

## RESEARCH ARTICLE

## Development of hidden Markov modeling method for molecular orientations and structure estimation from high-speed atomic force microscopy time-series images

Tomonori Ogane<sup>1</sup>, Daisuke Noshiro<sup>2,3</sup>, Toshio Ando<sup>2</sup>, Atsuko Yamashita<sup>4</sup>, Yuji Sugita<sup>5,6,7</sup>, Yasuhiro Matsunaga<sup>1\*</sup>

**1** Graduate School of Science and Engineering, Saitama University, Saitama, Japan, **2** Nano Life Science Institute (WPI-NanoLSI), Kanazawa University, Kanazawa, Japan, **3** Institute for Genetic Medicine, Hokkaido University, Sapporo, Japan, **4** Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan, **5** Computational Biophysics Research Team, RIKEN Center for Computational Science, Kobe, Japan, **6** Laboratory for Biomolecular Function Simulation, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan, **7** Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, Saitama, Japan

\* [ymatsunaga@mail.saitama-u.ac.jp](mailto:ymatsunaga@mail.saitama-u.ac.jp)

## OPEN ACCESS

**Citation:** Ogane T, Noshiro D, Ando T, Yamashita A, Sugita Y, Matsunaga Y (2022) Development of hidden Markov modeling method for molecular orientations and structure estimation from high-speed atomic force microscopy time-series images. *PLoS Comput Biol* 18(12): e1010384. <https://doi.org/10.1371/journal.pcbi.1010384>

**Editor:** Changbong Hyeon, Korea Institute for Advanced Study, REPUBLIC OF KOREA

**Received:** July 15, 2022

**Accepted:** December 20, 2022

**Published:** December 29, 2022

**Copyright:** © 2022 Ogane et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All relevant data are within the manuscript and its [Supporting Information](https://github.com/matsunagalab/MDToolbox.jl) files. The code is available at <https://github.com/matsunagalab/MDToolbox.jl>. The scripts, input and output files are available at [https://github.com/matsunagalab/paper\\_ogane2022](https://github.com/matsunagalab/paper_ogane2022).

**Funding:** This work was supported by MEXT as “Program for Promoting Researches on the Supercomputer Fugaku” (Biomolecular dynamics

## Abstract

High-speed atomic force microscopy (HS-AFM) is a powerful technique for capturing the time-resolved behavior of biomolecules. However, structural information in HS-AFM images is limited to the surface geometry of a sample molecule. Inferring latent three-dimensional structures from the surface geometry is thus important for getting more insights into conformational dynamics of a target biomolecule. Existing methods for estimating the structures are based on the rigid-body fitting of candidate structures to each frame of HS-AFM images. Here, we extend the existing frame-by-frame rigid-body fitting analysis to multiple frames to exploit orientational correlations of a sample molecule between adjacent frames in HS-AFM data due to the interaction with the stage. In the method, we treat HS-AFM data as time-series data, and they are analyzed with the hidden Markov modeling. Using simulated HS-AFM images of the taste receptor type 1 as a test case, the proposed method shows a more robust estimation of molecular orientations than the frame-by-frame analysis. The method is applicable in integrative modeling of conformational dynamics using HS-AFM data.

## Author summary

Biomolecules dynamically change their structures in cells and perform various functions that are important for life. It is difficult to observe such dynamic structural changes using conventional experimental techniques, such as X-ray crystallography, that mainly observe static atomic structures. Alternative approach is the atomic force microscopy (AFM), a technique to measure the surface geometry of a sample molecule by scanning with an acute tip. Recently, by increasing the imaging rate of the AFM, it has become possible to observe the dynamics of biomolecules at work. However, the information measured with

in a living cell, Grant number: JPMXP1020200101 to Y.S. and Y.M.), JST CREST (Grant numbers: JPMJCR1762 to Y.M., JPMJCR13M1 to T.A.), JSPS KAKENHI (Grant numbers: 20K21380 to Y.M., 20H03195 to A.Y., 19H05645 and 21H05249 to Y.S.), and the Cooperative Research Program of "Network Joint Research Center for Materials and Devices" (Grant number: 20221300 to Y.M.). We used the computational resources provided by the HPCI system research project (Project ID: hp200135, hp210177, and hp220170) and those in RIKEN Hokusai "BigWaterFall". The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

AFM is limited to the surface geometry of a molecule, so it is important to extract three-dimensional structural information from those data. In this study, we developed a more robust method to estimate structural information from AFM data.

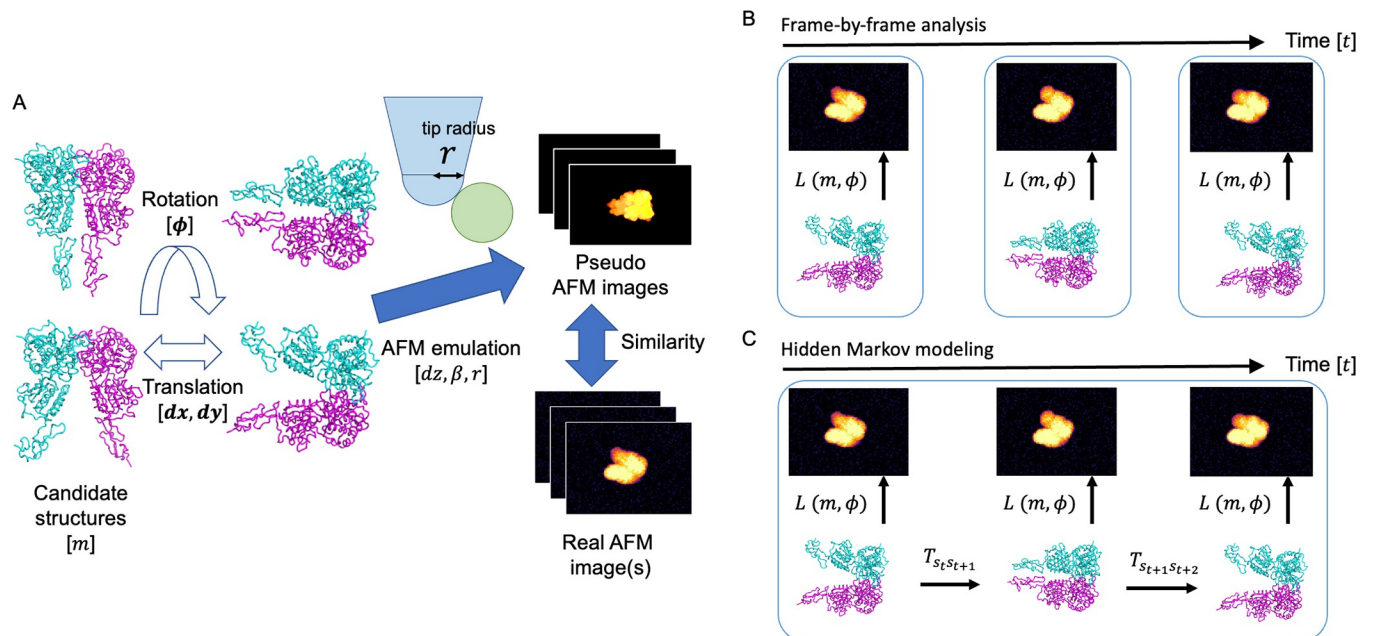
This is a *PLOS Computational Biology* Methods paper.

## Introduction

Observing conformational dynamics of biomolecules in action is important to deepen our understanding of biomolecular functions. Among the various types of experimental measurements, single-molecule measurements are powerful approaches since they can directly characterize heterogeneous fluctuations of biomolecules. The measurement techniques used in single-molecule measurements include Förster resonance energy transfer microscopies (FRET) [1,2], optical tweezers [3], and atomic force microscopy (AFM) [4]. AFM uses an acute tip as a probe for scanning the surface of a molecule and the information on the surface morphology can be obtained. In general, the imaging rate of conventional AFM is too slow to observe conformational dynamics of biomolecule in action. To overcome this limitation, the high-speed AFM (HS-AFM) has been developed by improving the resonant frequency of cantilevers, the speed of scanners [5,6]. HS-AFM enables us to observe the direct visualization of biomolecules in action at high spatiotemporal resolution. Indeed, HS-AFM has been widely applied to visualize conformational dynamics of various biomolecules; for example, myosin V walking [7], rotary catalysis of  $F_1$ -ATPase [8], the conformational dynamics of CRISPR-Cas9 in action [9], and the structure and dynamics of intrinsically disordered proteins [10].

A weak point of AFM is that the observed information is limited to the geometry on the molecular surface. Also, the obtained image profile depends on the tip shape as well as the molecular surface. Therefore, estimating the real 3D structural information hidden in the 2D AFM data is an important remaining problem. In HS-AFM and AFM observations, molecules tend to be fixed in specific orientations due to interactions with the stage, and the observed data are anisotropic. Generally, it is impossible to reconstruct structure from anisotropic data in a model-free way as is done in the single particle analysis of cryo-electron microscopy (CryoEM) in which a tremendous number of isotropic images are taken. Thus, a model-based approach is required to infer 3D structures from HS-AFM and AFM data. In this direction, the most popular model-based approach is the rigid-body fitting method [11–20]. In this method, assuming that target molecules in AFM have known structures already observed by X-ray, NMR, or CryoEM, many candidate structures are superimposed to AFM images to determine the best structure matching with the AFM image. The number of candidate structures is increased by, for example, performing molecular dynamics (MD) simulations using an experimental structure as an initial condition.

Niina et al. recently treated a candidate structural model as a rigid-body by exhaustively rotating and translating it to generate a pseudo AFM image and superimposing it on the HS-AFM image [19]. They successfully estimated the appropriate position and orientation of the molecule. Moreover, by repeating this calculation using various tip radii and half angles (see Fig 1A), these parameters were also successfully estimated. In another study [21], combined with MD simulations, a flexible fitting method was developed to search for the optimal structure by changing the structure to fit the AFM image, which enables estimation of the 3D structure even when there is no "correct" candidate structure.



**Fig 1. Scheme of analysis on atomic force microscopy images.** (A) Computation of likelihood of each atomic force microscopy image. A structure is rotated and translated, and then emulated to generate a pseudo image. The pseudo image is compared with real images with some similarity measures. (B) Schemes of the frame-by-frame rigid body fitting. (C) Schematic of hidden Markov modeling using multiple images as time-series data.

<https://doi.org/10.1371/journal.pcbi.1010384.g001>

Usually, the rigid-body fitting and flexible fitting are based on the idea that the 3D structure is independently estimated by frame-by-frame analysis of each image. On the other hand, a unique point of HS-AFM data is that they are time-series data where adjacent frames are temporally correlated with other. As mentioned above, because biomolecules interact with the stage, the orientations of the molecules tend to be fixed in a particular direction, thus orientational correlations exist even in the millisecond to second time-scale. Focusing on this unique feature of HS-AFM time series data, Fuchigami et al recently applied a time-series analysis method, called the particle filter [22] to artificially simulated AFM images. In the particle filter, a large number of copies (or replicas) are simultaneously simulated in MD, and the copies are duplicated or deleted according to the likelihood of the corresponding image. By repeating this process through the whole image frames, one can obtain a sequence of structures that well matches with time-series of AFM images. The particle filter is based on the state-space modeling. It assumes that experimental data are generated by projection from a hidden state space, using the three-dimensional structure space of a target molecule as a state space. It is theoretically guaranteed that a Bayesian inference (posterior density) of the latent space is obtained by the computational algorithm of the particle filter. The drawback of the particle filter, however, is that it is computationally very expensive. In fact, a number of simulation copies (replicas) required to accurately sample the posterior densities generally increases exponentially with the state space dimension.

