



(51) International Patent Classification:

C07H 21/00 (2006.01) *C12N 15/74* (2006.01)
C07K 14/00 (2006.01) *C12N 5/10* (2006.01)

(21) International Application Number:

PCT/US2018/029152

(22) International Filing Date:

24 April 2018 (24.04.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

62/489,115 24 April 2017 (24.04.2017) US

(71) Applicant: TRUSTEES OF PRINCETON UNIVERSITY [US/US]; 87 Prospect Avenue, Princeton, NJ 08544 (US).

(72) Inventor; and

(71) Applicant: ZHANG, Teng [US/US]; 2580 Greenhill Way, Apt. 108, Oviedo, FL 32765 (US).

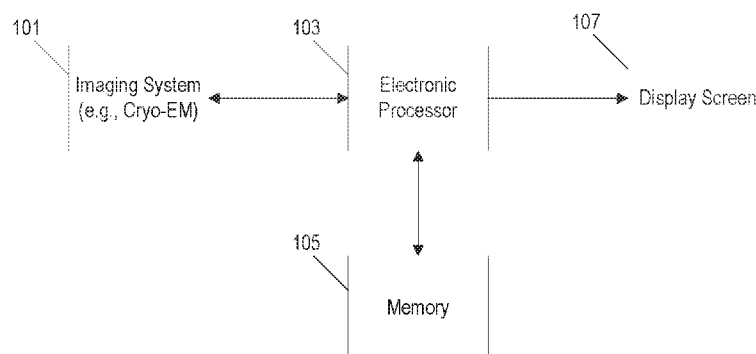
(72) Inventors: SINGER, Amit; 300 East 55th Street, Apt. 25C, New York, NY 10022 (US). BHAMRE, Tejal; 1929 Crisanto Avenue, Mountain View, CA 94040 (US).

(74) Agent: PAPROCKI, Andrew, J. et al.; Michael Best & Friedrich LLP, 100 East Wisconsin Avenue, Suite 3300, Milwaukee, WI 53202-4108 (US).

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) Title: ANISOTROPIC TWICING FOR SINGLE PARTICLE RECONSTRUCTION USING AUTOCORRELATION ANALYSIS

**FIG. 1**

(57) **Abstract:** Methods and systems are described for digitally reconstructing an unknown 3D structure of a target molecule using orthogonal extension. A plurality of 2D images of the target molecule are captured by an imaging system. An estimated low-order moment of the unknown 3D structure (e.g., a covariance matrix) is calculated based on the 2D images. A homologous molecule having a known 3D structure is identified and at least one expansion coefficient of the known structure of the homologous molecule is determined. At least one estimated expansion coefficient for the unknown structure is calculated based at least in part on the estimated low order moment of the unknown structure and the at least one expansion coefficient of the known structure. An estimated 3D reconstruction of the target molecule is then generated based on the at least one estimated expansion coefficient for the unknown structure of the target molecule.



MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,
TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW,
KM, ML, MR, NE, SN, TD, TG).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

ANISOTROPIC TWICING FOR SINGLE PARTICLE RECONSTRUCTION USING AUTOCORRELATION ANALYSIS

RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 62/489,115, filed April 24, 2017, entitled “ANISOTROPIC TWICING FOR SINGLE PARTICLE RECONSTRUCTION USING AUTOCORRELATION ANALYSIS,” the entire contents of which are incorporated herein by reference.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH OR DEVELOPMENT

[0002] This invention was made with government support under Grant No. GM-090200 awarded by the National Institutes of Health and Grant No. FA9550-12-1-0317 awarded by the Air Force Office of Scientific Research. The government has certain rights in the invention.

BACKGROUND

[0003] The present invention relates to systems and methods for determining a structure of a molecule using an imaging system such as, for example, x-ray free electron lasers and cryo-electron microscopy (cryo-EM).

SUMMARY

[0004] The three-dimensional structure of an object – for example, a molecule – can be determined using imaging techniques, such as, for example, cryo-electron microscopy (cryo-EM) in which multiple 2D images of the object are captured and then processed to generate a three-dimensional reconstruction of the object. In some implementations, the imaging system defines the 3D structure of the object using Cartesian coordinates while, in other implementations, the 3D structure of the object is defined formulaically, for example, using a function or a collection of functions (e.g., radial functions).

