; LINCS25 was used to constrain the protein with a Bussi thermostat26 and a Parrinello-Rahman barostat27. Long-range electrostatic fields were calculated using the Zero-Dipole Summation method28,29 and the electrostatic and van der Waals cutoffs were set at 12 Å.

Once the system was prepared, pre-run and productive runs of the McMD simulation were performed with the multicanonical temperature constrained to a specific target range between 280K and 700K. The pre-run procedure requires multiple iterations of sampling to estimate the density of states and correct bias to allow for random walks over a wide energy range. For Chignolin, the number of iterations was 32; the McMD weighting function is updated between iterations; after determining the proper bias for McMD in the prerun, a 3.2 microsecond (32 x 100 ns) production run was performed. Structures were extracted at equal time intervals from the entire ensemble of structures sampled by the McMD production run. The resulting data set contained variety of 292,693 Chignolin structures with the free energy value.

Simulations of synthetic particle images from prepared structures

We prepared a dataset of 1.5 million simulated particle images consisting of various molecular orientation with 128x128 (A/pix: 0.2) size from different structures by following steps 1-4 as described below from the structures.

STEP1. Calculate the Boltzmann factor, exp(-E/RT); T = 300 K, for each sampled structure from the free energy value calculated by McMD..

STEP2. Perform N (target total number of images 1.5 million) weighted restoration extractions from the population of sampled structures, using the Boltzmann factors as weights. This determines the number of 2D projection images, n, to be computed from that structure for each sampled structure.

STEP3. To determine the orientation of the projection image for each of the n assigned to each sampled structure, a uniform restoration extraction from a uniform hemispheric orientation list (10,240 orientations) is used to create a projection orientation list for each sample structure.

STEP4. Each structure is best fitted based on all atoms, excluding hydrogen atoms, and a 3D density map is calculated by electron density calculation. Based on the projection orientation list created in step 3, 1.5 million images were obtained from the 3D density map for each structure to create a synthetic cryoEM particle images.

5.6. Numerical experiment details

Details of Numerical Experiments on Result 2.2

We trained cryoTWIN using equation (5) of Method 1.(2) with 1.5 million Chignolin simulated particle images with 128x128 (Å/pixel: 0.2) size and their orientation information as training data. It takes about 31 hours 35 minutes (18 minutes 57 seconds per epoch) to train the data. To obtain Fig 2a,b, Principal Component Analysis, PCA , was performed using scikit-learn OSS library.

Details of Numerical Experiments on Result 2.3

We trained cryoTWIN using equation (5) of Method 1.(2) with 131,899 50S ribosome particle images of size 128x128 (Å/pixel: 3.275) stored in EMPIAR-10076 and their estimated orientation (orientation information estimated by cryoSPARC publicly shared by the author in [8]) as training data. It takes about 11 hours 36 minutes (6 minutes 57 seconds per epoch) to train the data. For the visualization in Fig. 3a, we first focused on the top 30 average vectors among the 100 average vectors of all GMM components in terms of the size of GMM density. Then, the MaxFlux value (see equation (4)) was calculated for all edges between the 30 nodes, and only the top 1/3 edges were visualized in order of decreasing value and plotted in the graph.

Details of Numerical Experiments on Result 2.4

Two particle image data sets were prepared from cryoEM raw data stored in EMPIAR using RELION. Using those particle image data sets, we performed 3D reconstruction in five classes to obtain orientation information for each image. Of the five classes, the subclass that showed the clearest density map was selected and used as the training data set. The wild-type (EMPIAR-10469) dataset contains 84,945 particle images with a size of 256 x 256 (Å/pixel: 1.087). Using these particle images and the estimated orientation, we trained cryoTWIN using equation (5) in Method 1.(2), and the training time was 7 hours 50 minutes (4 minutes 41 seconds per epoch). The mutants D614G (EMPIAR-10469) dataset contains 84,945 particle images with a size of 256 x 256 (Å/pixel: 0.8566). Using these particle images and the estimated orientation, we trained cryoTWIN using equation (5) in Method 1.(2), and the training time was 3 hours 46 minutes (2 minutes 15 seconds per epoch).