This study develops a method to treat multiple HS-AFM images together as time-series data for more accurate structure and orientation estimations, while keeping computational cost lower than the particle filter. Similar to the particle filter, our idea is based on the use of the 3D structure and orientation space as a state space, but instead of a continuous space, we use a discretized space by clustering the candidate structures. By using a discretized state space model, we can apply an efficient estimation algorithm based on hidden Markov modeling [23] instead of the computationally expensive MD simulations. We combine our method with the

existing rigid-body fitting method and show that our hidden Markov modeling is a natural extension of the frame-by-frame rigid-body fitting. By performing twin experiments, we see that our method can more accurately estimate the orientation of the structure than the existing rigid-body fitting method. Furthermore, we show that our estimation method is robust even when the geometry of the tip shape in modeling differs from the correct (ground-truth) one in experiment. Finally, we discuss the applicability of the method for integrative modeling on conformational dynamics to real experimental HS-AFM data.

## Methods

This study proposes a method to analyze multiple AFM images as a time series by extending the conventional rigid-body fitting method. First, we review the conventional rigid-body fitting, in which each AFM image is analyzed by a frame-by-frame manner, and then we describe our hidden Markov approach in the next subsection. The conventional rigid-body fitting estimates the parameters (structures, orientations, translations, and others) of the maximum likelihood for each image. As like the classical method, the hidden Markov method outputs the maximum likelihood parameters where the likelihood is not computed independently for each frame, but computed for the whole frames considering them as a time series.

### Frame-by-frame rigid-body fitting

In the rigid-body fitting, an ensemble of candidate structures is prepared from experimentally determined structures or structures obtained by MD simulations. The goal of the rigid-body fitting is to select the correct structure and to find its orientation and translation from a given experimental AFM image. To accomplish this, the rigid-body fitting generates a large number of synthetic AFM images with various structures  $m$ , 3D rotations  $\phi$  (quaternions are used in this study), translations  $(dx, dy)$ , and other parameters, and compares them to the experimental AFM images. The parameter set that produces the synthetic AFM image with the highest similarity to the experimental AFM image is considered as the best estimate. Hereafter, we will refer to these synthetic images as pseudo-AFM images.

To generate pseudo-AFM images, we employed a conventional collision-detection method [11,18,19]. This method calculates the height at which a tip perpendicular to the stage collides with a sample molecule. Let the stage be the  $xy$ -plane, then the height of the stage ( $z$ -coordinate) is aligned with the minimum value of the  $z$ -coordinate of the atoms of the molecule. Then, let us describe by  $dz$  an offset of the stage from the origin as an unknown parameter. The scaling parameter  $\beta$  for height  $z$  is also considered as unknown, if the calibration of the height scale is not perfect. Following Niina et al. [19], we approximate the tip by a hemisphere (with radius  $r$ ) combined with a circular frustum of a cone. In summary, unknown parameters when generating pseudo-AFM images are structure  $m$ , rotation  $\phi$ , and translation  $(dx, dy)$ , offset  $dz$ , scaling parameter  $\beta$  to calibrate the height, and apex radius  $r$  of the tip.

The rigid-body fitting measures similarity between the generated pseudo-AFM image and the experimental AFM image. In this approach, the parameters used to generate the pseudo-AFM image with the highest similarity are considered as the best estimate. The similarity measures include pixel-RMSD [22], cross correlation [18,20], cosine similarity [19], and structural similarity index measure (SSIM) [17]. For example, pixel-RMSD is defined as follows, essentially the same as peak signal-to-noise ratio (PSNR):

$$\text{RMSD}(I^{\text{ref}}, I^{\text{pseudo}}(m, \phi, dx, dy, dz, \beta, r)) = \sqrt{\frac{1}{N_{\text{pixel}}} \sum_i^{N_{\text{pixel}}} (I_i^{\text{ref}} - I_i^{\text{pseudo}}(m, \phi, dx, dy, dz, \beta, r))^2}$$

where  $I_i^{ref}$  and  $I_i^{pseudo}$  are the heights at pixel  $i$  of the experimental AFM and the pseudo-AFM images, respectively.  $N_{pixel}$  is the number of pixels in the image. In this case, a smaller pixel-RMSD means a higher similarity.

In general, the calibration for height is not always perfect in AFM data and thus the scaling parameter  $\beta$  for height and the stage offset  $dz$  should be treated carefully. In fact, Müller and Engel showed that ion concentration and the affected Debye length lead to different AFM measure heights in biomolecules [24]. Due to this problem, conventional rigid-body fitting studies use cross correlation or cosine similarity as the similarity measure between images. These measures are invariant to the scaling or translations by height. However, these measures are incompatible with the probabilistic modeling of images, and thus cannot be directly used for the likelihood analysis in this study. Considering this context, in the following, we introduce a measure developed by Cossio and Hummer in their analysis of CryoEM images [25], that is invariant to height scaling and translations, to the analysis of AFM images, which enabled us to perform probabilistic modeling of images and extend the conventional rigid-body fitting to multiple frames.

First, we introduce the concept of probability into the rigid-body fitting to extend the analysis to a likelihood-based analysis that allows more flexible modeling. We assume that an experimental AFM image can be generated by adding spatially independent Gaussian noise to a pseudo-AFM image. Then, the probability of observing the experimental AFM image or likelihood, given a set of parameters, can be written as follows

$$L(m, \phi, dx, dy, dz, \beta, r, \lambda) = p(I^{ref} | m, \phi, dx, dy, dz, \beta, r, \lambda) = \prod_i^{N_{pixel}} (2\pi\lambda)^{-1/2} e^{-\frac{|I_i^{ref} - I_i^{pseudo}(m, \phi, dx, dy, dz, \beta, r)|^2}{2\lambda^2}}, \quad (1)$$

where  $\lambda$  is the standard deviation of the Gaussian noise. From the above equation, it is obvious that this likelihood  $L$  is maximized for the parameter set with the smallest pixel-RMSD. In principle, the maximum likelihood parameters can be found by exhaustively searching possible parameters, but the large number of possible combinations of the parameters makes it infeasible to search all of them. Therefore, in this study, we reduce the number of parameters by integrating out some of unknown parameters, as was done by Cossio and Hummer in their CryoEM data analysis [25].

Starting from the same type of equation as Eq 1, Cossio and Hummer analytically integrated out the parameters,  $dz$ ,  $\beta$ , and  $\lambda$ . The derivation is very general and thus can be applied directly to AFM data, not just CryoEM. The Gaussian integral can be applied to the integration over  $\beta$  and  $dz$ . The integration over  $\lambda$  needs some approximation (a saddle-point-type approximation). Then, the marginalized likelihood becomes as follows,

$$L(m, \phi, dx, dy, r) \approx \sqrt{\pi}(2\pi e)^{1-N_{pixel}/2} \times [N_{pixel}(C_{cc}C_{oo} - C_{oc}^2) + 2C_oC_{oc}C_c - C_{cc}C_o^2 - C_{oo}C_c^2]^{3/2-N_{pixel}/2} \times [(N_{pixel} - 2)(N_{pixel}C_{cc} - C_c^2)]^{\frac{N_{pixel}}{2} - 2}. \quad (2)$$

Here,  $N_{pixel}$  is the total number of pixels in the image.  $C_o$ ,  $C_c$ ,  $C_{oo}$ ,  $C_{cc}$ ,  $C_{oc}$  are sums of pixel heights, sums of squares, and inner product of pixel heights, respectively:

$$C_o = \sum_i I_i^{ref},$$

$$C_c = \sum_i I_i^{pseudo},$$

$$C_{oo} = \sum_i (I_i^{\text{ref}})^2,$$

$$C_{cc} = \sum_i (I_i^{\text{pseudo}})^2,$$

$$C_{oc} = \sum_i I_i^{\text{ref}} I_i^{\text{pseudo}}.$$

This marginalized likelihood depends only on the candidate structure  $m$ , 3D rotation  $\phi$ , translation  $(dx, dy)$  and tip apex radius  $r$ , and now this is more feasible compared to Eq 1.

In this study, using the marginalized likelihood (Eq 2), we conducted maximum likelihood estimates for the candidate structure  $m$ , 3D rotation  $\phi$ , and tip apex radius  $r$ . For the integration over the 3D rotation  $\phi$ , we used 577 quaternions which are uniformly sampled in the SO (3) group space [26,27].  $C_{oo}$ ,  $C_c$ ,  $C_{oo}$  and  $C_{cc}$  are invariant with respect to translation  $(dx, dy)$ , thus only  $C_{oc}$  is needed to integrate over  $(dx, dy)$ .  $C_{oc}$  can be expressed in the form of convolution integrals, thus can be efficiently computed in the inverse space using Fast Fourier Transform (FFT) [27]. In this study, we employed FFT and the computational cost was reduced from  $O(N_{\text{pixel}}^3)$  to  $O(N_{\text{pixel}}^2 \log N_{\text{pixel}})$ . As for the translation, we just took the slice of the maximum point instead of marginalization. This is important for analyzing AFM movies which contain particles other than the target molecule because marginalization is affected by non-targeted particles.

### Hidden Markov modeling extension to multiple frames

To extend the rigid-body fitting described in the previous subsection to treat multiple AFM images as a time-series, we introduce a likelihood for multiple images consisting of temporally consecutive frames (Fig 1C). To define this, we first define state  $s$  ( $s = 1, \dots, M$ ) by the combination of the candidate structure  $m$  and the 3D rotation  $\phi$ , and then introduce the transition probability from state  $s_t$  at the  $t$ -th frame ( $t = 1, \dots, T$ ) to state  $s_{t+1}$  at the  $(t+1)$ -th frame as  $T_{s_t s_{t+1}}$ . We assume that the transitions between states satisfy the Markovian property so that the transitions can be fully characterized by only the transition probabilities. Furthermore, the transition probabilities are assumed to satisfy the detailed balance. Translations  $(dx, dy)$  are not included in the definition of the state because, unlike rotation, translations may not affect the conformational dynamics of a molecule because interactions with the stage do not change over translations. Then, the probability of being in state  $s_1$  at the 1st frame and observing image  $I^{\text{ref},1}$  can be written by  $p(s_1)L(m_1, \phi_1, dx_1, dy_1, r)$  where  $m_1, \phi_1, dx_1, dy_1$  are the parameter sets at the 1st frame,  $p(s_1)$  is the equilibrium probability of  $s_1$ . Repeating this, the probability of observing state  $s_2$  and image  $I^{\text{ref},2}$  in the 2nd frame is written by  $p(s_1)L(m_1, \phi_1, dx_1, dy_1, r)T_{s_1 s_2}L(m_2, \phi_2, dx_2, dy_2, r)$  using the transition probability. By repeating this process up to the  $T$ th frame, and marginalizing over the state  $s_t$  for all the frames, the likelihood of AFM time-series images becomes as follows,

$$L(m_1, \phi_1, dx_1, dy_1, \dots, m_T, \phi_T, dx_T, dy_T, r) = \sum_{s_1=1}^M \dots \sum_{s_T=1}^M p(s_1)L(m_1, \phi_1, dx_1, dy_1, r) \prod_{t=2}^T T_{s_{t-1} s_t} L(m_t, \phi_t, dx_t, dy_t, r). \quad (3)$$

Note that if the equilibrium probabilities and transition probabilities are uniform, the likelihood in Eq 3 is equivalent to simply multiplying the likelihoods of the frame-by-frame rigid-

body fitting. On the other hand, if the transition probabilities are not uniform, the likelihood is affected by not only the individual likelihood of rigid-body fitting, but also the transition probabilities.