[0005] In some situations, an imaging system might be utilized to determine the 3D structure of a molecule that has been modified in some way from a known structure of the molecule (or a known structure of a similar molecule). For example, the 3D structure of a molecule may be changed when the molecule binds to another molecule. One option for

determining the 3D structure of the modified molecule would be to capture a sufficient number of 2D images in order to generate a reconstruction of the 3D structure. However, in some situations, this type of reconstruction can require a large number of 2D images and computational resources. Particularly in situations where many different modifications of the molecule structure are being analyzed, this process can take too long and can hinder/limit research progress.

[0006] In various embodiments, the invention described in this disclosure provides a technique for reconstructing an unknown 3D structure of a molecule based on a combination of 2D image data for the molecule and a previously determined 3D structure of a homologous (i.e., similar) molecule. Instead of capturing enough 2D images to determine the structure of the new molecule on its own, the methods and system described herein capture only enough 2D images of the molecule in order to determine how the molecule is different from the homologous molecule. As a result, fewer 2D images of the molecule are required and the computational load (and, in turn, processing time) for determining the 3D structure of the molecule is also reduced.

[0007] In some implementations, the coefficients of the function defining the 3D structure of a homologous molecule are known (e.g., based on previously captured and reconstructed image data). Based on the assumption that the 3D structure of the new molecule is similar to the 3D structure of the homologous molecule, the coefficients of a function defining the 3D structure of the new molecule will be similar/close to the coefficients of the function defining the 3D structure of the homologous molecule. Accordingly, in some implementations, the invention provides methods and systems for determining an estimated difference between the coefficients of the function defining the 3D structure of the new molecule and the coefficients of the function defining the 3D structure of the homologous molecule.

[0008] In some embodiments, the system is configured to compute the first two moments from the image data of a plurality of 2D images of the new molecule. The computed first moment is indicative of the average image and the second moment is indicative of the 2D covariance matrix of the image data. The 2D covariance matrix is indicative of the correlation between pixel values across the image in all image of the plurality of images. The covariance matrix computed from the 2D image data gives information about the 3D structure. However, in some implementations, the covariance matrix alone is not enough to

reconstruct the 3D structure. Instead, the system is configured to calculate the estimated expansion coefficients for the unknown 3D structure of the new molecule based on the 2D covariance matrix and the expansion coefficients of the known 3D structure of the homologous molecule.

[0009] In one embodiment, the invention provides a method for generating an estimated 3D structure of a target molecule based on a plurality of 2D images of the target molecule and a stored representation of a known 3D structure of a homologous molecule. First and second order moments of the 3D structure of the target molecule are calculated based on the captured 2D images. Expansion coefficients for a known 3D structure of a homologous molecule are determined based on data stored to a memory. Estimated expansion coefficients of the unknown 3D structure of the target molecule are then calculated based on the expansion coefficients from the homologous molecule and at least the second order moment calculated for the unknown 3D structure. After the estimated expansion coefficients for the unknown 3D structure are estimated, an estimated 3D structure of the target molecule is generated and, in some embodiments, a visual representation of the estimated 3D structure is output on a display screen.

[0010] In another embodiment, the invention provides a method for generating an estimated 3D structure of a target molecule. A plurality of 2D raw cryo-EM images of the target molecule are captured and the 2D image data is expanded in a truncated steerable basis (e.g., Fourier-Bessel basis). A covariance matrix for the 3D structure of the target molecule is then estimated by applying a steerable principal component analysis and estimated autocorrelation matrices are calculated from the estimated covariance matrix. The spherical harmonic expansion coefficients for the 3D structure of the target molecule are then calculated up to one or more missing orthogonal matrices. The 3D Fourier transform of a homologous molecule is calculated in a truncated spherical harmonic basis from a known 3D structure of a homologous molecule stored to a memory. An estimation of the missing orthogonal matrices for the 3D structure of the target molecule is then calculated using orthogonal extension based in part on the 3D Fourier transform of the homologous molecule. The spherical harmonic expansion coefficients of the unknown 3D structure of the target molecule are then recovered based at least in part on the estimated spherical harmonic expansion coefficients calculated from the 2D image data and the estimated missing orthogonal matrices. The estimated 3D structure of the target molecule is then generated

based on the recovered spherical harmonic expansion coefficients and, in some embodiments, a visual representation of the estimated 3D structure of the target molecule is output on a display screen.

