NMMD: Efficient cryo-EM flexible fitting based on simultaneous Normal Mode and Molecular Dynamics atomic displacements

Rémi Vuillemot, Osamu Miyashita, Florence Tama, Isabelle Rouiller, Slavica Jonic

This is the author's version of an article accepted for publication in Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2022.167483

NMMD: Efficient cryo-EM flexible fitting based on simultaneous Normal Mode and Molecular Dynamics atomic displacements

Rémi Vuillemot [1,4], **Osamu Miyashita** [2], **Florence Tama** [3], **Isabelle Rouiller** [4], **Slavica Jonic** [1,*]

[1] *IMPMC - UMR 7590 CNRS, Sorbonne Université, Muséum National d'Histoire Naturelle, Paris, France*

[2] *RIKEN Center for Computational Science, Japan*

[3] *Institute of Transformative Biomolecules and Department of Physics, Graduate School of Science, Nagoya University, Japan*

[4] *Department of Biochemistry & Pharmacology and Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Victoria, Australia*

Correspondence*:
Dr. Slavica Jonic
Sorbonne Univerisité
IMPMC - UMR CNRS 7590
4 Place Jussieu, 75005 Paris, France
Phone: +33 1 44 27 72 05
Fax: +33 1 44 27 37 85
Email: slavica.jonic@upmc.fr

ABSTRACT

Atomic models of cryo electron microscopy (cryo-EM) maps of biomolecular conformations are often obtained by flexible fitting of the maps with available atomic structures of other conformations (e.g., obtained by X-ray crystallography). This article presents a new flexible fitting method, NMMD, which combines normal mode analysis (NMA) and molecular dynamics simulation (MD). Given an atomic structure and a cryo-EM map to fit, NMMD simultaneously estimates global atomic displacements based on NMA and local displacements based on MD. NMMD was implemented by modifying EMfit, a flexible fitting method using MD only, in GENESIS 1.4. As EMfit, NMMD can be run with replica exchange umbrella sampling procedure. The new method was tested using a variety of EM maps (synthetic and experimental, with different noise levels and resolutions). The results of the tests show that adding normal modes to MD-based fitting makes the fitting faster (40% in average) and, in the majority of cases, more accurate.

Keywords: Molecular modeling, cryo electron microscopy (cryo-EM), flexible fitting, normal mode analysis, molecular dynamics simulation, replica exchange umbrella sampling, EMfit, GENESIS

1 INTRODUCTION

The study of conformational variability of biological macromolecular complexes is the key to deciphering their biological functions and to structure-based drug development. Cryo Electron Microscopy (cryo-EM) is an experimental technique to image structures of biological macromolecular complexes in their close-to-native conditions. Unlike X-ray crystallography, cryo-EM allows imaging and reconstructing multiple conformational states of a complex from the same sample and does not require sample crystallization Jonić (2017). In the recent years, cryo-EM benefited from fast improvements in terms of the quality of three-dimensional (3D) reconstruction. Obtaining near-atomic resolution of structures by Single Particle Analysis (SPA) is now becoming more and more common Nakane et al. (2020); Bai et al. (2015); Frank (2017). SPA is based on collecting a large number of images of the complexes at unknown random positions and orientations, which are then estimated to reconstruct a 3D density volume (also known as cryo-EM map). Interpreting cryo-EM maps in terms of atomic positions is still a major obstacle, even for near-atomic resolution cryo-EM maps. To solve this problem, cryo-EM maps are frequently fitted with already available atomic structures. For instance, a cryo-EM map can be fitted with an atomic structure obtained by X-ray crystallography for a different conformation of the complex from the one present in the

cryo-EM map. Flexible fitting of such 3D atomic structures in the density volumes is not trivial, due to a considerable number of atomic degrees of freedom of biological macromolecular complexes. To perform flexible deformations of the atomic structure while preserving important properties of the system, several models of the potential energy of the system have been used, from standard semiempirical potentials used in fitting based on Molecular Dynamics (MD) simulations Wu et al. (2013); Trabuco et al. (2008); Orzechowski and Tama (2008); Miyashita et al. (2017); Igaev et al. (2019) to simpler pairwise Hookean potentials used in fitting based on Normal Mode Analysis (NMA) and Elastic Network Model (ENM) Tirion (1996); Tama et al. (2002, 2004a); Schröder et al. (2007); Lopéz-Blanco and Chacón (2013).

In an MD simulation, the conformational space is sampled using complex semiempirical potentials (force fields), with a large number of atomic degrees of freedom (usually, three Cartesian coordinates for each atom of the molecule), which makes the conformational space exploration very slow. Flexible fitting methods based on MD simulations employ an additional potential (biasing potential) that guides the simulation towards the target cryo-EM map, making the conformational space sampling more efficient.

The biasing potential is usually defined based on the correlation coefficient (CC) between the target EM map and the map simulated from the structure being fitted Orzechowski and Tama (2008); Miyashita et al. (2017); Igaev et al. (2019). However, some methods define it as a potential field that pushes the atoms to high-density regions of the EM map Trabuco et al. (2008), which may involve system-dependent manipulations of the EM map. For instance, MDFF and its interactive version iMDFF Trabuco et al. (2008) involve thresholding low density regions of the EM map, rescaling intensities of the EM map voxels, and incorporating restraint potentials to conserve the secondary structure elements and avoid over-fitting. On the contrary, the implementation of a CC-based biasing potential does not require any complicated EM map manipulations or imposing restraints on the secondary structures.

The approaches with the biasing potential based on the CC and those based the potential field produce comparable fitting results. In the case of high-resolution EM maps, the atomic structure tends to get trapped into local optima due to inefficient conformational sampling in rough density regions present in such maps. This problem can be alleviated by gradually increasing the intensity of the biasing potential during the fitting, to move atoms first globally and then locally. To this end, CDMD Igaev et al. (2019) proposes a multi-resolution CC-based potential, where the resolution of the map simulated from the atomic structure and the force constant controlling the contribution of the biasing potential are both gradually increased

during the fitting. Another approach to alleviate the problem of local optima is Replica Exchange Umbrella Sampling (REUS), where multiple replicas are run in parallel and exchange force constants (highest force constants to replicas with the highest CC), so as to gradually increase them Miyashita et al. (2017). Besides, the CC variation over the different replicas can be used to estimate the uncertainty of the fitting. A small CC variation would suggest that the obtained conformations are nearly the same for all the replicas and that the best-matching conformation in such a case may be trusted more than in the case with a large CC variation over the replicas.

Flexible fitting methods based on MD simulations suffer from high computational cost of using large number of atomic degrees of freedom, which slows down the sampling of large conformational changes. A fast but rough conformational space sampling can be obtained using NMA with ENM, which is based on reducing the number of atomic degrees of freedom to a small number of vectors (known as normal modes) along which atoms are allowed to move. The number of atoms moved with a normal mode is counted using the so-called collectivity measure Brüschweiler (1995). Highly collective modes are those along which all atoms move. Large-scale conformational changes have been described as global, collective motions represented by a few low-frequency normal modes Tama and Sanejouand (2001); Delarue and Dumas (2004); Wang et al. (2004); Ma (2005); Suhre et al. (2006); Tama and Brooks III (2006). Normal modes are obtained using a simplified potential based on Tirion's ENM Tirion (1996) and a linear combination of selected low-frequency normal modes is used to simulate global dynamics of the system. Therefore, flexible fitting approaches based on NMA with ENM drastically reduce the number of atomic degrees of freedom by using only a few lowest-frequency normal modes and neglecting all other modes Delarue and Dumas (2004); Tama et al. (2004a); Suhre et al. (2006). The optimal linear combination of the selected low-frequency normal modes is estimated by maximizing a measure of similarity between the target EM map and the map simulated from the atomic structure being fitted, for instance the CC Tama et al. (2004a). The reduction of the number of atomic degrees of freedom to a small number of lowest-frequency normal modes yields faster sampling of large conformational spaces. Unfortunately, neglecting high-frequency normal modes lowers the accuracy of describing more localized motions during the fitting. Moreover, large-amplitude motions along normal modes usually induce distortions that might alter the mechanical properties of the structure.

To take the best from both MD-based and NMA-based flexible fitting methods, one can think of combining them. A recent work (Costa et al. (2020)) proposed to initiate an MD simulation using a linear combination of normal modes whose amplitudes are sampled with a Monte Carlo approach so as to guide the simulation towards the conformation in the given EM map. Given an initial structure and a selected set of its low-frequency normal modes, this method tries randomly generated combinations of normal-mode amplitudes to find the one that produces the NMA-based structural displacement in the direction of the CC increase. The obtained linear combination of normal modes is then used to initiate (update the velocities) a short, 2-ps MD simulation. This procedure is iterated until the CC convergence, considering the structure resulting from the MD run at the previous iteration as the initial structure for the next iteration, which involves recalculating normal modes of the initial structure at each iteration. Such MD and NMA combinations may be computationally expensive, particularly for large conformational differences between the given EM map and atomic structure, as possibly requiring many recalculations of normal modes during the fitting process. Furthermore, the Monte Carlo estimation of normal-mode amplitudes at each iteration, to generate the ”excitation” direction, may not be the most efficient, inducing extra computation time.

In this article, we propose a method that efficiently combines MD-based and NMA-based approaches, resulting in a speed-up of the fitting and an accurate description of both global and local motions during the fitting. Instead of a stochastic estimation of the gradient of the CC, our method uses analytic expressions to calculate the gradient, which ensures fast and accurate estimation of normal-mode amplitudes. Also, our method requires a single calculation of normal modes (at the beginning of the fitting process) and a single MD simulation run, without any restriction regarding the duration of the MD simulation. Furthermore, our method updates normal-mode amplitudes in each MD simulation step, which improves accuracy of their estimation.

More precisely, we modified the atomic displacement used in MD-based fitting to incorporate a displacement due to a linear combination of normal modes. The amplitudes of the linear combination of normal modes are estimated simultaneously with the MD-based atomic displacements, by integrating the equation of motion in the same fashion as in MD-based fitting. The amplitudes of normal modes are updated at every step of the simulation with almost no extra integration cost, as the number of normal modes is much smaller than the number of atomic coordinates (Cartesian coordinates). The method has been coupled with a multireplica procedure using REUS to improve fitting by adjusting the force constant

of the biasing potential. This new fitting approach will be referred to as Normal Mode and Molecular Dynamics (NMMD).