Since the transition probabilities  $T_{s_t s_{t+1}}$  are usually unknown in advance of the analysis of AFM images, they should be estimated from the data using the Baum-Welch algorithm of hidden Markov modeling [23]. In the original Baum-Welch algorithm, the so-called *emission*, which is the probability of observing an image from a given state, is estimated together from the data. Since the emission is exactly the same as the likelihood  $L(m, \phi, dx, dy, r)$  of the frame-by-frame rigid-body fitting in this case, only the transition probabilities are estimated here. The Baum-Welch algorithm can estimate transition probabilities that reproduce the data with higher likelihood defined in Eq 3 (though it only locally optimizes the likelihood thus it could be trapped a local minimum). In the algorithm, the Forward-Backward algorithm is used for the calculation of the likelihood, which drastically reduces the computational cost from  $O(2TM^T)$  to  $O(2TM^2)$ .

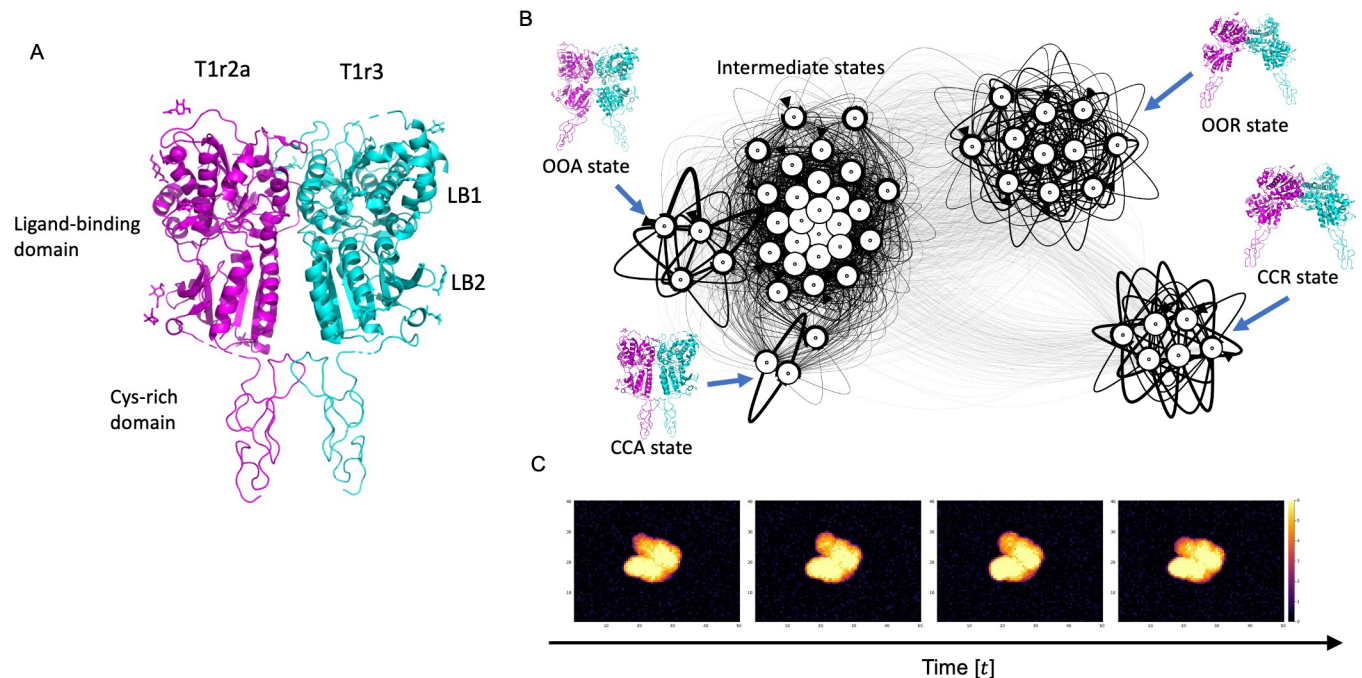
An important characteristic of the Baum-Welch algorithm is that the transition probability, initially sets to zero, remains zero even after the algorithm is applied [23]. By imposing these topological constraints as initial conditions, it is possible to introduce a prior knowledge that transitions between certain states completely unlikely happen in a single transition and further reduce computational cost [28]. In the case of AFM experiments, a target molecule is often interacting with the stage, and large rotations between adjacent frames do not likely happen. Thus, we propose to impose a prior knowledge that significantly change its orientations unlikely occur in a single transition. This can be achieved by imposing the transition probabilities to be zero between states with largely different orientations. Throughout this study, we imposed the probability of transitions between states whose quaternion distance is  $\|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\|^2 = 1 - \sum_{i=1}^4 (q_i^{(1)} - q_i^{(2)})^2 > 0.1$  to be zero.

In the calculation of the likelihood of AFM time-series images (Eq 3), we first need to compute the likelihood  $L(m_b, \phi_b, dx_b, dy_b, r)$  of the frame-by-frame rigid-body fitting (Eq 2). The results of the frame-by-frame likelihood can be used to further reduce the number of states used in Eq 3, because the structures and orientations with very small frame-by-frame likelihood (Eq 2) value are also very small even in the likelihood of the time-series (Eq 3). By removing the structures and orientations with very small frame-by-frame likelihoods, we can further reduce computational cost of the hidden Markov modeling.

Once the transition probabilities are estimated by the Baum-Welch algorithm, they can be used to estimate the maximum likelihood states that match the AFM time-series images. This maximum likelihood estimation can be done efficiently with the Viterbi algorithm in the Hidden Markov modeling [23]. The Viterbi algorithm find the maximum likelihood state sequences from the given AFM data by maximizing the likelihood of AFM time-series images over the parameters (sequence of structures, orientations,  $m_1, \phi_1, \dots, m_T, \phi_T$ ). In this study, the tip radius  $r$  is fixed during the analysis instead of optimizing it in order to further reduce the computational cost. In summary, by applying the hidden Markov modeling to AFM time-series images, (i) we first obtain the transition probabilities between states by the Baum-Welch algorithm, (ii) then we obtain the maximum likelihood sequence of structures, rotations with by the Viterbi algorithm.

## Twin experiment

To compare the performance of the Hidden Markov modeling method described in the previous subsection with the frame-by-frame rigid-body fitting, we performed twin experiments. In the experiments, the accuracy of the method is assessed by generating pseudo-AFM images



**Fig 2. Structure of the taste receptor and its Markov state model (MSM).** (A) The extracellular ligand-binding domains and the cysteine-rich domains of the medaka fish taste receptor. (B) MSM constructed from coarse-grained simulation data. (C) Pseudo-AFM data generated by a stochastic simulation with the MSM, which is used for the twin experiments.

<https://doi.org/10.1371/journal.pcbi.1010384.g002>

and compare the estimated structures and orientations with the ground-truths used to generate the images.

We used the extracellular ligand-binding domains (LBDs) of the medaka fish taste receptor (PDB ID: 5X2M [29]) as a molecular model in the twin experiments (Fig 2A). This is a heterodimer consisting of T1r2a and T1r3. The ligand-binding domains (LBDs) in T1r2a and T1r3 have a ligand-binding pocket. Previous biophysical studies have shown that the T1r2aLBD pocket is characterized by a broad yet discriminating chemical recognition, while the T1r3LBD pocket is rather loosely bound an amino acid [29]. The receptor is an ortholog of human sweet and umami taste receptors, in which each subunit of the heterodimer is known to play different functional roles [30]. In this context, it is important to correctly identify T1r2aLBD and T1r3LBD of this pseudo symmetric structure from AFM images and to characterize the structure and dynamics of each domain.

To prepare molecular structures used for generating pseudo-AFM images for twin experiment, we performed coarse-grained MD simulations of the taste receptor (without the stage of AFM), sampling all possible structures, and then constructed a Markov state model (MSM) [31–33]. Before starting MD simulations, to clarify the dynamics of the entire extracellular region not only with LBD but also with the downstream cysteine-rich domain, cysteine-rich domains were added to the atomistic structure of 5X2M by homology modeling using the structure of the same protein family with T1rs, class C G protein-coupled receptor (GPCR), the active form of human calcium-sensing receptor extracellular domain (PDB ID: 5K5S [34]) as a template. Then, the subdomains (LB1 and LB2) of LBDs were superimposed on the structure of the inactive form of human calcium-sensing receptor extracellular domain (PDB ID: 5K5T [34]) to create a structure where LBDs are open (called the OO state while the closed state is called the CC state). Also, cysteine-rich domain was also superimposed on the structure

of a single protomer of another class C GPCR, the metabotropic receptor (mGluR)3LBD (PDB ID: 2E4U [35]), to create an open structure of the whole dimer (called the R state while the closed state is called the A state). Simple superposition of the LBDs made the interface region of the LBDs sterically crashed, so the structure was superimposed again to the crashed structure by a targeted MD simulation [36] using an all-atom model from the 5X2M structure. In the targeted MD simulation, we did not impose any pulling forces to the interface region to make the regions relaxed according to naturally occurring inter-molecular interactions. Similar calculations were performed for structures with open LBDs (OO state). All obtained structures were then subjected to all-atom MD with positional restraints on their C $\alpha$  atoms to relax the other regions including side-chains and surrounding solvents. Then, using the obtained structures, C $\alpha$ -based Karanicolas-Brooks (KS) Go-model [37,38] was created for coarse-grained MD simulations. Based on KB Go-model, we created a quad-basin potential energy by macro-mixing the four states (CCA, CCR, OOR, OOA states, see Fig 2B) obtained by the all-atom modeling, which makes its energy landscape stabilizing these four states. Then, a number of coarse-grained MD simulations were conducted with the created model to sample all possible structures.

All the MD simulations were performed with GENESIS [39,40]. In the all-atom MD simulations, the simulation systems were prepared with the CHARMM-GUI web server [41] using the modeled structures. Missing residues were modelled with MODELLER [42]. CHARMM36 force field [43,44] was used with the TIP3P water model [45]. Electrostatic interactions were treated using the Smooth Particle Mesh Ewald method [46], and covalent bonds including hydrogen atoms were constrained using SHAKE [47] and SETTLE [48]. Temperature was kept at 303.15 K with Langevin dynamics. In the coarse-grained MD simulations, the parameter sets for KB Go-model were prepared by using the MMTSB [49] Go model web server. Quad-basin macro-mixing parameters were created from four single-basin KB Go-model parameters. All the bonds between coarse-grained atoms were constrained using SHAKE, and the temperature was controlled at 200 K by Langevin dynamics.