[0011] Other aspects of the invention will become apparent by consideration of the detailed description and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Fig. 1 is a block diagram of a system for determining a 3D structure of a target molecule according to one embodiment.

[0013] Fig. 2 is a flowchart of a method for generating an estimated 3D structure of a target molecule based on a plurality of 2D images of the target molecule and a known 3D structure of a homologous molecule using the system of Fig. 1.

[0014] Fig. 3A is a flowchart of another method for generating an estimated 3D structure of a target molecule by recovering spherical harmonic expansion coefficients of the 3D structure of the target molecule from a plurality of 2D images of the target molecule and a 3D Fourier transform of a known 3D structure of a homologous molecule using the system of Fig. 1.

[0015] Fig. 3B is a flowchart of a method for estimating the missing orthogonal matrices in the method of Fig. 3A using orthogonal extension by least squares.

[0016] Fig. 3C is a flowchart of a method for estimating the missing orthogonal matrices in the method of Fig. 3A using orthogonal extension by anisotropic twicing.

[0017] Fig. 4A is a graph of the bias of the anisotropic twicing, least squares, and twicing estimators and also the family of estimators with $t = 1, 5, 10$ averaged over 1000 experiments.

[0018] Fig. 4B is a graph of the RMSE of the anisotropic twicing, least squares, and twicing estimators and also the family of estimators with $t = 1, 5, 10$ averaged over 1000 experiments.

[0019] Fig. 5A is an example of a 3D structure of a target molecule.

[0020] Fig. 5B is an example of the 3D structure of the target molecule of Fig. 5A reconstructed using a least squares estimator.

[0021] Fig. 5C is an example of the 3D structure of the target molecule of Fig. 5A reconstructed using a twicing estimator.

[0022] Fig. 5D is an example of the 3D structure of the target molecule of Fig. 5A reconstructed using an anisotropic twicing estimator.

[0023] Fig. 6A is an elevation view of an example of a target molecule.

[0024] Fig. 6B is an elevation view of an example of a homologous molecule similar in structure to the target molecule of Fig. 6A.

[0025] Fig. 6C is a reconstruction of the target molecule of Fig. 6A.

[0026] Fig. 6D is a reconstruction of the homologous molecule of Fig. 6B.

[0027] Fig. 6E is an example of the 3D structure of the target molecule of Fig. 6A reconstructed using a least squares estimator.

[0028] Fig. 6F is an example of the 3D structure of the target molecule of Fig. 6A reconstructed using a twicing estimator.

[0029] Fig. 6G is an example of the 3D structure of the target molecule of Fig. 6A reconstructed using an anisotropic twicing estimator.

[0030] Fig. 7A is a graph of the non-uniform viewing angle distribution of the images used to reconstruct the 3D structures.

[0031] Fig. 7B is a graph of the uniform viewing angle distribution after sampling of the images used to reconstruct the 3D structures.

[0032] Fig. 8A is an example of the 3D structure of the target molecule of Fig. 6A reconstructed using a least square estimator and approximately 30,000 two-dimensional images of the target molecule with an approximately uniform viewing angle distribution.

[0033] Fig. 8B is an example of a 3D structure of the target molecule of Fig. 6A reconstructed using a twicing estimator and approximately 30,000 two-dimensional images of the target molecule with an approximately uniform viewing angle distribution.

[0034] Fig. 8C is an example of a 3D structure of the target molecule of Fig. 6A reconstructed using an anisotropic twicing estimator and approximately 30,000 two-dimensional images of the target molecule with an approximately uniform viewing angle distribution.

[0035] Fig. 9A is an example of the 3D structure of the target molecule of Fig. 6A reconstructed using a least square estimator and approximately 73,000 two-dimensional images of the target molecule with a non-uniform viewing angle distribution.

[0036] Fig. 9B is an example of a 3D structure of the target molecule of Fig. 6A reconstructed using a twicing estimator and approximately 73,000 two-dimensional images of the target molecule with a non-uniform viewing angle distribution.

[0037] Fig. 9C is an example of a 3D structure of the target molecule of Fig. 6A reconstructed using an anisotropic twicing estimator and approximately 73,000 two-dimensional images of the target molecule with a non-uniform viewing angle distribution.

[0038] Fig. 10A is a graph of the FCR curve for the reconstruction of the entire molecule corresponding to the examples of Figs. 8A, 8B, and 8C.

[0039] Fig. 10B is a graph of the FCR curve for the reconstruction of an unknown additional structural component of the molecule corresponding to the examples of Figs. 8A, 8B, and 8C.