NMMD has been implemented in GENESIS, an open-source MD simulation software Kobayashi et al. (2017). GENESIS includes an implementation of the MD-based flexible fitting method EMFit and an implementation of the REUS method Miyashita et al. (2017); Orzechowski and Tama (2008). In this article, we describe the NMMD approach and present its results using synthetic cryo-EM maps of four molecular complexes (LAO binding protein (LAO) Oh et al. (1993), Adenylate kinase (AK) Müller et al. (1996); Müller and Schulz (1992), Lactoferrin (LF) Haridas et al. (1995); Norris et al. (1991), and Elongation factor 2 (EF2) Jørgensen et al. (2003)) and experimental cryo-EM maps of two complexes (p97 ATPase (p97) Banerjee et al. (2016) and ABC exporter (ABC) Hofmann et al. (2019)). Also, we show that fitting using the combination of MD and NMA samples the conformational space more efficiently than fitting using only MD, by comparing the results of NMMD and EMFit using the same data.

2 METHODS

2.1 MD-based flexible fitting

To simulate the transition between the conformations in the given atomic structure and cryo-EM map, aligned in terms of rigid-body transformations, one can use the so-called biased MD. In this approach, a biasing potential is added to the standard MD-simulation force field potential U_{mol} in order to guide the simulation towards the target conformation Orzechowski and Tama (2008) Miyashita and Tama (2018). This biasing potential must incorporate information about the quality of the atomic structural model fitting with the cryo-EM map. One popular approach to define the biasing potential is to use the correlation coefficient (CC), in which case the total potential is as follows:

$$U = U_{mol} + k(1 - CC). \tag{1}$$

The CC has values between 0 and 1 and measures the similarity between the given cryo-EM map and the map simulated from the atomic structure during the fitting. A value of 1 means that the densities fit perfectly, whereas a low CC value corresponds to a fit of low quality. The potential is proportional to $1 - CC$ to guide the simulation to the highest CC values. The factor k is the force constant that defines the balance between the biasing potential and the classical potential and is expressed in *kcal/mol*. The output

of the fitting strongly depends on the choice of k and a procedure to automatically adjust k is described below. The CC is defined as follows:

$$CC = \frac{\sqrt{\sum_{l=1}^{N_{vox}} \rho_{sim}^{\mathbf{r}_{md}}(l)\rho_{exp}(l)}}{\sqrt{\sum_{l=1}^{N_{vox}} \rho_{sim}^{\mathbf{r}_{md}}(l)^2}\sqrt{\sum_{l=1}^{N_{vox}} \rho_{exp}(l)^2}}, \tag{2}$$

where ρ_{exp} is the given cryo-EM map, $\rho_{sim}^{\mathbf{r}_{md}}$ is the map simulated from the fitted atomic structure, and $\mathbf{r}_{md} = \{\mathbf{r}_{md}^n\}$ is the vector of $3 \times N$ atomic (Cartesian) coordinates for the structure with N atoms $(n = 1, N)$. A way to simulate such density maps is by placing a three-dimensional Gaussian function at the position of each atom and integrating these Gaussian functions over each voxel in the volume Orzechowski and Tama (2008). The fitted atomic positions correspond to a displacement $\mathbf{x}(t)$ from the initial atomic positions $\mathbf{r}_0$:

$$\mathbf{r}_{md}(t) = \mathbf{x}(t) + \mathbf{r}_0. \tag{3}$$

2.2 NMA-based flexible fitting

The biased MD fitting approach is computationally expensive as it relies on a large number of conformational degrees of freedom ($N \times 3$ atomic coordinates). One way to reduce these conformational degrees of freedom is by selecting a few normal modes that describe the most collective motions. NMA is commonly performed using Tirion's ENM Tirion (1996), where the potential energy is reduced to simple harmonic potentials between close atoms. In this case, the NMA calculation consists of diagonalizing the Hessian matrix of second derivatives of the potential energy function (a square matrix of dimension $N \times 3$). This produces a matrix of normal modes and their associated frequencies. The total number of normal modes is $N \times 3$ and the length of each normal mode is $N \times 3$. Usually, a small subset of M lowest-frequency normal modes (describing global, collective motions) is selected to displace atoms to fit the given EM map Tama et al. (2004b). The displaced atomic coordinates are determined by a linear combination of the selected normal modes, with normal-mode amplitudes as the coefficients of the linear combination:

$$\mathbf{r}_{nm} = \mathbf{q} \cdot \mathbf{A} + \mathbf{r}_0, \tag{4}$$

where $\mathbf{A} = \{\mathbf{a}_i\}$ is the matrix of the selected M normal modes (size $M \times (N \times 3)$) and $\mathbf{q} = \{q_i\}$ is the vector of M coefficients of the linear combination (M normal-mode amplitudes). Normal modes are ordered according to their frequency. The six lowest-frequency modes (modes 1-6) correspond to

rigid-body transformations and are not used for flexible fitting. NMA-based flexible fitting is done by optimizing a similarity measure like CC in equation (2) to tune the coefficients of the linear combination of normal modes that is applied on the given atomic structure until it fits the given EM map (Tama et al. (2004a); Miyashita and Tama (2018)). NMA-based fitting is much faster than MD-based fitting as the number of degrees of freedom is reduced to M (M normal-mode amplitudes q_i), where $M << N$ and, usually, $M < 10$. However, the selected lowest-frequency normal modes fit well global motions but not local motions.

2.3 NMMD: combined MD and NMA based flexible fitting

In this article, we propose NMMD approach, which combines NMA-based flexible fitting (small number of degrees of freedom well describing global motions) with MD-based flexible fitting (large number of degrees of freedom well describing local motions). In this approach, the computational cost of fitting large-scale conformational transitions is reduced thanks to the NMA-based fitting and the precision of fitting local dynamics is maintained thanks to the MD-based fitting. To this end, equation (3) was modified to add a NMA-based atomic displacement to the MD-based displacement $\mathbf{x}(t)$ from the initial atomic position $\mathbf{r}_0$, as follows:

$$\mathbf{r}(t) = \mathbf{q}(t) \cdot \mathbf{A} + \mathbf{x}(t) + \mathbf{r}_0, \tag{5}$$

where $\mathbf{q}(t) \cdot \mathbf{A}$ is the displacement of N atoms induced by a linear combination of M normal modes (given by matrix $\mathbf{A}$) with amplitudes $\mathbf{q}(t)$ at time t (M unknown parameters, $M << N$), $\mathbf{x}(t)$ is the atomic displacement at time t from classical MD ($N \times 3$ unknown parameters, equation (3)). It can be noted that the total number of unknown parameters in NMMD ($\mathbf{q}(t)$ and $\mathbf{x}(t)$) at a simulation step t is $M + N \times 3$, where $M << N$.

NMMD integrates over time both types of parameters, $\mathbf{q}(t)$ and $\mathbf{x}(t)$, whereas classical MD integrates $\mathbf{x}(t)$ only. For the numerical integration, NMMD uses the Velocity Verlet integrator, which has good numerical stability and is commonly used in classical MD-based approaches. The integration of parameters

$\mathbf{x}$(t) is given by :

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + \dot{\mathbf{x}}(t)\Delta t + \frac{1}{2}\ddot{\mathbf{x}}(t)\Delta t^2 \tag{6}$$

$$\dot{\mathbf{x}}(t+\Delta t) = \dot{\mathbf{x}}(t) + \frac{\ddot{\mathbf{x}}(t) + \ddot{\mathbf{x}}(t+\Delta t)}{2}\Delta t, \tag{7}$$

and the integration of parameters $\mathbf{q}(t)$ is given by:

$$\mathbf{q}(t+\Delta t) = \mathbf{q}(t) + \dot{\mathbf{q}}(t)\Delta t + \frac{1}{2}\ddot{\mathbf{q}}(t)\Delta t^2 \tag{8}$$

$$\dot{\mathbf{q}}(t+\Delta t) = \dot{\mathbf{q}}(t) + \frac{\ddot{\mathbf{q}}(t) + \ddot{\mathbf{q}}(t+\Delta t)}{2}\Delta t, \tag{9}$$

where $\dot{\mathbf{x}}(t)$ and $\ddot{\mathbf{x}}(t)$ respectively are the first and second derivatives of $\mathbf{x}(t)$ with respect to time, $\dot{\mathbf{q}}(t)$ and $\ddot{\mathbf{q}}(t)$ respectively are the first and second derivatives of $\mathbf{q}(t)$ with respect to time, and Δt is the time step of the numerical integration.

Recalling the Newton's second law of motion for $\mathbf{x}(t)$, $\ddot{\mathbf{x}}(t)$ is given by:

$$\ddot{\mathbf{x}}(t) = \mathbf{M}_x^{-1}\mathbf{F}_x(t), \tag{10}$$

where the force $\mathbf{F}_x(t)$ is the negative gradient of the potential energy $U(t)$ with respect to the atomic positions $\mathbf{x}(t)$ and $\mathbf{M}_x^{-1}$ is the inverse of the mass matrix $\mathbf{M}_x$ that is a diagonal matrix with the atomic masses m_x^i $(i = 1, ..., N)$ as the entries of the diagonal.