The obtained MD structures were aligned to the CCA structure, and principal component analysis (PCA [50]) was performed on the Cartesian coordinates of the aligned structures. The 1st PC corresponds to opening-closing structural motions in the entire dimer, and the 2nd PC corresponds to opening-closing motions in LB1 and LB2 of LBDs. We performed the  $k$ -means clustering in the space of these 1st and 2nd PCs to define the states in the Markov state model (MSM) with  $k = 50$ . Note that, “states” used here, contain only structures (does not contain rotations). By counting the transition between these states from the trajectories, transition probabilities for states (structures) were estimated. Finally, we constructed a MSM of the taste receptor to represent its conformational dynamics.

We used the constructed MSM to generate pseudo-AFM data for twin experiment (see S1 Fig for the computational protocol). To generate pseudo-AFM images, (i) we first conducted stochastic simulations with the MSM to calculate the transitions between the MSM’s discretized states for 100 frames. In the simulation, we first chose one state from the MSM’s discretized fifty states as the initial state according to the equilibrium probabilities. Then, calculate next state according to the transition probabilities of the MSM. By repeating this process, we obtained 100 state sequences. According to the obtained sequence of states, the sequence of corresponding structures was drawn from the centroid of  $k$ -means clustering. Then, (ii) 100 frames of AFM images were created from the sequence of structures using a collision-detection method [19]. We assumed the tip by a hemisphere (with the radius of  $r = 2.5$  nm and the half angle of 10 degrees, see S1 Fig for the definition of the half angle) combined with a circular frustum of a cone. When calculating the collision with a tip, the effective radius of the C $\alpha$  atom for its amino acid residue was taken from the CryoEM study [25]. The pixel resolution is 0.625

nm  $\times$  0.625 nm, and the image has 80 pixels along the X-axis and 60 pixels along the Y-axis. For simplicity, we assumed that a molecule is strongly interacted with the stage, so the orientation of the molecule is fixed to a specific angle during the 100 frames. To determine the orientation, a quaternion was randomly selected from 576 quaternions uniformly sampled in SO(3) group [26,27]. Irrespective the chosen orientations, the same transition probabilities between states (structures) were used for the stochastic simulation. Lateral drift of the molecule was simulated as a random walk where each step was sampled by 2D independent Gaussian distribution with the standard deviation of (0.1 nm). (iii) Then, to mimic experimental noise, spatially independent Gaussian noise with a standard deviation of 0.3 nm were added to the generated pseudo-AFM images. The standard deviation of 0.3 nm is a typical noise width of AFM images [19]. The above calculations were repeated 50 times to produce a total of 50 sets of 100-frame AFM data.

We performed the frame-by-frame rigid-body fitting and the maximum likelihood estimation with hidden Markov modeling on the pseudo-AFM images created by the above procedure. We used the same 50 structures with the MSM (i.e., the centroids of  $k$ -means clustering) as candidate structures. The same 576 quaternions and AFM image grids were used. RMSD between the ground-truth and estimated structures was used to access the accuracy of the estimation. In order to access the accuracy of not only the structure  $m$  but also the 3D rotation  $\phi$ , the RMSD calculation was performed without applying structural alignment.

It is noted again that the same 50 structure with the MSM were used for the analysis. This is mainly due to the requirement that the MSM clusters should be the same as the ground-truth in order to check the accuracy of the reconstructed MSM (relative entropy calculation, described later) in the latter part of our analysis. For checking the effect of this choice, we conducted an additional analysis and calculated RMSDs calculation using different 50 structures obtained by a clustering of MD simulation trajectories.

The actual experimental data often suggest that the shape of a tip is not simple as a circular frustum of a cone assumed in this study, but a more complex one. Therefore, to account for the possibility that the shape of the real tip might be different from the assumed one, we performed maximum likelihood estimations with intentionally different tip apex radii (ground truth used for generating pseudo-AFM images are 2.5 nm while we used 1.5 nm, 1.8 nm, 2.5 nm, 3.2 nm, and 3.5 nm for estimation). We tested how robustly the method estimates the structure and orientation in the situation where there was a discrepancy in the tip apex radius. Furthermore, the hidden Markov modeling can estimate the transition probabilities  $T_{s_t s_{t+1}}$  between states. We reconstructed the transition probabilities between structures by reducing orientations from  $T_{s_t s_{t+1}}$  obtained by the Baum-Welch algorithm. Then, compared the reduced transition probabilities between structures with those of the Markov state model.

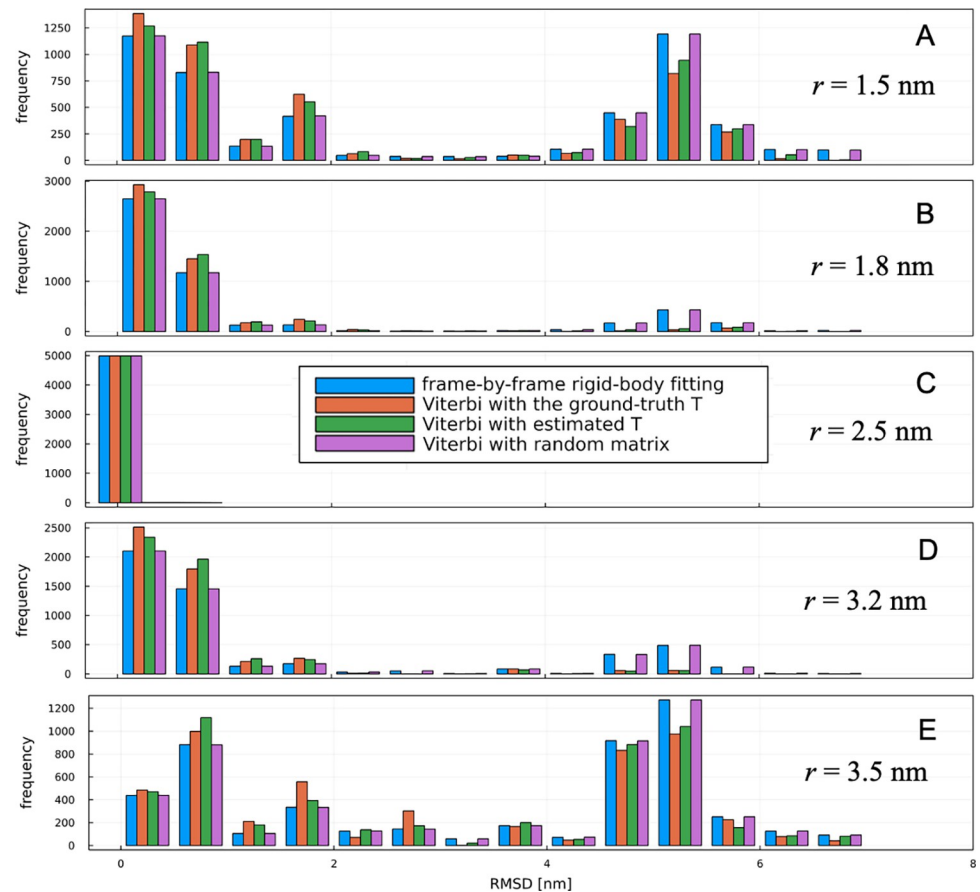
## Software availability

The Hidden Markov modeling-based analysis method described here is implemented and publicly available as part of the MDTtoolbox.jl package and can be downloaded from <https://github.com/matsunagalab/MDTtoolbox.jl>. In addition, Jupyter notebooks to reproduce the results of this paper are publicly available at [https://github.com/matsunagalab/paper\\_ogane2022](https://github.com/matsunagalab/paper_ogane2022).

## Results and discussion

### Twin experiments with identical radii

First, we examined the estimation accuracy when the ground-truth tip radius used to generate the pseudo-AFM data (2.5 nm) and the tip radius used for the maximum likelihood estimation



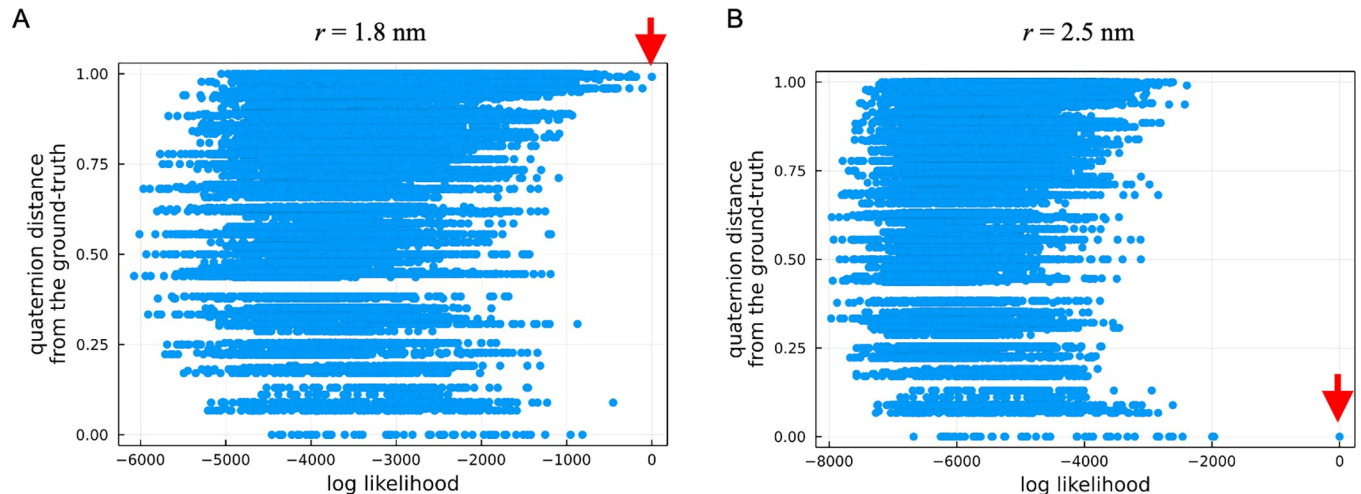
**Fig 3. Results of twin experiments.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).

<https://doi.org/10.1371/journal.pcbi.1010384.g003>

are identical. Fig 3C shows that both methods (the frame-by-frame rigid-body fitting and the hidden Markov modeling) are able to estimate the correct ones from the combinations of 50 candidate structures and 576 angles with almost 100% accuracy, despite the addition of Gaussian noise with a standard deviation of 0.3 nm. Specifically, the frame-by-frame rigid body fitting resulted in zero RMSD (without structural alignment) for 4991 frames out of 50 frames  $\times$  100 sets of images. The Hidden Markov modeling resulted in zero RMSD in the same number of frames (4991 frames). The accuracy of this estimation result may be dependent on the noise size; in the current setup, the spatial scale of the average RMSD ( $\sim 1.1$  nm) between the 50 structures and the resolution of the rotations is comparable or larger than the size of Gaussian noise.