AUTHOR INFORMATION

Corresponding Author

tokuhisa@riken.jp, okuno.yasushi.4c@kyoto-u.ac.jp.

ORCID

Atsushi Tokuhisa: 0000-0002-9584-1819

Kimihiro Yamazaki: 0000-0003-4827-9488

Yuichiro Wada: 0000-0002-5214-1265

Mutsuyo Wada: 0009-0002-8934-9972

Takashi Katoh: 0009-0003-3849-7649

Akira Nakagawa: 0009-0008-1563-1573

Yoshinobu Akinaga: 0000-0002-3495-1758

Yasushi Okuno: 0000-0003-3596-4208

Author Contributions

A.T. and Y.O. designed the study. A.T., Y.A. and Y.S. performed the data preparation of the simulated image data of chignolin protein and the experiment image data analysis of spike protein. A. N. and K. Y. enhanced mathematical foundation of PaStEL from our MICCAI previous work. Y. W., T. K. and K.Y. constructed designed practical PaStEL algorithm. K. Y. developed PaStEL software. K.Y., M. W., A.T., Y.A, and Y. S conducted pathway analysis using PaStEL. All authors reviewed the final manuscript.

Notes

The authors declare the following competing financial interest(s): This research was conducted as part of a collaborative research effort with Fujitsu Limited. Fujitsu Limited provided funding and technical support for this study. The authors ensured that the research design, data analysis, and interpretation remained unbiased and independent.

Funding Sources

This work was supported by the FOCUS Establishing Supercomputing Center of Excellence project; MEXT as "Program for Promoting Researches on the Supercomputer Fugaku" (Simulation- and AI-driven next-generation medicine and drug discovery based on "Fugaku", JPMXP1020230120).

This work used computational resources of the supercomputer Fugaku provided by RIKEN Center for Computational Science through the HPCI System Research Projects (Project IDs: hp220078, hp230102, hp230216, ra000018); the supercomputer system at the information initiative center, Hokkaido University, Sapporo, Japan through the HPCI System Research Projects(Project IDs: hp220078, hp230102, hp240109) .

Computational resources of AI Bridging Cloud Infrastructure (ABCI) provided by the National Institute of Advanced Industrial Science and Technology (AIST) was also used.

Acknowledgements

We would like to thank Prof. N. Kamiya (University of Hyogo) and G-J, Bekker (Osaka University) for help with data preparation of McMD.

References

[1] Chapelle, O., Schlkopf, B., and Zien, A. (2010). Semi-Supervised Learn$^{\text{in}}$g. The MIT Press, 1st edition.

[2] Chen, M. and Ludtke, S. J. (2021). Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryoEM. Nature Methods, 18(8):930–936.

[3] Huo, S. and Straub, J. E. (1997). The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. Journal of Chemical Physics, 107(13):5000–5006.

[4] Kato, K., Zhou, J., Sasaki, T., and Nakagawa, A. (2020). Rate-Distortion Optimization Guided Autoencoder for Isometric Embedding in Euclidean Latent Space. In $^{\text{Pr}}$oceedings of the 37th International Conference on Machine Learning, ICML 2020, volume 119, pages 5166–5176.

[5] Punjani, A., Rubinstein, J. L., Fleet, D. J., and Brubaker, M. A. (2017). cryoSPARC: Algorithms for rapid unsupervised cryoEM structure determination. Nature Methods, 14(3):290–296.

[6] Ray, S. and Lindsay, B. G. (2005). The topography of multivariate normal mixtures. The Annals of Statistics, 33(5):2042–2065.

[7] Yamazaki, K., Wada, Y., Tokuhisa, A., Wada, M., Katoh, T., Umeda, Y., Okuno, Y., and Nakagawa, A. (2023). An Auto-Encoder to Reconstruct Structure with CryoEM Images via Theoretically Guaranteed Isometric Latent Space, and Its Application for Automatically Computing the Conformational Pathway. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pages 394–404, Cham. Springer Nature Switzerland.