In NMMD, we consider that the atomic displacement in equation (5) due to normal modes, more precisely $\mathbf{q}(t)$ because $\mathbf{A}$ does not change with time in this equation, follows the motion equation:

$$\ddot{\mathbf{q}}(t) = \mathbf{M}_q^{-1}\mathbf{F}_q(t), \tag{11}$$

where $\mathbf{F}_q(t)$ is the negative gradient of the potential energy $U(t)$ with respect to $\mathbf{q}(t)$ and $\mathbf{M}_q^{-1}$ is the inverse of the matrix $\mathbf{M}_q$ that is a diagonal matrix with $m_q^i = m_q$ $(i = 1, ..., M)$ as the entries of the diagonal. In this equation, m_q^i could be interpreted as a "mass" assigned to the i-th normal-mode amplitude $(i = 1, ..., M)$. Different values of such "masses" for different normal modes could be used for giving more or less weights to some of the normal modes: a high value of this parameter would result in a low

contribution of the normal mode to the motion while its low value would result in a strong contribution of the normal mode. Here, we assigned the same value of this parameter for each normal mode ($i.e.$, $m_q^i = m_q, i = 1, ..., M$), to avoid imposing any prior information about the individual contribution of the different normal modes. During the simulation, NMMD automatically determines these contributions by estimating the amplitude of each normal mode. Therefore, equation (11) can be written as follows:

$$\ddot{\mathbf{q}}(t) = \frac{1}{m_q}\mathbf{F}_q(t). \tag{12}$$

Regarding the choice of the mass value m_q to assign to all the normal modes, it should be noted that a too large value of m_q slows down the integration of the normal-mode amplitudes $\mathbf{q}(t)$, whereas its too small value makes the system unstable. This parameter can be tuned manually to maximize the speed while ensuring the stability of the system. For all the structures presented in the study, we have obtained satisfactory speed and stability results using $m_q = 10$. The diversity of these structures suggests that this value shall give satisfactory results in the majority of cases. Therefore, we propose to use $m_q = 10$.

While $\mathbf{F}_x(t)$ is implemented in the integration scheme of any standard MD simulation software, the computation of $\mathbf{F}_q(t)$ from scratch is not trivial as the potential energy is the sum of multiple potentials including the biasing potential. Besides, the gradient computation is usually highly optimized in MD software as it corresponds to the main computational consumption. In this context, it is convenient to see that $\mathbf{F}_q(t)$ can be expressed as a function of the force vector on MD-induced atomic coordinate displacement $\mathbf{F}_x(t)$ by using the following chain rule :

$$\begin{aligned}
\mathbf{F}_q(t) &= -\frac{\partial U}{\partial \mathbf{q}}(t) \\
&= -\frac{\partial \mathbf{r}}{\partial \mathbf{q}}(t) \cdot \frac{\partial U}{\partial \mathbf{r}}(t) \\
&= -\frac{\partial \mathbf{r}}{\partial \mathbf{q}}(t) \cdot \left(\frac{\partial \mathbf{r}}{\partial \mathbf{x}}(t)\right)^{-1} \cdot \frac{\partial U}{\partial \mathbf{x}}(t) \\
&= \mathbf{A} \cdot \mathbf{F}_x(t).
\end{aligned} \tag{13}$$

NMMD method is implemented in GENESIS software and uses $\mathbf{F}_x(t)$ of GENESIS to implement $\mathbf{F}_q(t)$. More precisely, NMMD was implemented in GENESIS 1.4 by modifying EMfit method, which performs MD-based fitting without normal modes Orzechowski and Tama (2008); Miyashita et al. (2017). EMfit

minimizes the total potential consisting of the classical MD-based potential and a biasing EM-map potential, where the contribution of the biasing EM-map potential is defined by the force constant according to equation (1). We modified EMfit to include NMA-based atomic displacement. For NMA, we use ENM to compute the normal modes. To speed up the calculation of normal mode, we used the Rotation Translation Block (RTB) method Tama et al. (2000) as implemented in ElNemo Suhre and Sanejouand (2004). With the RTB method, the dimension of the Hessian matrix that should be diagonalized is reduced by splitting the atomic structure into blocks, each block containing a selected number of residues and having 6 degrees of freedom (3 rotations and 3 translations).

It should be noted that the matrix of normal modes $\mathbf{A}$ is constant over time in the equations above. Indeed, in our experiments, we have not found useful to recalculate normal modes during the fitting with NMMD. If one wants to recalculate normal modes, the simulation should be reinitialized using $\mathbf{r}_0$ (equation (5)) set to be the best-fitting conformation from the previous run as well as using the corresponding normal-mode matrix of that new $\mathbf{r}_0$.

2.4 REUS: Replica Exchange Umbrella Sampling

The choice of the biasing force constant k (equation (1)) may strongly affect the fitting results. If k is high, the given EM map will influence too much the fitting direction and the simulation will rapidly reach high values of CC, with a risk of over-fitting the EM map and deriving atomic models of lower quality. Inversely, if k is low, the simulation will take more time to dock the atomic structure into the EM map, with a risk of getting trapped into local minima during the process.

The force constant can be tuned by REUS approach. In this approach, multiple parallel MD simulations are performed (called replicas), each with a different value of force constant k. During the fitting process, the force constants are periodically exchanged to adjust the optimal k value for each replica. In the REUS method, the probability W for the replica exchange is given by:

$$W(X \rightarrow X') = \begin{cases} 1 & \text{for } \Delta \leq 0 \\ \exp(-\Delta) & \text{for } \Delta > 0 \end{cases} \tag{14}$$

$$\Delta = \frac{1}{k_B T}(k_n - k_m)(CC^j - CC^i) \tag{15}$$

where CC^i is the CC value for the i-th replica and k_n is the n-th force constant, T the temperature and k_B the Boltzmann constant.

This algorithm tends to exchange highest force constants to replicas with the highest CC. The effect of these exchanges is to increase gradually the force constant and improve the results of the fitting. As EMfit, NMMD can be run with REUS procedure implemented in GENESIS to automatically adjust the value of the force constant Miyashita et al. (2017).

When searching larger conformational spaces, MD-based flexible fitting has been shown to yield better results with than without REUS. However, REUS is computationally expensive when used with all-atom MD simulations in explicit solvent Miyashita et al. (2017); Kulik et al. (2021). Therefore, it is usually used with coarse-grained models, such as Cα-based Go model Miyashita et al. (2017). In the same context, a multiscale approach has been proposed, which combines a Cα-based Go model with an implicit-solvent all-atom model via a targeted MD between the fitted coarse-grained structure and the all-atom structure Kulik et al. (2021). In our study, REUS was used with all-atom MD simulations in vacuum, which reduces the computational cost while maintaining the quality of the structure during the fitting.

3 RESULTS

In this section, we show the fitting performance of NMMD (MD with NMA). We compare the NMMD results with the results of EMfit (MD without NMA) on which NMMD is based. The two methods were compared regarding computational time and conformational sampling in GENESIS 1.4 using REUS procedure. The simulations were performed on two Intel Xeon Silver 4214 CPUs (24 cores at 2.60 GHz per CPU) with 64 GB RAM, using a single core per replica.

The comparison of the two methods is shown using synthetic data sets of four complexes (LAO, AK, LF, and EF2) and experimental data sets of two complexes (p97 and ABC). The size of these complexes goes from small (26 kDa for LAO) to large (542 kDa for p97) (Table 1). Each synthetic data set consists of two atomic structures corresponding to two different molecular conformations. One of the two atomic structures was used as the initial conformation to fit the EM map simulated from the other atomic structure (target conformation). Each experimental data set contains one atomic structure and one EM map, which correspond to two different conformations and were used as the initial and target conformations for fitting, respectively. Additionally, each experimental data set contains one atomic model derived from the given EM map, which was used for comparison with the atomic models obtained by fitting with NMMD and

EMfit. All atomic structures and cryo-EM maps used in this study are available in the Protein Data Bank (PDB) and Electron Microscopy Data Base (EMDB) databases. Their PDB and EMDB codes are provided in Table 1.

NMMD and EMfit are compared using the following two measures: 1) CC between the target EM map and the map simulated from the fitted atomic conformation (normalized CC, with values between 0 and 1); and 2) Root Mean Square Deviation between the atomic positions in the target and fitted atomic conformations (RMSD, in angstroms). In the case of synthetic EM maps, the atomic structure used to simulate the target-conformation EM map is the target (ground-truth) atomic conformation that should be retrieved by fitting of an initial atomic conformation into the simulated EM map. In the case of experimental EM maps, such unique ground-truth atomic conformation is unavailable but existing fitted atomic models are available in the PDB and used here as the target atomic conformations for the RMSD calculations.

In the rest of this article, EMfit is usually referred to as MD fitting.

Biomolecular complex	**Initial PDB**	**Target PDB**	**Target EMDB**	**Weight (kDa)**
LAO binding protein	1LST	2LAO	-	26
Adenylate kinase	4AKE	1AKE	-	47
Lactoferrin	1LFG	1LFH	-	77
Elongation factor 2	1N0V	1N0U	-	186
ABC exporter	6RAF	6RAH	4775	150
p97 ATPase	5FTM	5FTN	3299	542

Table 1. PDB and EMDB codes of the available structural data used for the tests of NMMD fitting method and its comparison with MD fitting method EMfit in GENESIS.

3.1 Synthetic EM maps of LAO, AK, LF, and EF2

To synthesize an EM map of a complex, we reproduced the reconstruction procedure used in single particle analysis. First, we obtained a high-resolution density map (map size: $100 \times 100 \times 100$ voxels, voxel size: $1.5 \times 1.5 \times 1.5$ Å) by converting a given atomic structure (target conformation for fitting) using atomic scattering factors Peng et al. (1996). Then, a library of 2D projections of the obtained map was calculated using a quasi-uniform distribution of the projection directions determined by a tilt-angle step of 5°, which produced 1647 projection images. To simulate the effect of the electron microscope, we applied on the images a contrast transfer function with a set of parameters that simulate a 200 kV microscope with a magnification of 50,000 and a defocus of -0.5 µm, and added Gaussian noise resulting in the Signal-to-Noise Ratio (SNR) of 0.5, following the method of Velazquez-Muriel et al. (2003). The SNR

of 0.5 was chosen as the noise level typically encountered in decent-quality single particle class averages (considering 1647 projection images as the class averages). Finally, a synthetic EM map was reconstructed from the images (using Fourier interpolation method) and low-pass filtered to 5 Å. Such procedure for EM map synthesis and the fact that NMMD and MD fitting methods use a different method for converting atoms into density (3D Gaussian functions to obtain density volumes at each iteration of the fitting) make fitting more difficult. All steps of the procedure were performed using Xmipp 3 command-line programs De la Rosa-Trevín et al. (2013).