### Twin experiments with discrepant radii

Next, we examined the estimation accuracy when the ground-truth tip radius (2.5 nm) used in generating the AFM data is different from the tip radius used for the maximum likelihood estimation. As described in Methods, the actual experimental probe geometry is more complex than the assumed shape, so we intentionally used a different radius for the estimation to assess how the estimations are robust against such discrepancies.



**Fig 4. Likelihood distribution over all possible orientations and structures computed using the first single frame of the first set AFM data (defined by Eq 2).** Likelihoods are projected onto the quaternion distance from the ground-truth. The maximum likelihoods are indicated by red arrows. (A) Likelihoods computed with the tip apex radius of 1.8 nm. (B) 2.5 nm (the ground-truth tip radius).

<https://doi.org/10.1371/journal.pcbi.1010384.g004>

Fig 3A and 3B and 3D and 3E show the accuracies of the frame-by-frame rigid-body fitting with different tip radii, measured by structural RMSD from the ground-truth structure. When the radius difference is close, the RMSDs are distributed lower than 1.0 nm. However, as the radius difference increases, higher peaks appear at around the RMSD of 5.0 nm, indicating that the estimation accuracy is becoming worse. The structures around the RMSD of 5.0 nm are flipped 180 degrees from the ground-truth structure, indicating that the estimations failed to assign each monomer of the taste receptor to the AFM image correctly. This is because, as the radius difference increases, the likelihood fails to catch the subtle surface differences in the two monomers.

To further investigate the cause of the double peaks in the RMSD distribution, we examined the relation between the orientations and the likelihood values (Eq 2) using the first frame of the first set in the AFM data (shown in Fig 4). Fig 4 shows the likelihood values for each 3D rotation  $\phi$  and candidate structures  $m$ . Fig 4B shows that, even when the radius of the tip apex is identical (2.5 nm), the likelihood of the structures flipped about 180 degrees have already higher likelihoods (compared to other angles), but the tip distinguishes the difference in surface geometry of each monomer and the correct angle is chosen as the maximum likelihood. On the other hand, when the radii are different from each other (Fig 4A), the likelihood at the ground-truth angle decreases whereas the likelihoods of the 180-degree flipped structures become competitive, resulting in the flipped structures to be selected as the maximum likelihood.

Next, we compared the results of the frame-by-frame rigid-body fitting with those of the hidden Markov modeling (Fig 3). We first used the ground-truth transition probabilities of MSM used for generating pseudo-AFM data. A sequence of maximum likelihood states was computed by the Viterbi algorithm. Fig 3 shows that the RMSDs of the hidden Markov modeling with the ground-truth transition matrix are overall improved compared to the RMSDs of the frame-by-frame rigid-body fitting. In particular, when the probe radius is 1.8 nm, there is a noticeable reduction in the peak around RMSD of 5.0 nm corresponding to 180-degree flipped structures. This means that hidden Markov modeling enables a robust correct estimation of the orientation of the molecule against the choice of tip radius. As described in Eq 3, the hidden Markov uses all frames to find the maximum likelihood sequence, not just one

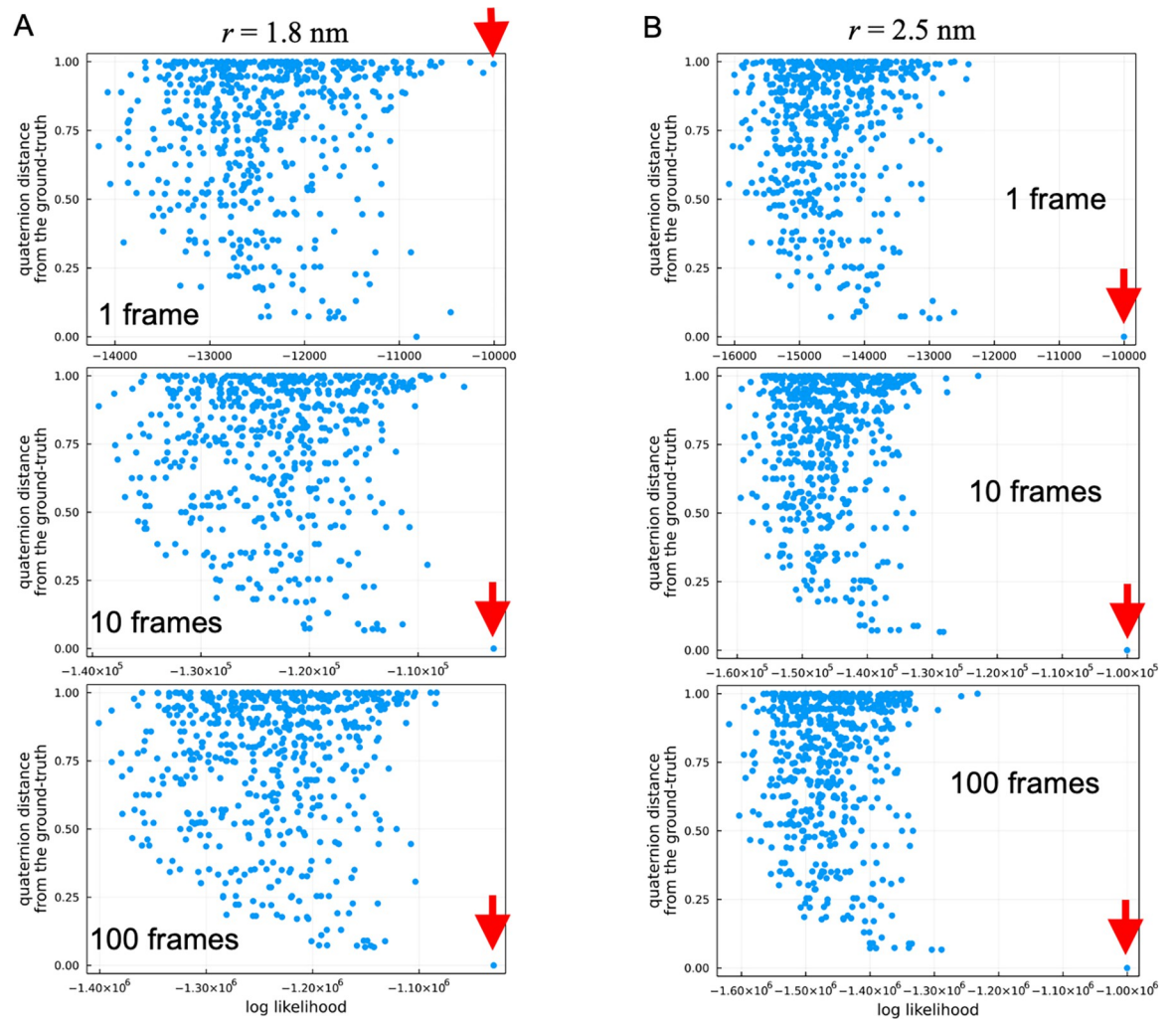
frame. The use of a transition matrix with rotational constraints gives penalties for large rotational flips. As a result, a fixed orientation throughout all frames is favored as the maximum likelihood sequence by a majority rule; minor orientations are not selected. This mechanism can be considered as the same principle as in the so-called ensemble learning [51]; the variance of estimation becomes large when a single discriminator is used, while the variance becomes small when multiple discriminators are combined. Note, however, when the bias getting larger due to large inconsistencies between tip radii, the majority rule cannot decrease the error as shown in Fig 3A (1.5 nm) and 3E (3.5 nm).

In order to check the sensitivity of the results against noise realizations (Gaussian noise added to the pseudo-AFM images), we conducted the same analyses using different noise realizations (S2 and S3 Figs). The same trend was observed regardless of the noise realization values. Furthermore, to check the sensitivity of the results against noise levels, we conducted the same analyses using the pseudo-AFM images with different noise levels (the standard deviations of Gaussian noise, 0.6, 0.9, and 1.2 nm, shown in S4–S6 Figs). Interestingly, we did not see a significant decrease in RMSDs with these noise levels. Although this may be due in part to the still small noise levels, this result suggests that the tip radius (shown in Fig 3) has a stronger influence on AFM image analysis than the noise level. Moreover, to check the sensitivity of the results against tip radius used for generating pseudo-AFM image, we conducted the same analysis with the tip radius of 3.5 nm, which shows the same tendency as that of the tip radius of 2.5 nm (S7 Fig).

S8 Fig shows the RMSDs from the ground-truth structures calculated using 50 different structures different from those used for generating the pseudo-AFM structures. This corresponds to a more general situation for analyzing HS-AFM data. Although the accuracies of estimations have deteriorated by using the different structures, the overall story that hidden Markov modeling improves estimation accuracy remains unchanged.

To investigate how the majority rule mechanism works in multiple frames, we again examined the relation between orientations and likelihood values, changing the numbers of frames. By imposing a constraint being at a specific 3D angle  $\phi$  in the Viterbi algorithm, we computed the maximum likelihood at each orientation using 1 frame, 10 frames, and 100 frames of the first set of the AFM data. Fig 5B shows the maximum likelihoods at each orientation when both radii are identical (2.5 nm). In this case, the correct orientation has already been selected as the maximum likelihood of the frame-by-frame analysis. As the number of frames increases, the difference between the largest likelihood of the correct orientation and the second largest one increases, indicating that the correct orientation is more likely chosen by using a greater number of frames. Fig 5A shows the results when the probe radius is not identical with the ground truth (2.5 nm and 1.8 nm). When only single frame is used, the maximum likelihood orientation is not correct, a 180 degrees flipped structure is chosen. Interestingly, as the number of frames increases, the relationship switches and the likelihood of the correct orientation begins to be chosen because the other frames have higher likelihoods for the correct orientation. As the number of frames increases, the likelihood of the correct orientation becomes more prominent. In summary, when the likelihood is calculated for only a single frame, it is probable that a wrong orientation is chosen due to the estimation error (variance), but when the number of frames is increased, the correct orientation is robustly chosen by decreasing the variance.

To verify further whether the transition probabilities work to improve estimation accuracy, we computed the maximum likelihood sequence using a random matrix generated by uniform random numbers (normalized over rows) as a transition probability matrix. Interestingly, the estimation accuracy of hidden Markov modeling using the random matrix is almost identical to the results of the frame-by-frame rigid-body fitting. This is due to the fact that the



**Fig 5. Maximum likelihood of pseudo-AFM time-series images in each orientation computed using the first single frame, the first 10 frames, the first 100 frames of the first set AFM data (defined by Eq 3).** Note that only the maximum likelihood for each orientation is shown due to computational costs (calculated by the Viterbi algorithm). The maximum likelihoods among all orientations are indicated by red arrows. (A) Likelihoods computed with the tip apex radius of 1.8 nm. (B) 2.5 nm (the ground-truth).

<https://doi.org/10.1371/journal.pcbi.1010384.g005>

mechanism of the ensemble learning does not work because the transition probabilities can no longer give penalty for large rotational flips. As a result, the likelihood of each frame is independently selected in the maximum likelihood sequence.