DETAILED DESCRIPTION

[0040] Before any embodiments of the invention are explained in detail, it is to be understood that the invention is not limited in its application to the details of construction and the arrangement of components set forth in the following description or illustrated in the following drawings. The invention is capable of other embodiments and of being practiced or of being carried out in various ways.

[0041] Fig. 1 illustrates an example of a system for generating reconstructions of the 3D structure of an object – for example, a molecule. An imaging system 101 captures image data of the object. The imaging system 101 may include, for example, a cryo-electron microscopy (cryo-EM) imaging system configured to capture a plurality of 2D projection images of the object (e.g., a target molecule) from a plurality of different viewing angles. Image data captured by the imaging system 101 is transmitted to an image processing system including an electronic processor 103 and a non-transitory computer-readable memory 105. The electronic processor 103 is configured to execute instructions stored on the memory 105 to provide system functionality such as, for example, described herein. The electronic processor 103 is configured to process the image data captured by and received from the imaging system 101 in order to reconstruct the 3D structure of the object (e.g., the target molecule). The electronic processor 103 is also configured to store the reconstruction of the 3D structure of the object to the memory 105 and, in some implementations, to store other reconstructions of other 3D structures to the memory 105. Similarly, the electronic processor 103 is communicatively coupled to a display screen 107 and is configured to selectively output a visual representation of a reconstruction of a 3D structure to the display screen 107 causing the visual representation to be shown on the display screen 107.

[0042] In some implementations, the system of Fig. 1 is configured to capture a sufficient number of 2D images of a target object such that a 3D reconstruction of the structure of the object can be generated entirely from the 2D image data captured by the imaging system 101. Alternatively or additionally, in some implementations, the system of Fig. 1 may be configured to generate a 3D reconstruction of the structure of the object based on a smaller set of 2D images and information derived from another 3D structure that is already stored on the memory 105. For example, the system of Fig. 1 may be configured to access a reconstruction of a 3D structure stored to memory that is similar to the target object (e.g., a homologous object). Instead of capturing and processing new images in order to reconstruct the entire 3D structure of the target object, the system is instead configured to capture and process new images of the target object in order to identify differences between the target object and the homologous object.

[0043] Fig. 2 is one example of a method for determining an estimated 3D structure of a target molecule using the system of Fig. 1. A plurality of 2D cryo-EM projection images of the target molecule are captured using the imaging system 101 (step 201). The system of Fig.

I then processes the captured image data to calculate a first order moment and a second order moment for the unknown 3D structure of the target molecule (step 203). In some implementations, the first order moment is an average image calculated from the raw image data captured by the imaging system and, in some implementations, the second order moment is a 2D covariance matrix indicating a correlation between pixel values across the image in all of the images of the target molecule captured by the imaging system 101. The 2D covariance matrix provides information about the 3D structure of the target molecule. However, the 2D covariance matrix alone does not provide enough data in order to reconstruct the entire 3D structure of the target molecule. Instead, the system access a previously stored 3D structure of a similar molecule (e.g., a homologous molecule) from the memory 105 (step 205) and calculates expansion coefficients of the known 3D structure (step 207). Based on the expansion coefficients for the known 3D structure of the homologous molecule and the second order moment calculated for the unknown 3D structure of the target molecule, the system calculates estimated expansion coefficients for the unknown 3D structure of the target molecule (step 209). Based on the estimated expansion coefficients, the system generates an estimated reconstruction of the 3D structure of the target molecule (step 211). The estimated reconstruction is then stored to memory and a visual representation of the estimated reconstruction of the 3D structure of the target molecule is output on the display screen 107.

[0044] The missing phase problem in crystallography entails recovering information about a crystal structure that is lost during the process of imaging. In X-ray crystallography, the measured diffraction patterns provide information about the modulus of the 3D Fourier transform of the crystal. The phases of the Fourier coefficients need to be recovered by other means, in order to reconstruct the 3D electron density map of the crystal. One method to solve the missing phase problem is Molecular Replacement (MR), which relies on a previously solved homologous structure which is similar to the unknown structure. The unknown structure is then estimated using the Fourier magnitudes of its diffraction data, along with phases from the homologous structure.