3.2 Experimental EM maps of p97 and ABC

For ABC, fitting was performed using the following two conformations: 1) extracellular-side closed (intracellular-side open) conformation, given by PDB:6RAF (an atomic model derived from a cryo-EM map of this conformation); and 2) extracellular-side open (intracellular-side closed) conformation, given by EMD-4775 (a cryo-EM map of this conformation at 2.8 Å resolution). For p97, fitting was performed using the following two conformations: 1) conformation with N-terminal domains in ”down” position, given by PDB:5FTM (an atomic model derived from a cryo-EM map of this conformation); and 2) conformation with N-terminal domains in ”up” position, given by EMD-3299 (a cryo-EM map of this conformation at 3.3 Å resolution). Both cryo-EM maps used as the target conformations for fitting were down-sampled to the size of $128 \times 128 \times 128$ voxels to speed up the computation.

In our preliminary experiments, we identified mode 10 of the 5FTM p97 structure as the mode that moves up and down the N domains of all six monomers while preserving the overall p97 symmetry. Our preliminary experiments with ABC showed that mode 9 of the 6RAF ABC structure is one of the modes that contribute to opening and closing motions of ABC. A description of complex conformational transitions, such as those observed in the data used in this article, requires using more than one normal mode. We show below that NMMD achieves good fitting results for these two molecular complexes using a small set of normal modes (ten lowest-frequency normal modes, i.e., modes 7-17) in combination with MD.

3.3 Parameters for running NMMD and MD fitting methods

For each of the six molecular complexes (Table 1), we ran 16 replicas of both NMMD and MD fitting methods with a time step of 2 femtoseconds. The value of the force constant for each complex was adjusted with REUS, from a linear distribution of values in the range determined in preliminary experiments with

each complex, as the optimal range is specific to each system (5000-10000 kcal/mol for AK and LAO, 10000-30000 kcal/mol for LF, EF2 and ABC, 50000-100000 kcal/mol for p97). For all complexes and both fitting methods, all-atom simulations including hydrogen atoms were preformed using CHARMM 36 force fields and the temperature of 300 K regulated with the Langevin thermostat with friction coefficient of 1 ps^{-1}. EMfit and NMMD do not search for rotations and translations during the fitting, meaning that the atomic structure and the EM map must be aligned using rigid-body transformations prior to fitting. The rigid-body transformations were performed using ChimeraX function "Fit in map" Goddard et al. (2018). To simulate the map from the fitted atomic structure using Gaussian functions, the standard deviation of the Gaussian functions was set to 2.5, resulting in a map of 5 Å resolution (the simulated map to be compared with the given map using the CC). NMMD fitting was run using normal modes 7-17, which are ten lowest-frequency non-rigid-body modes that usually describe collective motions. Normal modes were calculated using the RTB block size of 10 residues and the ENM with the interaction cutoff radius of 8 Å (the cutoff defining the radius beyond which the nodes of the ENM are not connected with elastic springs).

3.4 Inclusion of normal modes speeds up fitting

For each of the six molecular complexes (Table 1), the mean and the standard deviation of the CC and the RMSD over all replicas of NMMD and MD are plotted in Figure 1. For the replica that reached the lowest RMSD, Table 2 shows the achieved values of RMSD and CC, the total execution time (in the case of NMMD, the total time also includes the time required for computing normal modes), the convergence time (the time until the RMSD starts to change by less than 1% between the successive steps), and the measures of the obtained atomic structure quality (MolProbility score).

One can note that, generally, the CC increases (the RMSD decreases) faster with NMMD than with MD (Figure 1). The faster convergence of NMMD can be explained by the use of normal modes, as this is the main difference between the two fitting methods. Therefore, the addition of normal modes to MD-based fitting improves the sampling efficiency. The computation and integration of normal modes induces a small computational cost, included in the total execution time reported in Table 2. This computational cost is insignificant compared to the speed increase induced by the inclusion of normal modes. Indeed, the convergence time is, in average, around 40% shorter for NMMD than for MD (Table 2). Additionally, it should be noted that NMMD generally reaches lower RMSD values than MD (in all tested cases of

https://doi.org/10.1016/j.jmb.2022.167483

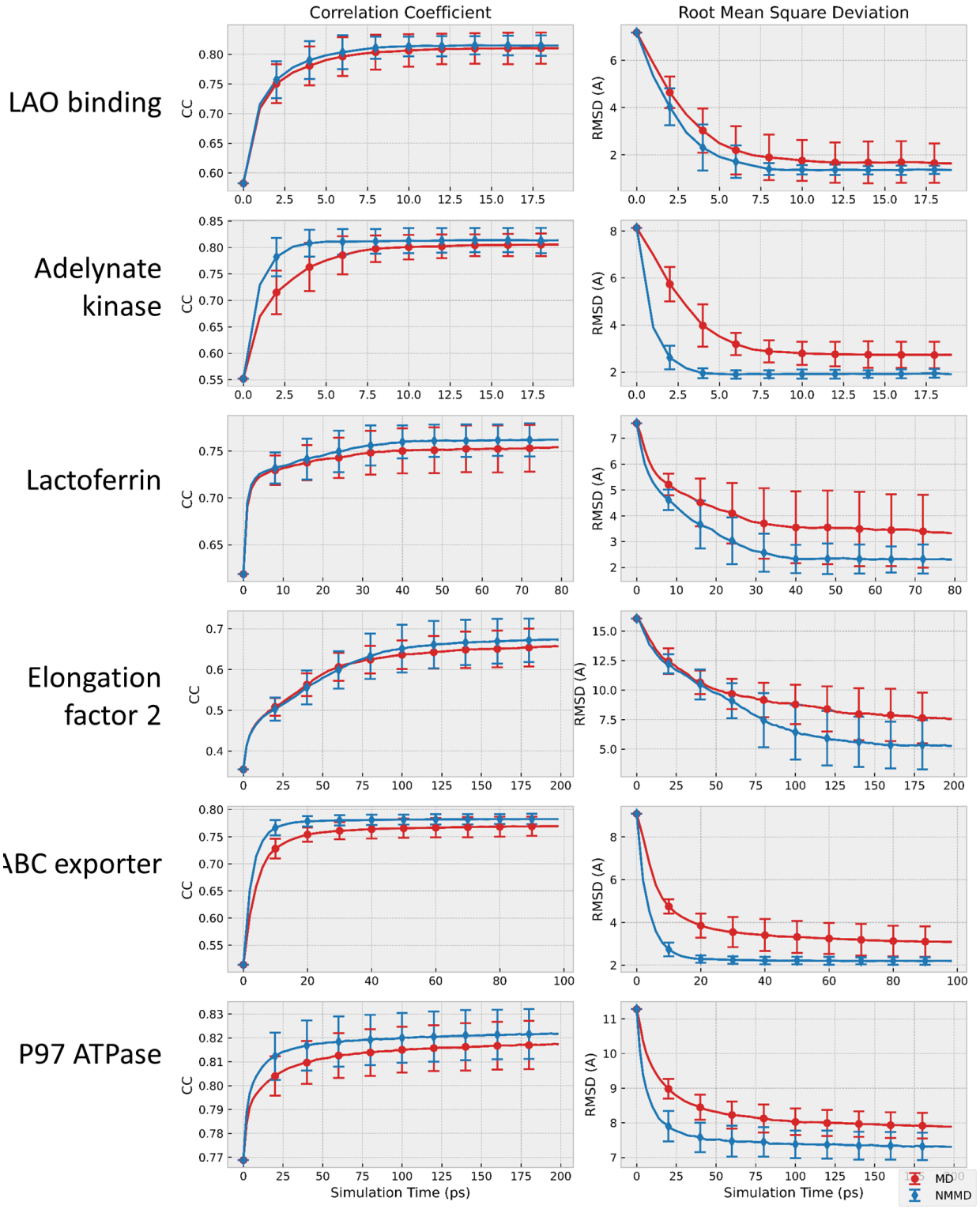


Figure 1. Mean (curve) and standard deviation (error bar) of CC (left panels) and of RMSD (right panels) as a function of simulation time, for 16 replicas of NMMD fitting (blue) and the MD fitting (red). The results are shown for four synthetic data sets (LAO binding protein, Adenylate kinase, Lactoferrin, and Elongation factor 2) and two experimental data sets (ABC exporter and p97 ATPase). See Table 2 for additional information regarding the fitting results.

Biomolecular complex	Fitting method	CC	RMSD (Å)	Total time (min)	Convergence time (min)	Speed increase	MolProbity score
LAO binding protein	MD	0.83	1.04	10.5	6.2	+32%	2.84
	NMMD	0.83	1.06	10.6	4.2		2.79
Adenylate kinase	MD	0.83	1.88	9.8	5.3	+62%	2.61
	NMMD	0.84	1.58	10.0	2.0		2.59
Lactoferrin	MD	0.79	1.57	132.2	122.2	+46%	2.33
	NMMD	0.79	1.45	133.2	64.9		2.37
Elongation factor 2	MD	0.73	3.67	402.5	358.2	+10%	2.67
	NMMD	0.75	2.13	405.0	322.6		2.49
ABC exporter	MD	0.79	2.18	229.9	179.3	+66%	2.21
	NMMD	0.79	1.86	235.3	61.1		2.21
p97 ATPase	MD	0.83	7.17	1754.8	1526.7	+17%	2.00
	NMMD	0.81	6.57	1796.0	1257.2		1.99

Table 2. Comparison of NMMD and MD fitting results for each of four synthetic data sets (LAO binding protein, Adenylate kinase, Lactoferrin, and Elongation factor 2) and two experimental data sets (ABC exporter and p97 ATPase). The table shows the achieved CC and RMSD values for the replica with the lowest achieved value of RMSD (from 16 replicas), the total time of execution (for NMMD, this time includes the time required for computing normal modes), the convergence time (the time until the RMSD starts to change by less than 1% between the successive steps), the speed increase between NMMD and MD convergence time in percentage, and the measure of the obtained atomic structure quality (MolProbility score). See Figure 1 for the CC and RMSD plots over the simulation for all 16 replicas.

molecules except for LAO, where the two methods reached almost the same value of RMSD, Table 2). In the case of EF2, the speed increase is less important than for the other structures, but NMMD reaches a much lower RMSD value (2.13 Å) than MD (3.67 Å), probably thanks to this extra time compared to other structures.