[8] Zhong, E. D., Bepler, T., Berger, B., and Davis, J. H. (2021). CryoDRGN: reconstruction of heterogeneous cryoEM structures using neural networks. Nature Methods, 18(2):176–185.

[9] Punjani, A. and Fleet, D. J. (2023). 3DFlex: determining structure and motion of flexible proteins from cryoEM. Nature Methods, 20(6):860–870.

[10] Kinman, L., Powell, B., Zhong, E., Berger, B., and Davis, J. (2023). Uncovering structural ensembles from single-particle cryoEM data using cryoDRGN. Nature Protocols, 18(2):319–339.

[11] Davis, J. H., Tan, Y. Z., Carragher, B., Potter, C. S., Lyumkis, D., and Williamson, J. R. (2016). Modular assembly of the bacterial large ribosomal subunit. Cell, 167(6):1610–1622.

[12] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. In: International Conference on Learning Representations, ICLR 2020 (2020)

[13] Shaw, D. E. et al. Anton, a special-purpose machine for molecular dynamics simulation. Communications of the ACM 2008, 51 .

[14] Fujita, K.; Iwaki, M.; Iwane, A. H.; Marcucci, L.; Yanagida, T. Switching of myosin-V motion between the lever-arm swing and Brownian search-and-catch. Nature Communications 2012, 3 20

[15] Djumagulov, M.; Demeshkina, N.; Jenner, L.; Rozov, A.; Yusupov, M.; Yusupova, G. Accuracy mechanism of eukaryotic ribosome translocation. Nature 2021, 600, 543–546.

[16] Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and "En-route" intermediates for protein folding? an investigation for small globular proteins. Journal of Molecular Biology 2000, 298 .

[17] Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. Multiple-basinenergy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America 2006, 103, 11844–9.

[18] Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Physical Review Letters 1996, 77.

[19] Zheng, W.; Doniach, S. A comparative study of motor-protein motions by using a simple elastic-network model. Proceedings of the National Academy of Sciences 2003, 100. 21.

[20] Bekker, G.-J.; Fukuda, I.; Higo, J.; Kamiya, N. Mutual population-shift driven antibody-peptide binding elucidated by molecular dynamics simulations. Scientific Reports 2020, 10, 1406.

[21] Bekker, G.-J.; Kamiya, N. Advancing the field of computational drug design using multicanonical molecular dynamics-based dynamic docking. Biophysical Reviews 2022, 14, 1349–1358.

[22] Bekker, G.-J.; Araki, M.; Oshima, K.; Okuno, Y.; Kamiya, N. Exhaustive search of the configurational space of heat-shock protein 90 with its inhibitor by multicanonical molecular dynamics based dynamic docking. Journal of Computational Chemistry 2020, 41, 1606–1615.

[23] Barua, B.; Lin, J. C.; Williams, V. D.; Kummler, P.; Neidigh, J. W.; Andersen, N. H. The Trp-cage: optimizing the stability of a globular miniprotein. Protein Engineering Design and Selection 2008, 21, 171–185.

[24] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015, 1-2, 19–25.

[25] Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins: Structure, Function, and Bioinformatics 2010, 78 .

[26] Fukuda, I.; Yonezawa, Y.; Nakamura, H. Molecular dynamics scheme for precise estimation of electrostatic interaction via zero-dipole summation principle. The Journal of Chemical Physics 2011, 134, 164107.

[27] Kamiya, N.; Fukuda, I.; Nakamura, H. Application of zero-dipole summation method to molecular dynamics simulations of a membrane protein system. Chemical Physics Letters 2013, 568-569, 26–32.

[28] Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. Journal of Computational Chemistry 2005, 26, 1668–1688.

[29] Li, W.; Wang, W.; Takada, S. Energy landscape views for interplays among folding, binding, and allostery of calmodulin domains. Proceedings of the National Academy of Sciences 2014, 111, 10550–10555.

[30] Li, W. F.; Terakawa, T.; Wang, W.; Takada, S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proceedings of the National Academy of Sciences of the United States of America 2012, 109, 17789–17794.