[0045] The missing phase problem can be formulated mathematically using matrix notation that enables generalization as follows. Each Fourier coefficient A is a complex-valued scalar, i.e., $A \in \mathbb{C}^{1 \times 1}$ that we wish to estimate, given measurements of $C = A A^*$ (A^* denotes the complex conjugate transpose of A , i.e., $A_{ij}^* = \overline{A_{ji}}$), corresponding to the

Fourier squared magnitudes, and $\mathbf{B}$ corresponds to a previously solved homologous structure such that $\mathbf{A} = \mathbf{B} + \mathbf{E}$, where $\mathbf{E}$ is a small perturbation. We denote an estimator of $\mathbf{A}$ as $\hat{\mathbf{A}}$.

There are many possible choices for such an estimator. One such choice is the solution to the least squares problem

$$\hat{\mathbf{A}}_{LS} = \arg_{\mathbf{A}} \min \|\mathbf{A} - \mathbf{B}\|_F, \text{ subject to } \mathbf{A} \mathbf{A}^* = \mathbf{C} \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. However, $\hat{\mathbf{A}}_{LS}$ does not reveal the correct relative magnitude of the unknown part of the crystal structure, and the recovered magnitude is about half of the actual value. As a magnitude correction scheme, it was empirically found that setting the magnitude to be twice the experimentally measured magnitude minus the magnitude of the homologous structure has the desired effect of approximately resolving the issue. That is, the estimator $2\hat{\mathbf{A}}_{LS} - \mathbf{B}$ is used instead. This procedure is referred to as twicing.

[0046] In this disclosure, we consider the following problem: How to estimate $\mathbf{A} \in \mathbb{R}^{N \times D}$ (or $\mathbb{C}^{N \times D}$) from $\mathbf{C}$ and $\mathbf{B}$, where $\mathbf{C} = \mathbf{A} \mathbf{A}^*$ and $\mathbf{A} = \mathbf{B} + \mathbf{E}$ for matrix $\mathbf{E}$ of small magnitude? When $N = D$, the result described below for an asymptotically consistent estimator of $\mathbf{A}$ is given by $\hat{\mathbf{A}}_{AT} = \mathbf{B} + \mathbf{U} \mathbf{W} \mathbf{U}^* (\hat{\mathbf{A}}_{LS} - \mathbf{B})$.

[0047] The motivation to study this problem is 3D structure determination in single particle reconstruction (SPR) without estimating the viewing angle associated with each image. Although the examples discussed in this disclosure focus on cryo-electron microscopy (cryo-EM), the methods and systems can also be applied to SPR using other imaging techniques including, for example, x-ray free electron lasers (XFEL). In SPR using XFEL, short but intense pulses of x-rays are scattered from the molecule. The measured 2D diffraction patterns in random orientations are used to reconstruct the 3D diffraction volume by an iterative refinement procedure, akin to the approach in cryo-EM.

[0048] Some of the examples discussed below utilized orthogonal extension (OE) in cryo-EM that combines ideas from MR and Kam's autocorrelation analysis for the purpose of 3D homology modeling, that is, for reconstruction of an unknown complex directly from its raw, noisy images when a previously solved similar complex exists. In SPR using cryo-EM, the 3D structure of a macromolecule is reconstructed from its noisy, contrast transfer function (CTF) affected 2D projection images. Individual particles images are picked from

micrographs, preprocessed, and used in further parts of the cryo-EM pipeline to obtain the 3D density map of the macromolecule. Algorithms such as RELION, XMIPP, SPIDER, EMAN2, FREALIGN refine a starting 3D structure using noisy 2D projection images. The result of the refinement procedure is often dependent on the choice of the initial model. It is therefore important to have a procedure to provide a good starting model for refinement. Also, a high quality starting model may significantly reduce the computational time associated with the refinement procedure. Such a high quality starting model can be obtained using OE. The main computational component of autocorrelation analysis is the estimation of the covariance matrix of the 2D images. This computation requires only a single pass over the experimental images. Autocorrelation analysis is therefore much faster than iterative refinement, which typically takes many iterations to converge. In fact, the computational cost of autocorrelation analysis is even lower than that of a single refinement iteration, as the latter involves comparison of image pairs (noisy raw images with volume projections). OE can also be used for the purpose of model validation, being a complementary method for structure prediction.