3.5 Inclusion of normal modes improves accuracy of fitting

We observe (Table 2) that NMMD and MD retrieved the target structure (relatively low RMSD) for five out of six complexes (for all except for the experimental data of p97). As noted earlier, NMMD achieves a lower RMSD value than MD in all cases except in the case of LAO, where the achieved RMSD is almost the same for the two methods. In some cases, the achieved RMSD values differ more between the two methods (e.g, in the synthetic EF2 case and the experimental p97 case, Table 2).

The best achieved RMSD for the synthetic case of EF2 is 3.67 Å for MD and 2.13 Å for NMMD. The target and fitted conformations are different in the case of MD but very similar in the case of

NMMD, meaning that NMMD produced better fitting than MD. For a visual assessment of the obtained conformations, we present the results of EF2 fitting in Figure 2. NMMD retrieved the right conformation (Figure 2c) while MD was able to retrieve the global target shape but not the details of the structure (some secondary structure elements are different or missing, Figure 2b). To assess the quality of the obtained atomic models, we evaluated their MolProbity scores Davis et al. (2007) (Table 2). The Molprobity score is commonly used to assess quality of structures and corresponds to a combination of the number of atomic clashes, rotamer evaluations, and Ramachandran evaluations. In the case of EF2, the Molprobity score is significantly lower for NMMD than for MD, indicating that the conformation obtained with NMMD has better quality than the one obtained with MD. Generally, the quality of the atomic models obtained with NMMD is comparable to the quality of the models obtained with MD and, in some cases, NMMD produces models of better quality (see Molprobity score columns of Table 2).

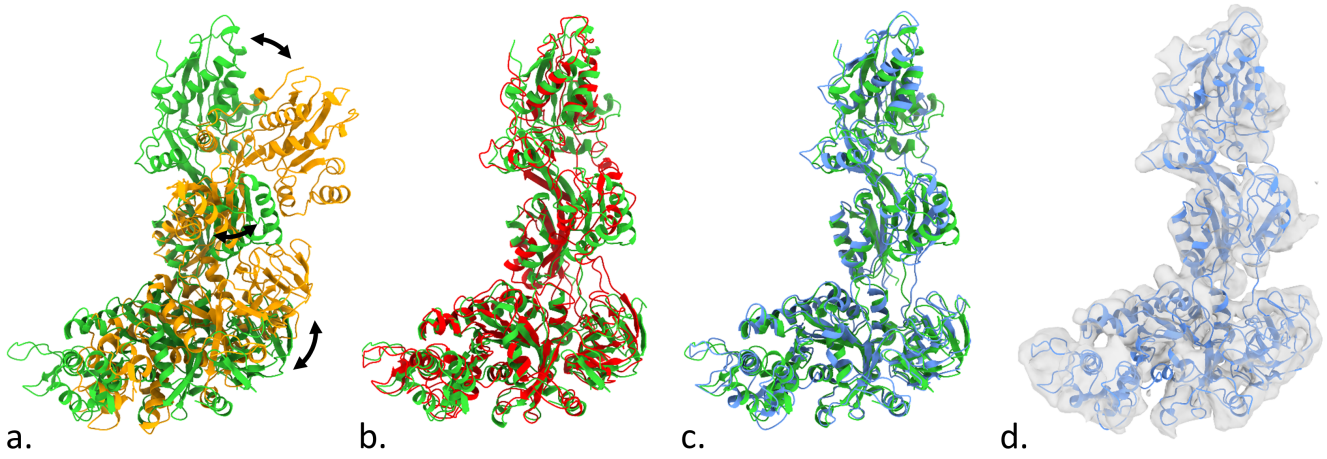


Figure 2. Flexible fitting of a synthetic EM map of Elongation factor 2 for the NMMD and MD replicas reaching the lowest RMSD value. Target structure (green) is overlapped with the initial structure (a), MD fitted structure (b), NMMD fitted structure (c). NMMD-fitted structure is overlapped with the target EM map (d). The black arrows show the main conformational changes.

Figure 3 shows the results of fitting in the case of ABC experimental data. Despite the large conformational rearrangements of ABC, NMMD successfully retrieved the target conformation (RMSD $= 1.86$ Å), with less clashes than MD (Table 2). The MD achieved a slightly worse RMSD than NMMD (by around 0.3 Å) (Table 2).

Concerning the experimental p97 case, we observe that the achieved RMSD is high for both methods (RMSD > 6 Å) but still smaller for NMMD than for MD. These results indicate that the fitted structures are largely different from the target structure. Figure 4 presents the p97 fitting results. We observe that

https://doi.org/10.1016/j.jmb.2022.167483

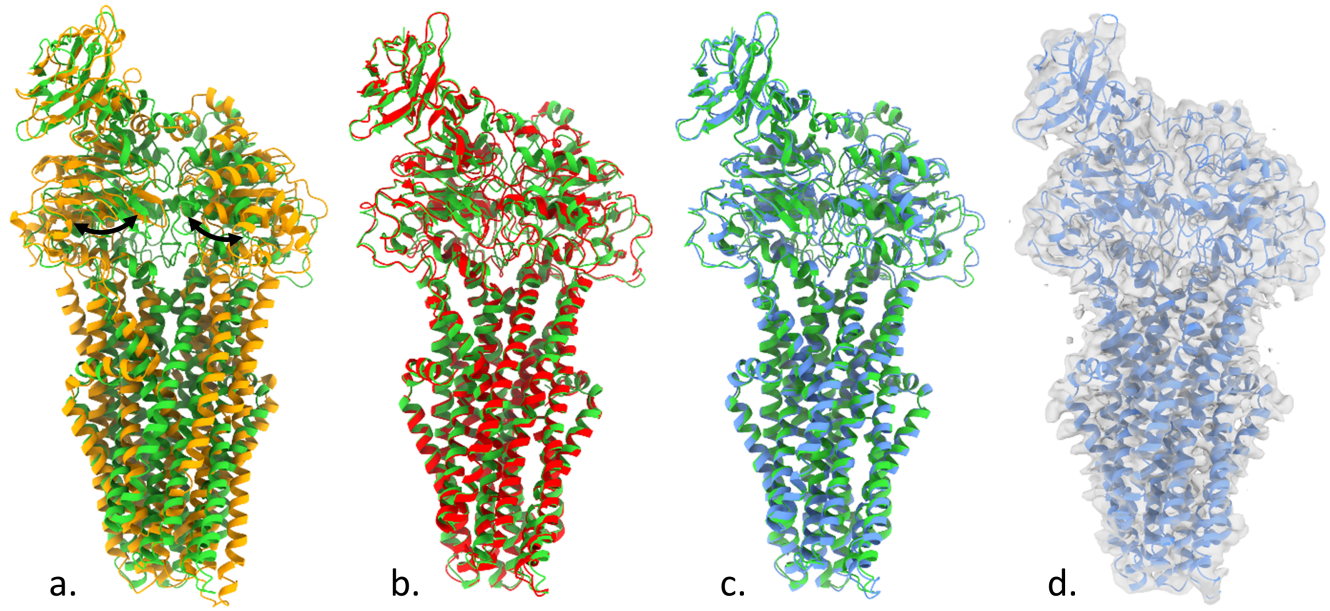


Figure 3. Flexible fitting of an experimental EM map of ABC exporter for the NMMD and MD replicas reaching the lowest RMSD value. Target structure (green) is overlapped with the initial structure (a), MD fitted structure (b), NMMD fitted structure (c). NMMD-fitted structure is overlapped with the target EM map (d). The black arrows show the main conformational changes.

both MD and NMMD were able to lift the N domain up (the N-domain motion is shown by black arrow in Figure 4a) and to fit the protein globally. However, the final conformations obtained by both methods are different from the target conformation. Figure 4 shows significant conformational differences in the N-terminal domain of these three structures (Figure 4b,c).

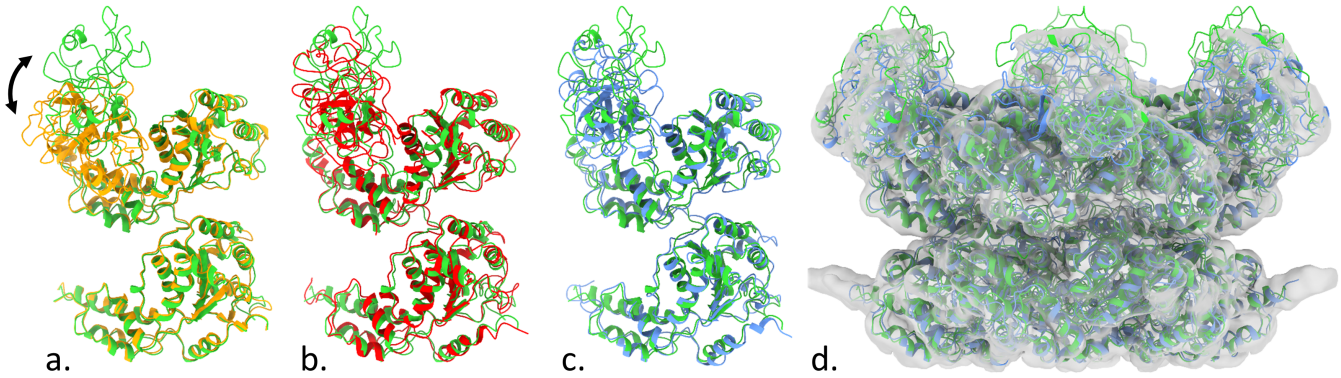


Figure 4. Flexible fitting of an experimental EM map of p97 ATPase for the NMMD and MD replicas reaching the lowest RMSD value (a single monomer is shown for better visibility in a-c). Target structure (green) of one monomer is overlapped with the initial structure (a), MD fitted structure (b), and NMMD fitted structure (c). Target (green) and NMMD-fitted (blue) structures of all six monomers are overlapped with the target EM map (d). The black arrows show the main conformational changes.

To assess and compare the local resolutions of the two fitted experimental cryo-EM maps (p97 and ABC), we employed MonoRes Vilas et al. (2018), an automatic method to determine the local resolution of an EM map (Figure 5). We can note that the ABC map has a similar local resolution for the whole structure, which is around 3 Å (Figure 5b). On the contrary, the p97 map has a much lower local resolution in the regions that correspond to the N-terminal domains (top parts of the map) than in the other regions (Figure 5a). More precisely, the local resolution in the major part of the p97 map is around 3 Å, but it is between 7 and 15 Å in the regions corresponding to the N domains. The differences in local resolutions of the cryo-EM maps of ABC and p97, especially for the N domains of p97, may explain the differences in the fitting results obtained for these two test cases.