Finally, we performed estimations where the transition probabilities were estimated from the data without providing ground-truth transition probabilities in advance. Starting from a random matrix as the initial condition and imposing a rotational constraint, we estimated transition probabilities from the data using the Baum-Welch algorithm, and subsequently obtained the maximum likelihood sequence to verify the accuracy of the estimation. As the constraint on rotation, the probability of transitions between states whose quaternion distance is  $\|\mathbf{q}^{(1)} - \mathbf{q}^{(2)}\|^2 = 1 - \sum_{i=1}^4 (q_i^{(1)} - q_i^{(2)})^2 > 0.1$  was imposed to be zero. Fig 3 shows that the estimation accuracy is improved compared with those of the rigid-body fitting, even when the transition probabilities are estimated from the data, although they are not as good as those with the ground-truth transition probabilities. This improvement in accuracy is mainly contributed by rotational constraints on the transition probabilities.

The confidence intervals of the estimated transition probabilities were evaluated by the bootstrap method [52] for the radii of 2.5 and 2.0 nm (S11 and S12 Figs). The results indicate that probabilities can be estimated in the range of 0.1 to 0.2 from the current data set.

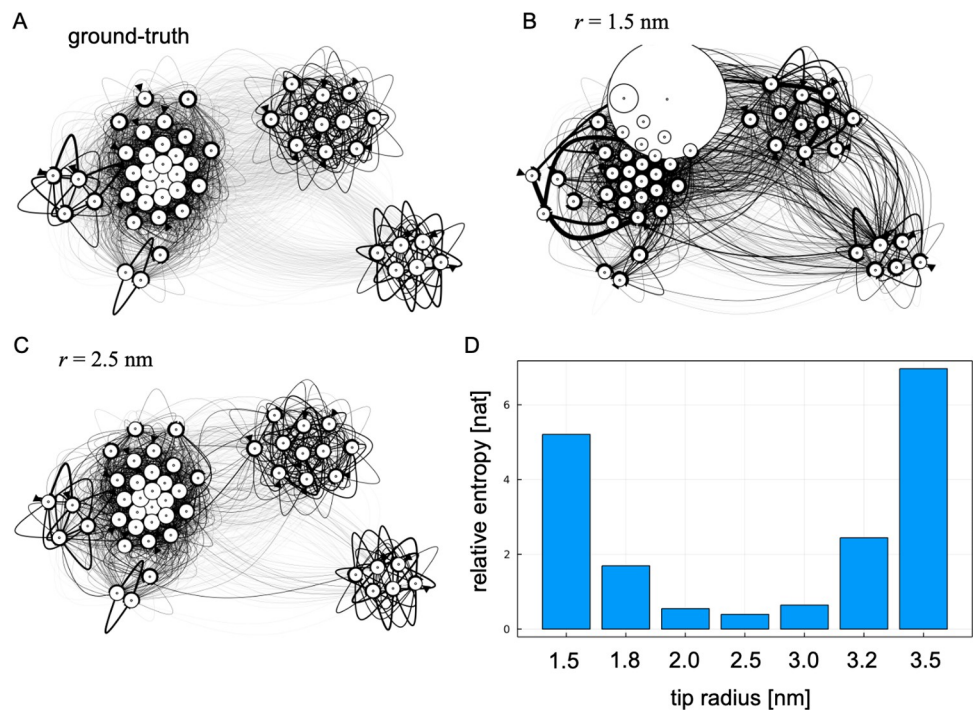
### Reconstruction of Markov state model

As calculated in the previous subsection, Hidden Markov modeling can estimate transition probabilities from the data using the Baum-Welch algorithm. Estimating transition probabilities and constructing more accurate MSM from experimental data is an important topic for application studies. Here, we examined how accurately our proposed method can estimate the transition probabilities between structures using the same twin experiment data. Again, using a random matrix as the initial condition and imposing a rotation constraint, we estimated transition probabilities from the data using the Baum-Welch algorithm. Then, we marginalized the estimated transition probabilities between states (structures and rotations) to the transition probabilities between structures. After the marginalization, the probabilities were post-processed to suffice the detailed balance. Fig 6 visualizes the transition probabilities between states estimated with various tip radii. The equilibrium probability of each structure (indicated by the area of each node in Fig 6) computed from the detailed balance.

To access the accuracy of the reconstructed Markov state model, we used the relative entropy between two models [53],

$$D(T||T^{\text{est}}) = \sum_{i=1}^M \sum_{j=1}^M p(i) T_{ij} \log \frac{T_{ij}}{T_{ij}^{\text{est}}},$$

where  $T_{ij}$  is the transition probability of the ground-truth model, and  $p(i)$  is its equilibrium



**Fig 6. Reconstruction of MSM from atomic force microscopy data.** (A) Graph representation of the ground-truth MSM used for twin experiments. (B) Reconstructed Markov state model with the tip apex radius of 1.5 nm. (C) Reconstructed MSM with the tip apex radius of 2.5 nm (the ground-truth radius). (D) Accuracies of reconstructed msms evaluated by relative entropies for various tip apex radii.

<https://doi.org/10.1371/journal.pcbi.1010384.g006>

probability of state  $i$ .  $T_{ij}^{\text{est}}$  is the transition probability of a reconstructed model. The relative entropy is zero if the two models are identical and takes increasingly large values the more the two models differ. To avoid singular behavior due to zero divide, we here added a very small value ( $10^{-10}$ ) uniformly to  $T_{ij}^{\text{est}}$ .

Fig 6 shows that the estimated transition probabilities are almost the same, especially when the probe radii used for estimation are ground truth and identical (2.5 nm). In fact, as in the ground-truth MSM, CCR state is kinetically isolated from other states and the transition to this state is rare. This is true even when the probe radius is 2.0 nm, indicating that our proposed method is robust to the discrepancies in the probe geometry. According to the frame-by-frame rigid-body fitting of real AFM data (actin filament) by Nina et al. [19], the error size in the tip radii estimation from real data looks less than  $\pm 1.0$  nm. Thus, the robustness of our method would be applicable to real experimental data in the future study.

On the other hand, when 1.8 nm or 1.5 nm is used as the tip radius, the distinction between structural states becomes blurred, indicating that the kinetic isolation of CCA state is significantly lost at the radius of 1.5 nm. This observation is quantified by the relative entropies between two models which takes the minimum value when two radii are identical.

The relative entropies calculated with various conditions (different noise realizations, noise levels, tip radii for generating pseudo-AFM images) are shown in S13 Fig.

## Concluding remarks

In this study, we have developed a hidden Markov modeling-based method to analyze multiple AFM images as time series data as a natural extension of the conventional frame-by-frame rigid-body fitting. In modeling AFM data, molecular states can be defined mainly by a combination of structure and rotation. Here, we have proposed a modeling that constrains transitions in rotational space to take advantage of the tendency of molecules to interact with the substrate and remain time-correlated in the rotational direction. In contrast to the conventional frame-by-frame rigid-body fitting, the rotationally constrained Hidden Markov modeling can make estimations by using the statistics of multi-frame images. By twin experiments with simulated AFM data, we have shown that the proposed hidden Markov modeling is able to estimate the molecular orientations of a taste receptor with higher accuracy.

Using the proposed method, the estimation accuracy is expected to be improved the number of frames contained in a single AFM video increases, as shown in Fig 5. Thus, actual experimental AFM data could be modeled with higher accuracy than existing methods. However, there are several issues that must be overcome before the method applied to actual experimental data. The first issue is that a threshold for rotational constraint must be determined in advance. Although we used an arbitrary threshold for the quaternion distance in this study, it is difficult to know in advance how much the rotation of a sample molecule changes per frame. A possible prescription for this is to perform cross-correlation analysis (used in CryoEM data analysis [54]) between images of adjacent frames in AFM data, and determine a threshold from the distribution of angular differences between adjacent frames.

Another issue is that the candidate structures for rigid-body fitting may not contain the “correct” structures corresponding to the experimental data. In this case, it is necessary to reconsider the candidate structures. A direct solution is to perform flexible fitting to AFM images using MD simulations to search for structures that are closer to the experimental images. Although the flexible fitting method can search the structures fitted to the AFM images, the method is prone to overfit in the case of sparse data like AFM images. The use of a Bayesian engine such as MELD [55] may work in this situation for finding physically reasonable structures as well as avoiding overfitting, as was done in CryoEM data analysis (CryoFold

[56]). On the other hand, a variety of ensemble refinements and approaches, including the maximum entropy method [57] and a Bayesian method [58], would be useful in generating structural ensembles that match AFM images. A challenge in applying the ensemble refinements method to single-molecule measurement data like HS-AFM is how to extract temporal heterogeneous information resolved by single-molecule measurements. Time-series modeling based on state space modeling [59] investigated in this study combine with the ensemble refinements method should be developed to well capture the heterogeneous information.

In this study, we focused mainly on orientation estimations, but in order to improve the accuracy of not only orientations but also structure estimations, it is important to use the tip shape closer to the correct one. The tip can be directly imaged by using scanning or tunnelling electron microscopy (SEM and TEM). However, since SEM and TEM provide only 2D projections of a sample, it is difficult to routinely reconstruct the 3D morphology of the tip using these approaches. Moreover, during AFM experiments, there is a possibility that the tip is partially damaged and its shape is changed. Thus, it is necessary to estimate the tip shape at the time of measurement using only AFM data. The blind tip reconstruction is a famous algorithm to estimate tip shape only from AFM image(s) [60]. The drawback of the algorithm, however, is susceptible to noise. Since HS-AFM is more prone to noise than conventional AFM, it would be necessary to develop a new method that is more robust against noise.

The estimation with the hidden Markov modeling is expected to be improved when the more time series data are available. Therefore, if multiple molecules are observed in a movie, the accuracy can be improved by collecting multiple time series using the particle detection algorithm [61], such as the Hessian blob algorithm. In addition, estimation accuracy is greatly affected by spatial resolution. Recently, Heath et al. developed localization AFM (LAFM) [62], an elegant image reconstruction technique to overcome resolution limitations in AFM. Combined with these super-resolution methods is an important future direction to investigate more detailed dynamics.

Although this paper described the proposed method as a natural extension of the rigid-body fitting, the method could be also used in the context of integrative structural modeling [63]. By integrating both experimental and simulated data, errors caused by model parameters in the simulations can be corrected, or experimental data can be interpreted in unprecedented details by using simulation structures. In the analysis of CryoEM density map, flexible fitting simulations has been successfully used for the refinement and interpretation of experimental data [64,65]. In the context of single-molecule measurements, Matsunaga and Sugita recently refined MSM originally constructed from simulation data using single molecule FRET measurement data to give a detailed interpretation of the experimental data [66,59]. In the same way, it would be possible to construct an accurate MSM by integrating HS-AFM and simulation data. An issue, in constructing a MSM from AFM data, is that possible dependence of structural dynamics on orientations. For some molecules, the interaction with the stage in a certain orientation may change the structural dynamics. In such cases, it is not possible to simply reduce the transition probability with respect to the orientation, as was done in this study (Fig 6). It is necessary to check the transition probability and MSM for each orientation and to check whether the interactions with the stage cause outliers in the structural dynamics. To do this, the proposed method in this work would be useful to make reliable estimations for molecular orientations.