[0049] In some implementations, OE predicts the structure directly from the raw, noisy images without any averaging. The method is analogous to MR in x-ray crystallography for solving the missing phase problem. In OE, the homologous structure is used for estimating the missing orthogonal matrices associated with the spherical harmonic expansion of the 3D structure in reciprocal space. It is important to note that the missing orthogonal matrices in OE are not associated with the unknown pose of the particles, but with the spherical harmonics expansion coefficients. The missing coefficient matrices are, in general, rectangular of size $N \times D$, which serves as the motivation to extend twicing to the general case of finding an estimator when $(N,D) \neq (1,1)$.

[0050] Fig. 3A illustrates a method for estimating a 3D structure of a target molecule using orthogonal extension. First, a plurality of raw 2D cryo-EM projection images are captured of a target molecule with an unknown 3D structure (step 301). Next, the 2D images are expanded in a truncated steerable basis such as Fourier-Bessel basis (step 303). The system then applies a steerable principal component analysis to estimate the covariance matrix of the underlying clean images and their in-plane rotations (step 305). The autocorrelation matrices of the unknown 3D structure are then estimated from the covariance matrix, for example, using Kam's theory (step 307). The Cholesky decomposition of the

autocorrelation matrices is then computed to estimate the spherical harmonic expansion coefficients of the unknown 3D structure of the target molecule up to the missing orthogonal matrices (step 309). In order to estimate/recover the missing orthogonal matrices for the unknown 3D structure of the target molecule, a reconstruction of a similar molecule (e.g., a homologous molecule) with a known 3D structure is accessed from memory (step 311) and the 3D Fourier transform of the homologous molecule is computed in a truncated spherical harmonic basis (e.g., spherical-Bessel) (step 313). Orthogonal extension is then applied (as discussed in further detail below) to estimate the missing orthogonal matrices (step 315) and the spherical harmonic expansion coefficients of the unknown 3D structure of the target molecule are recovered (step 317). Finally, based on the recovered spherical harmonic expansion coefficients, a reconstruction of the 3D structure of the target molecule is generated (step 319). In some implementations, the reconstruction of the 3D structure is stored to memory and/or a visual representation of the reconstruction of the 3D structure is output on a display screen.

[0051] “Orthogonal Retrieval” is an approach for 3D homology modeling based on Kam’s theory. Orthogonal Retrieval can be regarded as a generalization of the MR method from x-ray crystallography to cryo-EM. Let $\Phi_A : \mathbb{R}^3 \rightarrow \mathbb{R}$ be the electron scattering density of the unknown structure, and let $F(\Phi_A) : \mathbb{R}^3 \rightarrow \mathbb{C}$ be its 3D Fourier transform. Consider the spherical harmonics expansion of $F(\Phi_A)$

$$F(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi) \quad (2)$$

where k is the radial frequency and Y_l^m are the real spherical harmonics. The autocorrelation matrices

$$C_l(k_1, k_2) = \sum_{m=-l}^l A_{lm}(k_1) \overline{A_{lm}(k_2)}, \quad l = 0, 1, \dots \quad (3)$$

can be estimated from the covariance matrix of the 2D projection images whose viewing angles are uniformly distributed over the sphere. This can be achieved with both clean as well as noisy images, as long as the number of noisy images is large enough to allow estimation of the underlying population covariance matrix of the clean images to the desired level of accuracy. The decomposition (equation 3) suggests that the l^{th} order autocorrelation matrix C_l has a maximum rank of $2l+1$, and the maximum rank is even smaller in the presence of symmetry.

[0052] While equation (2) holds true if the molecule is represented to infinitely high resolution, in practice the images are sampled on a finite pixel grid and information cannot be recovered beyond the Nyquist frequency. In addition, the molecule is compactly support in $\mathbb{R}^3$, and the support size can also be estimated from the images. It therefore follows to expand the volume in a truncated basis of spherical Bessel functions or 3D prolates. This leads to

$$F(\Phi_A)(k, \theta, \varphi) = \sum_{l=0}^L \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi), \quad l = 0, 1, \dots, L \quad (4)$$

where the truncation L is based on the resolution limit that can be achieved by the reconstruction. Our specific choice of L is described after equation (8) below. We can expand $A_{lm}(k)$ in a truncated basis of radial functions, chosen here as the spherical Bessel functions, as follows:

$$A_{lm}(k) = \sum_{s=1}^{S_l} a_{lms} j_{ls}(k) \quad (5)$$

[0053] Here the normalized spherical Bessel functions are

$$A_{lm}(k) = \frac{1}{c\sqrt{\pi}[j_{l+1}(R_{l,s})]} j_l\left(R_{l,s} \frac{k}{c}\right), \quad 0 < k < c, \quad s = 1, 2, \dots, S_l \quad (6)$$