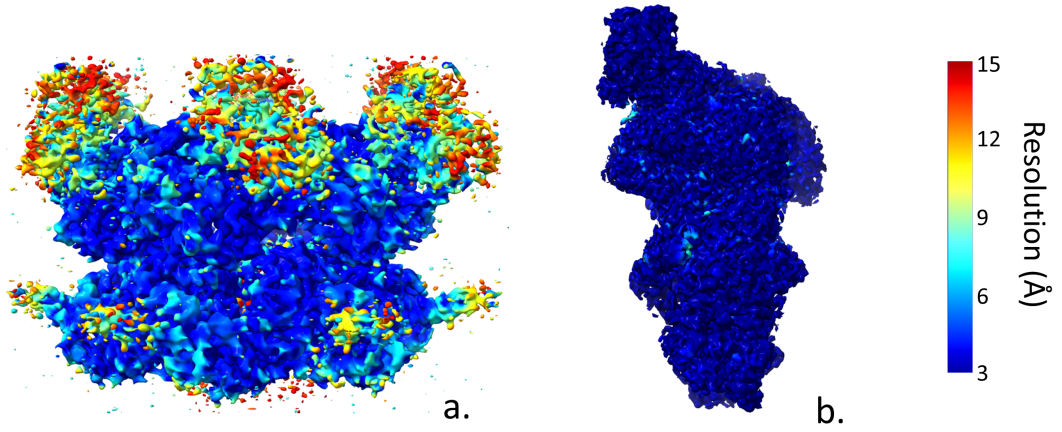


Figure 5. Local resolution obtained with MonoRes for the experimental cryo-EM maps fitted with MD and NMMD in this article. (a) Local resolution of p97 ATPase. (b) Local resolution of ABC exporter.

3.6 Inclusion of normal modes improves REUS adjustment of the force constant

The atomic models of the synthetic and experimental EM maps obtained by 16 replicas of NMMD and MD were analyzed by Principal Component Analysis (PCA). Figure 6 shows low-dimensional (2D) conformational spaces of LF, EF2, ABC and p97 (two synthetic and two experimental data cases), determined by the first two principal axes. In the case of LF, EF2 and ABC, the NMMD replicas (blue dots) are more concentrated around the target conformation (green dot), whereas the MD replica (red dots) are more scattered. This result indicates that REUS force constant adjustment produces more consistent fitting results with NMMD than with MD. In the case of p97, both MD and NMMD replicas are spread

between the initial and target conformations, which is consistent with our previous observation that both approaches converged to a conformation that is different from the target conformation.

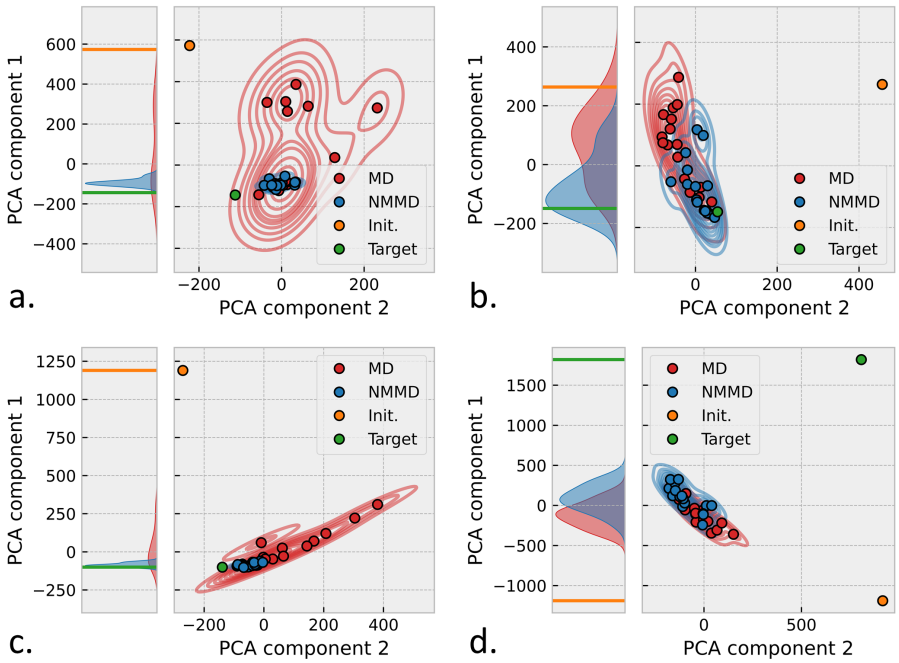


Figure 6. Two-dimensional conformational spaces determined by PCA of the atomic models from 16 replicas of MD (red) and NMMD (blue), together with the target structure (green) and the initial structure (orange), for Lactoferrin (a), Elongation factor 2 (b), ABC exporter (c), and p97 ATPase (d). The one-dimensional plots at the left side show the data distribution along the first PCA axis.

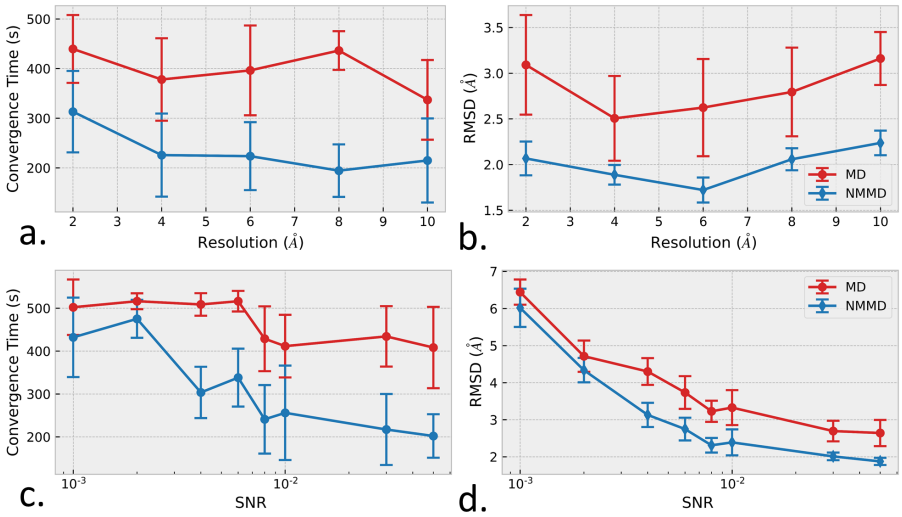


Figure 7. Impact of noise and map resolution on the results of fitting with MD (red curves) and with NMMD (blue curves). (a,c) Evolution of the convergence time with respect to the resolution (a) and the SNR (c). (b,d) Evolution of the minimum RMSD with respect to the resolution (b) and the SNR (d).

3.7 Impact of noise and map resolution

Additional tests of NMMD and MD sensitivity were performed with synthetic EM maps to study the impact of noise and map resolution. To this end, we performed several fitting processes on AK by varying the SNR of the images used for 3D reconstruction and the resolution of a synthetic map of AK. For the study of the noise impact, we employed the same process to synthesize the maps as the one described in section 3.1, with the SNR of images in the range from 0.5 to 0.001. Note that it is almost impossible to see the particle in the synthetic map for the SNR of 0.001. For the study of the resolution impact, we applied a low-pass filter onto the high-resolution synthetic map obtained by converting the AK atomic structure into density as described in section 3.1. The cut-off frequency of the low-pass filter was varied from 2 to 10 Å, with a step of 2 Å. For each of the resulting maps, we run ten NMMD and ten MD fitting processes with a force constant of 10000 kcal/mol (ten runs with random initial velocities). The average resulting convergence time and minimum RMSD are shown in Figure 7a,b for the resolution variation, respectively, and in Figure 7c,d for the SNR variation, respectively.

Figure 7a,b shows similar RMSD and convergence time values over the tested resolution range for each of the two methods. Figure 7c,d shows that the RMSD and the convergence time tend to increase with the decrease in the SNR. Also, Figure 7 shows that both RMSD and convergence time values are lower for NMMD than for MD for the entire tested resolution and SNR ranges, meaning that NMMD is faster, more accurate, and has a better robustness to noise than MD over the entire tested resolution and SNR ranges.

4 DISCUSSION AND CONCLUSION

In this article, we introduced NMMD, a new flexible fitting method for cryo-EM, which combines NMA and MD. Given an atomic structure and a cryo-EM map to fit, NMMD simultaneously estimates global atomic displacements based on NMA and local atomic displacements based on MD, by their simultaneous numerical integration.

A previous attempt to combine MD and NMA was based on an alternation between a stochastic estimation of a linear combination of normal modes (the combination that moves the structure in the direction of the CC increase) and a short MD simulation initiated with this normal-mode combination Costa et al. (2020). This normal-mode combination was updated with a time interval (delay) of the order of a picosecond and each update required a new calculation of normal modes and a new stochastic estimation of their amplitudes. NMMD, proposed in this article, is the first method that combines the displacements based on

MD and NMA via their simultaneous integration. In NMMD, the integration of normal-mode amplitudes at each time step is based on an analytic expression for the gradient of the potential energy, which ensures accurate and fast results. Furthermore, our NMMD experiments have shown that normal modes can remain constant during the fitting while normal-mode amplitudes are updated at each step with no delay, which saves computing time as multiple NMA runs are not required.

The NMMD is implemented as part of the open source software GENESIS version 1.4 by modifying its MD-based fitting method EMfit. In the tests described in this article, we compared NMMD with EMfit using a variety of maps (synthetic and experimental EM maps of six molecular complexes, with different noise levels and resolutions). These tests showed that the combination of MD and NMA has generally better performance than MD alone. Both NMMD and EMfit were run with REUS sampling procedure to automatically adjust the value of the force constant.

The results showed that adding normal modes to MD-based fitting makes the fitting faster by around 40% (in average). In the study shown, the computing time was balanced between the computing of the biasing potential and the computing of the classical MD potential (e.g., for AK, NMMD took 218 s to compute the biasing potential and 381 s to compute the CHARMM-based potential). Also, the results showed that adding normal modes to MD-based fitting improves the fitting accuracy (the obtained RMSD values were lower for NMMD than for EMfit in all cases except in one where the achieved RMSD value was almost the same for the two methods).