## Supporting information

**S1 Fig. Computational protocol for generating data for twin experiment.** (A) Markov state model constructed from coarse-grained molecular dynamics simulation of the taste receptor

type 1. (B) A series of structures were generated by stochastic simulation of the MSM and by drawing the centroid structures of the MSM states. (C) After rotating the structures to a specific orientation, the series of AFM image were generated by the collision-detection with the tip. Finally, spatially independent Gaussian noise were added to pixels.  
(TIF)

**S2 Fig. Results of twin experiments using different pseudo-random generator seeds for Gaussian noise in pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).  
(TIF)

**S3 Fig. Another set of results of twin experiments using another set of different pseudo-random generator seeds for Gaussian noise in pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).  
(TIF)

**S4 Fig. Results of twin experiments using Gaussian noise with a standard deviation of 0.6 nm for pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).  
(TIF)

**S5 Fig. Results of twin experiments using Gaussian noise with a standard deviation of 0.9 nm for pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).  
(TIF)

**S6 Fig. Results of twin experiments using Gaussian noise with a standard deviation of 1.2 nm for pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi

algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).

(TIF)

**S7 Fig. Results of twin experiments using the ground-truth tip radius of 3.5 nm.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (2.5 nm, 2.8 nm, 3.5 nm that is the ground-truth, 4.2 nm, 4.5 nm).

(TIF)

**S8 Fig. Results of twin experiments using different cluster centers from those of the Markov state model used for generating pseudo-AFM images.** Histograms of root-mean square deviations (RMSDs) of estimated structures from the ground-truth structures in twin experiments. Note that the RMSDs were computed without structural alignment. Structural estimations were performed with various conditions: with different algorithms (the frame-by-frame rigid-body fitting, the Viterbi algorithm using the ground-truth transition probabilities, estimated probabilities by the Baum-Welch algorithm, and a random matrix), and different tip radii (1.5 nm, 1.8 nm, 2.5 nm that is the ground-truth, 3.2 nm, 3.5 nm).

(TIF)

**S9 Fig. Likelihood distribution over all possible orientations and structures computed using the first single frame of the first set AFM data (defined by Eq 2).** Likelihoods are projected on to the 1st principal components (PC) of the Cartesian coordinates of structures. The maximum likelihoods are indicated by red arrows. The 1st PC of the ground-truth structure is indicated by red solid lines. (A) Likelihoods computed with the tip apex radius of 1.8 nm. (B) 2.5 nm (the ground-truth tip radius).

(TIF)

**S10 Fig. Maximum likelihood of pseudo-AFM time-series images in each orientation computed using the first single frame, the first 10 frames, the first 100 frames of the first set AFM data (defined by Eq 3).** Likelihood values are projected onto the 1st principal component of the structure of the first frame. Note that only the maximum likelihood of each orientation is shown due to computational costs (calculated by the Viterbi algorithm). The maximum likelihoods are indicated by red arrows. The 1st PC of the ground-truth structure is indicated by red solid lines. (A) Likelihoods computed with the tip apex radius of 1.5 nm. (B) 1.8 nm.

(TIF)

**S11 Fig. Confidence interval of transition probabilities  $T_{i,j}$  obtained by the Baum-Welch algorithm evaluated by bootstrap sampling using the tip radius of 2.5 nm.** Blue bars indicate 25 estimated transition probabilities from the bootstrap sampling. Red bars indicate the ground-truth transition probabilities.

(TIF)

**S12 Fig. Confidence interval of transition probabilities  $T_{i,j}$  obtained by the Baum-Welch algorithm evaluated by bootstrap sampling using the tip radius of 2.0 nm.** Blue bars indicate 25 estimated transition probabilities from the bootstrap sampling. Red bars indicate the

ground-truth transition probabilities.  
(TIF)

**S13 Fig. Accuracies of reconstructed MSMs evaluated by relative entropies for various conditions.** (A) Different pseudo-random number seeds used for Gaussian noise in pseudo-AFM images. (B) Another set of different pseudo-random number seeds for Gaussian noise. (C) Tip radius of 3.5 nm (the ground-truth is 2.5 nm). (D) Standard deviation of 0.6 nm for Gaussian noise in pseudo-AFM images. (E) Standard deviation of 0.9 nm. (F) Standard deviation of 1.2 nm.  
(TIF)

## Acknowledgments

We thank Shoji Takada and Sotaro Fuchigami for stimulating discussions. We also thank Toru Niina as our collision-detection code is based on his *afmize*.

## Author Contributions

**Conceptualization:** Yasuhiro Matsunaga.

**Data curation:** Daisuke Noshiro, Yasuhiro Matsunaga.

**Formal analysis:** Yasuhiro Matsunaga.

**Funding acquisition:** Toshio Ando, Atsuko Yamashita, Yuji Sugita, Yasuhiro Matsunaga.

**Investigation:** Tomonori Ogane, Daisuke Noshiro, Toshio Ando, Atsuko Yamashita, Yuji Sugita, Yasuhiro Matsunaga.

**Methodology:** Yuji Sugita, Yasuhiro Matsunaga.

**Project administration:** Toshio Ando, Atsuko Yamashita, Yuji Sugita.

**Resources:** Daisuke Noshiro, Toshio Ando, Atsuko Yamashita.

**Software:** Tomonori Ogane, Yasuhiro Matsunaga.

**Supervision:** Toshio Ando, Atsuko Yamashita, Yuji Sugita, Yasuhiro Matsunaga.

**Validation:** Yasuhiro Matsunaga.

**Visualization:** Tomonori Ogane, Yasuhiro Matsunaga.

**Writing – original draft:** Atsuko Yamashita, Yuji Sugita, Yasuhiro Matsunaga.

**Writing – review & editing:** Tomonori Ogane, Daisuke Noshiro, Toshio Ando, Atsuko Yamashita, Yuji Sugita, Yasuhiro Matsunaga.

## References

1. Ha T. Single-Molecule Fluorescence Resonance Energy Transfer. *Methods*. 2001; 25: 78–86. <https://doi.org/10.1006/meth.2001.1217> PMID: 11558999
2. Schuler B, Eaton WA. Protein folding studied by single-molecule FRET. *Current Opinion in Structural Biology*. 2008; 18: 16–26. <https://doi.org/10.1016/j.sbi.2007.12.003> PMID: 18221865
3. Moffitt JR, Chemla YR, Smith SB, Bustamante C. Recent Advances in Optical Tweezers. *Annu Rev Biochem*. 2008; 77: 205–228. <https://doi.org/10.1146/annurev.biochem.77.043007.090225> PMID: 18307407
4. Binnig G, Quate CF, Gerber Ch. Atomic Force Microscope. *Phys Rev Lett*. 1986; 56: 930–933. <https://doi.org/10.1103/PhysRevLett.56.930> PMID: 10033323

5. Ando T, Uchihashi T, Fukuma T. High-speed atomic force microscopy for nano-visualization of dynamic biomolecular processes. *Progress in Surface Science*. 2008; 83: 337–437. <https://doi.org/10.1016/j.progsurf.2008.09.001>
6. Ando T. *High-Speed Atomic Force Microscopy in Biology*. Springer; 2022.
7. Kodera N, Yamamoto D, Ishikawa R, Ando T. Video imaging of walking myosin V by high-speed atomic force microscopy. *Nature*. 2010; 468: 72–76. <https://doi.org/10.1038/nature09450> PMID: 20935627
8. Uchihashi T, Iino R, Ando T, Noji H. High-Speed Atomic Force Microscopy Reveals Rotary Catalysis of Rotorless F<sub>1</sub>-ATPase. *Science*. 2011; 333: 755–758. <https://doi.org/10.1126/science.1205510> PMID: 21817054
9. Shibata M, Nishimasu H, Kodera N, Hirano S, Ando T, Uchihashi T, et al. Real-space and real-time dynamics of CRISPR-Cas9 visualized by high-speed atomic force microscopy. *Nat Commun*. 2017; 8: 1430. <https://doi.org/10.1038/s41467-017-01466-8> PMID: 29127285
10. Kodera N, Noshiro D, Dora SK, Mori T, Habchi J, Blocquel D, et al. Structural and dynamics analysis of intrinsically disordered proteins by high-speed atomic force microscopy. *Nat Nanotechnol*. 2021; 16: 181–189. <https://doi.org/10.1038/s41565-020-00798-9> PMID: 33230318
11. Ando T, Kodera N, Uchihashi T, Miyagi A, Nakakita R, Yamashita H, et al. High-speed Atomic Force Microscopy for Capturing Dynamic Behavior of Protein Molecules at Work. *e-J Surf Sci Nanotechnol*. 2005; 3: 384–392. <https://doi.org/10.1380/ejssnt.2005.384>
12. Scheuring S, Busselez J, Lévy D. Structure of the Dimeric PufX-containing Core Complex of *Rhodobacter blasticus* by in Situ Atomic Force Microscopy. *Journal of Biological Chemistry*. 2005; 280: 1426–1431. <https://doi.org/10.1074/jbc.M411334200> PMID: 15522874
13. Scheuring S, Boudier T, Sturgis JN. From high-resolution AFM topographs to atomic models of supra-molecular assemblies. *Journal of Structural Biology*. 2007; 159: 268–276. <https://doi.org/10.1016/j.jsb.2007.01.021> PMID: 17399998
14. Asakawa H, Ikegami K, Setou M, Watanabe N, Tsukada M, Fukuma T. Submolecular-Scale Imaging of  $\alpha$ -Helices and C-Terminal Domains of Tubulins by Frequency Modulation Atomic Force Microscopy in Liquid. *Biophysical Journal*. 2011; 101: 1270–1276. <https://doi.org/10.1016/j.bpj.2011.07.020> PMID: 21889465
15. Trinh M-H, Odorico M, Pique ME, Teulon J-M, Roberts VA, Ten Eyck LF, et al. Computational Reconstruction of Multidomain Proteins Using Atomic Force Microscopy Data. *Structure*. 2012; 20: 113–120. <https://doi.org/10.1016/j.str.2011.10.023> PMID: 22244760
16. Chaves RC, Teulon J-M, Odorico M, Parot P, Chen SW, Pellequer J-L. Conformational dynamics of individual antibodies using computational docking and AFM: CONFORMATIONAL DYNAMICS OF IGG USING DOCKING AND AFM. *J Mol Recognit*. 2013; 26: 596–604. <https://doi.org/10.1002/jmr.2310> PMID: 24089367
17. Dasgupta B, Miyashita O, Tama F. Reconstruction of low-resolution molecular structures from simulated atomic force microscopy images. *Biochimica et Biophysica Acta (BBA)—General Subjects*. 2020; 1864: 129420. <https://doi.org/10.1016/j.bbagen.2019.129420> PMID: 31472175
18. Amyot R, Flechsig H. BioAFMviewer: An interactive interface for simulated AFM scanning of biomolecular structures and dynamics. Schneidman-Duhovny D, editor. *PLoS Comput Biol*. 2020; 16: e1008444. <https://doi.org/10.1371/journal.pcbi.1008444> PMID: 33206646
19. Niina T, Matsunaga Y, Takada S. Rigid-body fitting to atomic force microscopy images for inferring probe shape and biomolecular structure. *PLOS Computational Biology*. 2021; 17: e1009215. <https://doi.org/10.1371/journal.pcbi.1009215> PMID: 34283829
20. Amyot R, Marchesi A, Franz CM, Casuso I, Flechsig H. Simulation atomic force microscopy for atomic reconstruction of biomolecular structures from resolution-limited experimental images. *PLoS Comput Biol*. 2022; 18: e1009970. <https://doi.org/10.1371/journal.pcbi.1009970> PMID: 35294442
21. Niina T, Fuchigami S, Takada S. Flexible Fitting of Biomolecular Structures to Atomic Force Microscopy Images via Biased Molecular Simulations. *J Chem Theory Comput*. 2020; 16: 1349–1358. <https://doi.org/10.1021/acs.jctc.9b00991> PMID: 31909999
22. Fuchigami S, Niina T, Takada S. Particle Filter Method to Integrate High-Speed Atomic Force Microscopy Measurements with Biomolecular Simulations. *J Chem Theory Comput*. 2020; 16: 6609–6619. <https://doi.org/10.1021/acs.jctc.0c00234> PMID: 32805119
23. Rabiner L, Juang B. An introduction to hidden Markov models. *IEEE ASSP Magazine*. 1986; 3: 4–16. <https://doi.org/10.1109/MASSP.1986.1165342>
24. Müller DJ, Engel A. The height of biomolecules measured with the atomic force microscope depends on electrostatic interactions. *Biophysical Journal*. 1997; 73: 1633–1644. [https://doi.org/10.1016/S0006-3495\(97\)78195-5](https://doi.org/10.1016/S0006-3495(97)78195-5) PMID: 9284330