In contrast to an NMA-based fitting, which generally induces structural distortions, NMMD produces biochemically meaningful structures whose quality can, in some cases, even be slightly better than the quality of the structures obtained with an MD-based fitting (observed in this study using MolProbity scores). In this context, a simultaneous use of NMA and MD is an advantage of NMMD over a two-step approach in which an NMA-based fitting is followed by an MD-based fitting. In general, such two-step strategies are less likely to produce good quality structures, as the NMA-based fitting may create undesirable distortions of the structure that are impossible to correct with the MD-based flexible fitting.

Interestingly enough, adding normal modes to MD-based fitting improves the selection of the value of the force constant by REUS. Our NMMD approach could be used in a coarse-to-fine multiresolution scheme, which could further enhance the REUS-based sampling strategy. Such schemes do not necessarily reduce the computing time, but increase robustness against noise and minimize the risk of getting trapped into

local minima. A multiresolution scheme could be to generate different resolutions of the same EM map by its low-pass filtering or down-sampling and, then, run NMMD to perform the EM map fitting at the different resolutions by propagating and refining the solution from a coarser resolution level to the next finer level, to fit the map first globally than locally, in each of the different replicas.

In this study, no prior information was imposed on the contribution of normal modes, by using the same "mass" value (m_q) for all the normal modes. Also, the value of this parameter was manually tuned to $m_q = 10$, which was the value that ensured good speed and stability for all the experiments presented here. Based on the diversity of the structures used in our experiments, it can be expected that this value gives good results in most cases. However, finding the optimal value of this parameter for each particular case may increase the NMMD efficiency further. Future developments could include an automated optimization of this parameter, for instance through REUS-based parameter estimation during the fitting.

NMMD uses the CC to measure the similarity between the EM map and the map simulated from the atomic structure. The lower the EM-map resolution (global and local), the higher the uncertainty of the EM fitting with an atomic structure will be, independently of the similarity measure used. Multiple local minima in the potential (the potential that guides the fitting based on the chosen similarity measure) can lead the fitting to a local-minimum conformation that is different from the global-minimum conformation. We addressed this issue by running multiple parallel replicas with REUS. The distribution of the final values of the similarity measure (here, CC) obtained by the replicas (Figure 1) is informative of the fitting uncertainty: large CC variance suggests high uncertainty (different conformations obtained by different replicas), whereas small CC variance suggests low uncertainty (similar conformations obtained by different replicas). Beside the use of the CC variance over the different replicas, the uncertainty of the fit could be evaluated by observing the distances between the structures from the different replicas in a PCA-based low-dimensional space (Figure 6). Another way to visualize distances between the structures from different replicas (in a low-dimensional space) could be to perform a multivariate analysis of a matrix of pairwise RMSD-based distances between these structures.

The GENESIS software includes an efficient parallelization strategy for EMfit, based on a combination of MPI and OpenMP, which is well suited for computation over multiple nodes. NMMD in GENESIS offers the same parallelization strategy as EMfit. In this study, we decided not to use this parallelization scheme,

but a parallelization over replicas, using a single core per replica. However, for more time demanding fitting tasks, one could use NMMD with a larger number of cores per replica.

NMMD software code will be publicly available as part of ContinuousFlex plugin Harastani et al. (2020) (https://github.com/scipion-em/scipion-em-continuousflex) for Scipion 3 De la Rosa-Trevín et al. (2016), including a graphical user interface giving the user the opportunity to easily use NMMD on parallel systems.

Deep Mind's AlphaFold2 is a demonstration that AI-based methods for protein structure prediction from protein sequence are able to produce similar models to those that can be obtained with the gold-standard experimental methods. However, they are still limited to static structures and the protein dynamics is still studied by experimental or hybrid methods, such as cryo-EM single particle analysis in conjunction with flexible fitting of existing atomic models into cryo-EM maps. AI-based methods could be useful for obtaining the initial models for fitting, when such models are unavailable. Yet, the use of AI-based methods to predict challenging structures, such as large (multisubunit) and flexible complexes remains to be demonstrated.

ACKNOWLEDGMENT

We acknowledge the support of the French National Research Agency — ANR (ANR-19-CE11-0008-01 and ANR-20-CE11-0020-03 to SJ), cooperation between the CNRS and the University of Melbourne (The Melbourne-CNRS Network, CNRS PRC 2889 to SJ and IR), The University of Melbourne start-up fund (to IR), and access to HPC resources of CINES and IDRIS granted by GENCI (A0100710998, A0070710998, AP010712190, AD011012188 to SJ). This work was supported by FOCUS for Establishing Supercomputing Center of Excellence to FT, Nagoya University Fund to FT, and JSPS KAKENHI Grant Number JP20H05453 to OM.

5 CREDIT AUTHOR STATEMENT

Rémi Vuillemot: Conceptualization, Methodology, Software, Investigation, Validation, Writing- Original draft preparation. Osamu Miyashita: Methodology, Validation, Writing - Review & Editing. Florence Tama: Methodology, Validation, Writing - Review & Editing. Isabelle Rouiller: Validation, Writing - Review & Editing, Funding acquisition. Slavica Jonic: Conceptualization, Methodology, Validation, Writing - Review & Editing, Project administration, Funding acquisition.

6 DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary files. NMMD software code is publicly available on Github (https://github.com/mms29/nmmd) and will also be available as part of ContinuousFlex plugin (https://github.com/scipion-em/scipion-em-continuousflex) for Scipion 3. Further inquiries can be directed to the corresponding author.

7 CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Bai, X.-C., McMullan, G., and Scheres, S. (2015). How cryo-em is revolutionizing structural biology. *Trends in biochemical sciences* 40, 49–57

Banerjee, S., Bartesaghi, A., Merk, A., Rao, P., Bulfer, S. L., Yan, Y., et al. (2016). 2.3 å resolution cryo-em structure of human p97 and mechanism of allosteric inhibition. *Science* 351, 871–875

Brüschweiler, R. (1995). Collective protein dynamics and nuclear spin relaxation. *The Journal of chemical physics* 102, 3396–3403

Costa, M. G., Fagnen, C., Vénien-Bryan, C., and Perahia, D. (2020). A new strategy for atomic flexible fitting in cryo-em maps by molecular dynamics with excited normal modes (mdenm-emfit). *Journal of chemical information and modeling* 60, 2419–2423

Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., et al. (2007). Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic acids research* 35, W375–W383

De la Rosa-Trevín, J., Otón, J., Marabini, R., Zaldivar, A., Vargas, J., Carazo, J., et al. (2013). Xmipp 3.0: an improved software suite for image processing in electron microscopy. *Journal of structural biology* 184, 321–328

De la Rosa-Trevín, J., Quintana, A., Del Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., et al. (2016). Scipion: A software framework toward integration, reproducibility and validation in 3d electron microscopy. *Journal of structural biology* 195, 93–99

Delarue, M. and Dumas, P. (2004). On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. *Proceedings of the National Academy of Sciences* 101, 6957–6962

Frank, J. (2017). Advances in the field of single-particle cryo-electron microscopy over the last decade. *Nature protocols* 12, 209–212

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H., et al. (2018). Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Science* 27, 14–25

Harastani, M., Sorzano, C. O. S., and Jonić, S. (2020). Hybrid electron microscopy normal mode analysis with scipion. *Protein Science* 29, 223–236

Haridas, M., Anderson, B., and Baker, E. (1995). Structure of human diferric lactoferrin refined at 2.2 å resolution. *Acta Crystallographica Section D: Biological Crystallography* 51, 629–646

Hofmann, S., Januliene, D., Mehdipour, A. R., Thomas, C., Stefan, E., Brüchert, S., et al. (2019). Conformation space of a heterodimeric abc exporter under turnover conditions. *Nature* 571, 580–583

Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C., and Grubmüller, H. (2019). Automated cryo-em structure refinement using correlation-driven molecular dynamics. *Elife* 8, e43542

Jonić, S. (2017). Computational methods for analyzing conformational variability of macromolecular complexes from cryo-electron microscopy images. *Current Opinion in Structural Biology* 43, 114 – 121

Jørgensen, R., Ortiz, P. A., Carr-Schmid, A., Nissen, P., Kinzy, T. G., and Andersen, G. R. (2003). Two crystal structures demonstrate large conformational changes in the eukaryotic ribosomal translocase. *Nature Structural & Molecular Biology* 10, 379–385

Kobayashi, C., Jung, J., Matsunaga, Y., Mori, T., Ando, T., Tamura, K., et al. (2017). Genesis 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *Journal of computational chemistry* 38, 2193– 2206

Kulik, M., Mori, T., and Sugita, Y. (2021). Multi-scale flexible fitting of proteins to cryo-em density maps at medium resolution. *Frontiers in Molecular Biosciences* 8, 61

Lopéz-Blanco, J. R. and Chacón, P. (2013). imodfit: efficient and robust flexible fitting based on vibrational analysis in internal coordinates. *Journal of structural biology* 184, 261–270

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. *Structure* 13, 373–380

Miyashita, O., Kobayashi, C., Mori, T., Sugita, Y., and Tama, F. (2017). Flexible fitting to cryo-em density map using ensemble molecular dynamics simulations. *Journal of computational chemistry* 38, 1447–1461

Miyashita, O. and Tama, F. (2018). Hybrid methods for macromolecular modeling by molecular mechanics simulations with experimental data. *Integrative Structural Biology with Hybrid Methods* , 199–217

Müller, C., Schlauderer, G., Reinstein, J., and Schulz, G. E. (1996). Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* 4, 147–156

Müller, C. W. and Schulz, G. E. (1992). Structure of the complex between adenylate kinase from escherichia coli and the inhibitor ap5a refined at 1.9 å resolution: A model for a catalytic transition state. *Journal of molecular biology* 224, 159–177

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M., et al. (2020). Single-particle cryo-em at atomic resolution. *Nature* 587, 152–156

Norris, G., Anderson, B., and Baker, E. (1991). Molecular replacement solution of the structure of apolactoferrin, a protein displaying large-scale conformational change. *Acta Crystallographica Section B: Structural Science* 47, 998–1004

Oh, B.-H., Pandit, J., Kang, C.-H., Nikaido, K., Gokcen, S., Ames, G., et al. (1993). Three-dimensional structures of the periplasmic lysine/arginine/ornithine-binding protein with and without a ligand. *Journal of Biological Chemistry* 268, 11348–11355

Orzechowski, M. and Tama, F. (2008). Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophysical journal* 95, 5692–5705

Peng, L.-M., Ren, G., Dudarev, S., and Whelan, M. (1996). Robust parameterization of elastic and absorptive electron atomic scattering factors. *Acta Crystallographica Section A: Foundations of Crystallography* 52, 257–276

Schröder, G. F., Brunger, A. T., and Levitt, M. (2007). Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* 15, 1630–1641

Suhre, K., Navaza, J., and Sanejouand, Y.-H. (2006). Norma: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. *Acta crystallographica section D: biological crystallography* 62, 1098–1100

Suhre, K. and Sanejouand, Y.-H. (2004). ElNémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement. *Nucleic Acids Research* 32, W610–W614

Tama, F. and Brooks III, C. L. (2006). Symmetry, form, and shape: guiding principles for robustness in macromolecular machines. *Annu. Rev. Biophys. Biomol. Struct.* 35, 115–133

Tama, F., Gadéa, F., Marques, O., and Sanejouand, Y. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. *Proteins* 41, 1–7

Tama, F., Miyashita, O., and Brooks III, C. (2004a). Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-em. *Journal of Structural Biology* 147, 315 – 326

Tama, F., Miyashita, O., and Brooks III, C. L. (2004b). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *Journal of molecular biology* 337, 985–999

Tama, F. and Sanejouand, Y.-H. (2001). Conformational change of proteins arising from normal mode calculations. *Protein engineering* 14, 1–6

Tama, F., Wriggers, W., and Brooks III, C. (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. *Journal of molecular biology* 321, 297–305

Tirion, M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters* 77, 1905

Trabuco, L. G., Villa, E., Mitra, K., Frank, J., and Schulten, K. (2008). Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* 16, 673–683

Velazquez-Muriel, J., Sorzano, C., Fernandez, J., and Carazo, J. (2003). A method for estimating the ctf in electron microscopy based on arma models and parameter adjustment. *Ultramicroscopy* 96, 17–135

Vilas, J. L., Gómez-Blanco, J., Conesa, P., Melero, R., de la Rosa-Trevín, J. M., Otón, J., et al. (2018). Monores: automatic and accurate estimation of local resolution for electron microscopy maps. *Structure* 26, 337–344

Wang, Y., Rader, A., Bahar, I., and Jernigan, R. L. (2004). Global ribosome motions revealed with elastic network model. *Journal of structural biology* 147, 302–314

Wu, X., Subramaniam, S., Case, D. A., Wu, K. W., and Brooks, B. R. (2013). Targeted conformational search with map-restrained self-guided langevin dynamics: application to flexible fitting into electron

microscopic density maps. *Journal of structural biology* 183, 429–440

This is the Supplementary Material of an article accepted for publication in Journal of Molecular Biology, https://doi.org/10.1016/j.jmb.2022.167483

NMMD: Efficient cryo-EM flexible fitting based on simultaneous Normal Mode and Molecular Dynamics atomic displacements

Rémi Vuillemot [1,4], Osamu Miyashita [2], Florence Tama [3], Isabelle Rouiller [4], Slavica Jonic [1,*]

[1] IMPMC - UMR 7590 CNRS, Sorbonne Université, Muséum National d'Histoire Naturelle, Paris, France

[2] RIKEN Center for Computational Science, Japan

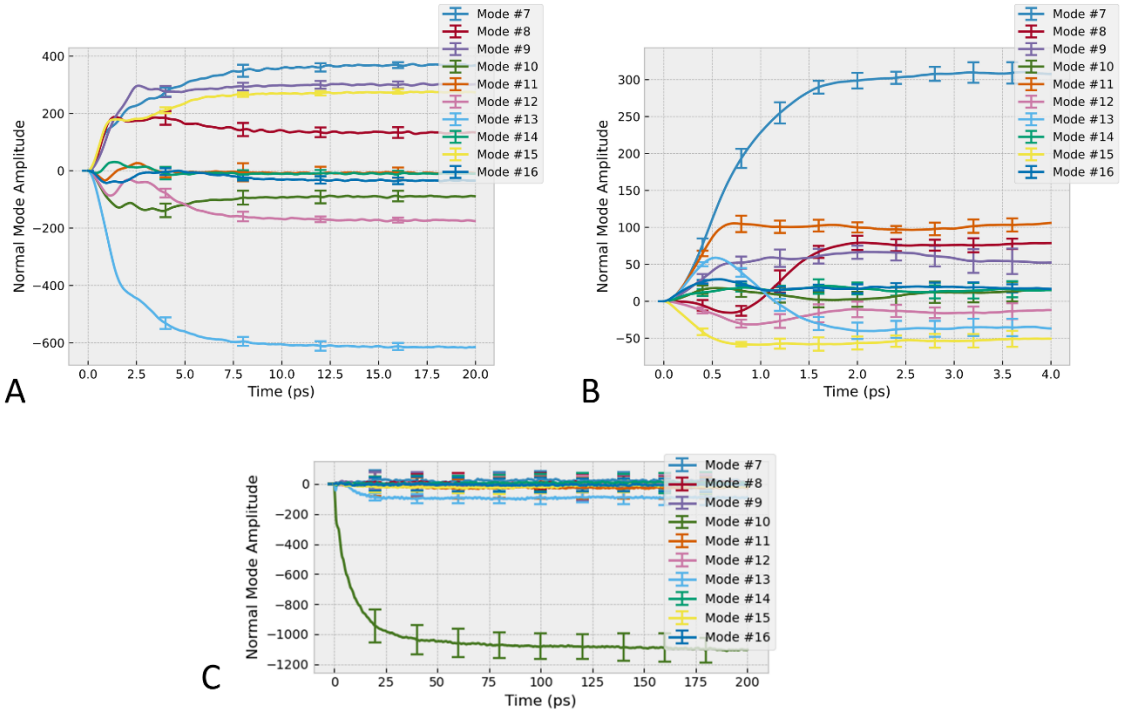
[3] Institute of Transformative Biomolecules and Department of Physics, Graduate School of Science, Nagoya University, Japan

[4] Department of Biochemistry & Pharmacology and Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Victoria, Australia

Contact details of the corresponding author:
Dr. Slavica Jonić
Sorbonne Université
IMPMC - CNRS UMR 7590, CC 115
4 Place Jussieu, 75005 Paris, France
Phone: +33 1 44 27 72 05
Fax: +33 1 44 27 37 85
E-mail: slavica.jonic@upmc.fr

1 Contribution of individual normal modes to a combination over the simulation time

To understand the contribution of individual normal modes to the fitting, we monitored the evolution of the normal-mode amplitudes (***q****(t)* in Eq 5 in the main text) over the simulation for Adenylate kinase (AK), ABC exporter (ABC), and p97 ATPase (p97). We reported the results in Supplementary Figure 1. For AK (Supplementary Figure 1B), we observe that the normal mode with the highest amplitude is mode 7, meaning that it contributes the most to the fitting, which is in the agreement with our preliminary experiments that have shown that mode 7 was well suited to describe the conformational change between the two given AK conformations. The same behavior can be observed for p97 (Supplementary Figure 1C), where mode 10 has the largest contribution from all of the used modes, as expected from our preliminary experiments. This mode moves the N domain of p97 up and down, which is the expected motion between the two given p97 conformations. For ABC (Supplementary Figure 1A), we observe that many modes, combined, contributed to the fitting, but still some modes contributed more than the others. Mode 9 that we observed in our preliminary experiments is among the four modes with the highest contribution to the fitting (modes 7, 9, 13, and 15). The ABC example shows the advantage of an automatic "selection" of modes by NMMD over a manual selection, as the modes are selected objectively and without a risk of missing to select some of the important modes. For each of the six experiments (six biomolecular complexes), Supplementary Table 1 reports the normal mode amplitude at the end of the simulation (referred to as final amplitude) for the replica with the lowest RMSD.



Supplementary Figure 1: *Normal-mode amplitudes over simulation time, averaged over the 16 replicas. ABC exporter (**A**), Adenylate kinase (**B**), p97 ATPase (**C**). See also Figure 1 and Table 2 in the main text.*

Biomolecular complex	Mode number									
	7	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
AK	303.4	66.0	44.7	48.2	110.9	-4.2	-51.4	13.4	-69.3	32.0
LAO	-108.8	-68.6	128.5	-21.1	-14.4	15.2	-5.0	-20.9	6.8	-29.9
LF	-247.7	-124.9	3.6	-139.8	-32.8	56.4	25.3	12.2	-35.3	2.7
EF2	15.7	-364.3	-31.9	71.6	218.5	66.5	-13.2	-9.2	27.1	19.8
ABC	371.3	162.5	273.6	116.2	-17.3	177.4	581.7	3.5	257.8	-57.5
p97	-25.0	-31.4	39.0	1216.5	20.1	43.7	148.0	16.0	19.8	-11.8

Supplementary Table 1 *: Final normal mode amplitude for the replica with the lowest RMSD, for all the six biomolecular complexes analyzed (AK: Adenylate kinase; LAO: LAO binding protein; LF: Lactoferrin; EF2: Elongation factor 2; ABC: ABC exporter; p97: p97 ATPase). See also Figure 1 and Table 2 in the main text.*

2 Additional simulation parameters

		AK	LAO	LF	EF2	ABC	p97
Total simulation time (ps)	MD	20	20	80	200	100	200
	NMMD	20	20	80	200	100	200
Convergence simulation time (ps)	MD	12	11	74	178	78	174
	NMMD	8	4	39	160	26	140

***Supplementary Table 2**: Total and convergence simulation times for the replica with the lowest RMSD, for all the six biomolecular complexes analyzed (AK: Adenylate kinase; LAO: LAO binding protein; LF: Lactoferrin; EF2: Elongation factor 2; ABC: ABC exporter; p97: p97 ATPase). See also Table 2 in the main text.*