25. Cossio P, Hummer G. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of Structural Biology*. 2013; 184: 427–437. <https://doi.org/10.1016/j.jsb.2013.10.006> PMID: 24161733
26. Yershova A, LaValle SM, Mitchell JC. Generating Uniform Incremental Grids on SO(3) Using the Hopf Fibration.: 18.
27. Cossio P, Rohr D, Baruffa F, Rampp M, Lindenstruth V, Hummer G. BioEM: GPU-accelerated computing of Bayesian inference of electron microscopy images. *Computer Physics Communications*. 2017; 210: 163–171. <https://doi.org/10.1016/j.cpc.2016.09.014>
28. Roweis S. Constrained Hidden Markov Models.: 7.
29. Nuemket N, Yasui N, Kusakabe Y, Nomura Y, Atsumi N, Akiyama S, et al. Structural basis for perception of diverse chemical substances by T1r taste receptors. *Nat Commun*. 2017; 8: 15530. <https://doi.org/10.1038/ncomms15530> PMID: 28534491
30. Xu H, Staszewski L, Tang H, Adler E, Zoller M, Li X. Different functional roles of T1R subunits in the heteromeric taste receptors. *Proc Natl Acad Sci USA*. 2004; 101: 14258–14263. <https://doi.org/10.1073/pnas.0404384101> PMID: 15353592
31. Husic BE, Pande VS. Markov State Models: From an Art to a Science. *J Am Chem Soc*. 2018; 140: 2386–2396. <https://doi.org/10.1021/jacs.7b12191> PMID: 29323881
32. Wang W, Cao S, Zhu L, Huang X. Constructing Markov State Models to elucidate the functional conformational changes of complex biomolecules. *WIREs Comput Mol Sci*. 2018;8. <https://doi.org/10.1002/wcms.1343>
33. Chodera JD, Noé F. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*. 2014; 25: 135–144. <https://doi.org/10.1016/j.sbi.2014.04.002> PMID: 24836551
34. Geng Y, Mosyak L, Kurinov I, Zuo H, Sturchler E, Cheng TC, et al. Structural mechanism of ligand activation in human calcium-sensing receptor. *eLife*. 2016; 5: e13662. <https://doi.org/10.7554/eLife.13662> PMID: 27434672
35. Muto T, Tsuchiya D, Morikawa K, Jingami H. Structures of the extracellular regions of the group II/III metabotropic glutamate receptors. *Proc Natl Acad Sci USA*. 2007; 104: 3759–3764. <https://doi.org/10.1073/pnas.0611577104> PMID: 17360426
36. Schlitter J, Engels M, Krüger P. Targeted molecular dynamics: A new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics*. 1994; 12: 84–89. [https://doi.org/10.1016/0263-7855\(94\)80072-3](https://doi.org/10.1016/0263-7855(94)80072-3) PMID: 7918256
37. Karanicolas J, Brooks CL. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science*. 2009; 11: 2351–2361. <https://doi.org/10.1110/ps.0205402> PMID: 12237457
38. Karanicolas J, Brooks CL. Improved Gō-like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-native Interactions. *Journal of Molecular Biology*. 2003; 334: 309–325. <https://doi.org/10.1016/j.jmb.2003.09.047> PMID: 14607121
39. Jung J, Mori T, Kobayashi C, Matsunaga Y, Yoda T, Feig M, et al. GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations: GENESIS. *WIREs Comput Mol Sci*. 2015; 5: 310–323. <https://doi.org/10.1002/wcms.1220> PMID: 26753008
40. Kobayashi C, Jung J, Matsunaga Y, Mori T, Ando T, Tamura K, et al. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *Journal of Computational Chemistry*. 2017; 38: 2193–2206. <https://doi.org/10.1002/jcc.24874> PMID: 28718930
41. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*. 2008; 29: 1859–1865. <https://doi.org/10.1002/jcc.20945> PMID: 18351591
42. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*. 2016;54. <https://doi.org/10.1002/cpbi.3> PMID: 27322406
43. Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, et al. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone  $\phi$ ,  $\psi$  and Side-Chain  $\chi_1$  and  $\chi_2$  Dihedral Angles. *J Chem Theory Comput*. 2012; 8: 3257–3273. <https://doi.org/10.1021/ct300400x> PMID: 23341755
44. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*. 2013; 34: 2135–2145. <https://doi.org/10.1002/jcc.23354> PMID: 23832629
45. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983; 79: 926–935. <https://doi.org/10.1063/1.445869>

46. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*. 1995; 103: 8577–8593. <https://doi.org/10.1063/1.470117>
47. Ryckaert J-P, Ciccotti G, Berendsen HJC. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*. 1977; 23: 327–341. [https://doi.org/10.1016/0021-9991\(77\)90098-5](https://doi.org/10.1016/0021-9991(77)90098-5)
48. Miyamoto S, Kollman PA. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*. 1992; 13: 952–962. <https://doi.org/10.1002/jcc.540130805>
49. Feig M, Karanicolas J, Brooks CL. MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *Journal of Molecular Graphics and Modelling*. 2004; 22: 377–395. <https://doi.org/10.1016/j.jmgs.2003.12.005> PMID: 15099834
50. Kitao A, Hirata F, Gō N. The effects of solvent on the conformation and the collective motions of protein: Normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum. *Chemical Physics*. 1991; 158: 447–472. [https://doi.org/10.1016/0301-0104\(91\)87082-7](https://doi.org/10.1016/0301-0104(91)87082-7)
51. Sagi O, Rokach L. Ensemble learning: A survey. *WIREs Data Mining Knowl Discov*. 2018;8. <https://doi.org/10.1002/widm.1249>
52. Visser I, Raijmakers MEJ, Molenaar PCM. Confidence intervals for hidden Markov model parameters. *British Journal of Mathematical and Statistical Psychology*. 2000; 53: 317–327. <https://doi.org/10.1348/000711000159240> PMID: 11109710
53. Bowman GR, Ensign DL, Pande VS. Enhanced Modeling via Network Theory: Adaptive Sampling of Markov State Models. *J Chem Theory Comput*. 2010; 6: 787–794. <https://doi.org/10.1021/ct900620b> PMID: 23626502
54. Tan YZ, Baldwin PR, Davis JH, Williamson JR, Potter CS, Carragher B, et al. Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat Methods*. 2017; 14: 793–796. <https://doi.org/10.1038/nmeth.4347> PMID: 28671674
55. MacCallum JL, Perez A, Dill KA. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc Natl Acad Sci USA*. 2015; 112: 6985–6990. <https://doi.org/10.1073/pnas.1506788112> PMID: 26038552
56. Shekhar M, Terashi G, Gupta C, Sarkar D, Debussche G, Sisco NJ, et al. CryoFold: Determining protein structures and data-guided ensembles from cryo-EM density maps. *Matter*. 2021; 4: 3195–3216. <https://doi.org/10.1016/j.matt.2021.09.004> PMID: 35874311
57. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K. Combining Experiments and Simulations Using the Maximum Entropy Principle. Levitt M, editor. *PLoS Comput Biol*. 2014; 10: e1003406. <https://doi.org/10.1371/journal.pcbi.1003406> PMID: 24586124
58. Hummer G, Köfinger J. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics*. 2015; 143: 243150. <https://doi.org/10.1063/1.4937786> PMID: 26723635
59. Matsunaga Y, Sugita Y. Use of single-molecule time-series data for refining conformational dynamics in molecular simulations. *Current Opinion in Structural Biology*. 2020; 61: 153–159. <https://doi.org/10.1016/j.sbi.2019.12.022> PMID: 32004808
60. Villarrubia JS. Algorithms for scanned probe microscope image simulation, surface reconstruction, and tip estimation. *J Res Natl Inst Stand Technol*. 1997; 102: 425. <https://doi.org/10.6028/jres.102.030> PMID: 27805154
61. Marsh BP, Chada N, Sanganna Gari RR, Sigdel KP, King GM. The Hessian Blob Algorithm: Precise Particle Detection in Atomic Force Microscopy Imagery. *Sci Rep*. 2018; 8: 978. <https://doi.org/10.1038/s41598-018-19379-x> PMID: 29343783
62. Heath GR, Kots E, Robertson JL, Lansky S, Khelashvili G, Weinstein H, et al. Localization atomic force microscopy. *Nature*. 2021; 594: 385–390. <https://doi.org/10.1038/s41586-021-03551-x> PMID: 34135520
63. van den Bedem H, Fraser JS. Integrative, dynamic structural biology at atomic resolution—it's about time. *Nat Methods*. 2015; 12: 307–318. <https://doi.org/10.1038/nmeth.3324> PMID: 25825836
64. Singharoy A, Teo I, McGreevy R, Stone JE, Zhao J, Schulten K. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife*. 2016; 5: e16105. <https://doi.org/10.7554/eLife.16105> PMID: 27383269
65. Mori T, Terashi G, Matsuoka D, Kihara D, Sugita Y. Efficient Flexible Fitting Refinement with Automatic Error Fixing for De Novo Structure Modeling from Cryo-EM Density Maps. *J Chem Inf Model*. 2021; 61: 3516–3528. <https://doi.org/10.1021/acs.jcim.1c00230> PMID: 34142833
66. Matsunaga Y, Sugita Y. Linking time-series of single-molecule experiments with molecular dynamics simulations by machine learning. *eLife*. 2018; 7: e32668. <https://doi.org/10.7554/eLife.32668> PMID: 29723137