where c is the bandlimit of the images, and $R_{l,s}$ is the s^{th} positive root of the equation $j_l(x) = 0$. The functions j_{ls} are normalized such that

$$\int_0^c j_{ls}(k) j_{ls}^*(k) k^2 dk = 1 \quad (7)$$

[0054] The number of the radial basis functions S_l in equation (5) is determined using the Nyquist criterion. We assume that the 2D images, and hence the 3D volume are compactly supported on a disk of radius R and have a bandlimit of $0 < c \leq 0.5$. We require that the maximum of the inverse Fourier transform of the spherical Bessel function and its first zero after this maximum are both inside the sphere of compact support radius R . The truncation limit S_l in equation (5) is then defined by the sampling criterion as the largest integer s that satisfies

$$R_{l,(s+1)} \leq 2\pi c R \quad (8)$$

[0055] L in equation (4) is the largest integer l for which equation (8) has only one solution, that is S_l in equation (5) is at least 1. Each C_l is a matrix of size $S_l \times S_l$ when using the representation of equation (5) in equation (3). S_l is a monotonically decreasing function of l with approximately linear decay that we compute numerically. In matrix notation, equation (3) can be written as

$$C_l = A_l A_l^* \quad (9)$$

where A_l is a matrix of size $S_l \times (2l+1)$, with $A_l(s, m) = a_{lms}$ in equation (5). From equation (9), we note that A_l can be obtained from the Cholesky decomposition of C_l up to a unitary matrix $U_l \in U(2l+1)$ (the group of unitary matrices of size $(2l+1) \times (2l+1)$). Since Φ_A is real-valued, one can show using properties of its Fourier transform together with properties of the real spherical harmonics, that $A_{lm}(k)$ (and hence A_l) is real for even l and purely imaginary for odd l . So A_l is unique up to an orthogonal matrix $O_l \in O(2l+1)$ (the group of unitary matrices of size $(2l+1) \times (2l+1)$). Determining O_l is an orthogonal retrieval problem.

[0056] If $S_l > 2l+1$, estimating the missing orthogonal matrix O_l is equivalent to estimating A_l . Since S_l is a decreasing function of l , for some large enough l we would have $S_l < 2l+1$, that is, it has $O(L)$ degrees of freedom. In such cases, it does not make sense to estimate O_L which has $O(L^2)$ degrees of freedom. But we can still estimate A_L closest to B_L using equation (1).

[0057] In order to determine the 3D Fourier transform $F(\Phi_A)$ and thereby the 3D density Φ_A , we need to determine the coefficient matrices A_l of the spherical harmonic expansion. Fig. 3B illustrates an example using OE in which the coefficient matrices A_l are estimated with the aid of a homologous structure Φ_B using a least squares estimator. Suppose Φ_B is a known homologous structure whose 3D Fourier transform $F(\Phi_B)$ has the following spherical harmonic expansion:

$$F(\Phi_B)(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l B_{lm}(k) Y_l^m(\theta, \varphi) \quad (10)$$

[0058] In practice, the homologous structure Φ_B is available at some finite resolution, therefore only a finite number of coefficient matrices B_l ($l=0, 1, \dots, L_B$) are given. We show how to estimate the unknown structure Φ_A up to the resolution dictated by the input images and the resolution of the homologous structure through estimating the coefficient matrices A_l for $l=0, 1, \dots, L_A$ where $L_A = \min(L, L_B)$.

[0059] Let F_l be any matrix of size $S_l \times 2l+1$ satisfying $C_l = F_l F_l^*$, determined from the Cholesky decomposition of C_l (step 321). Then, using equation (9)

$$A_l = F_l O_l \quad (11)$$

where $O_l \in O(2l+1)$ for $S_l > 2l+1$. Using the assumption that the structures are homologous, $A_l \approx B_l$, one can determine O_l as the solution to the least squares problem

$$O_l = \arg \min_{O \in O} \min_{(2l+1)} \|F_l O - B_l\|_F^2 \quad (12)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Although the orthogonal group is non-convex, there is a closed form solution to equation (12) given by

$$O_l = V_l U_l^T \quad (13)$$

where

$$B_l^* F_l = U_l \Sigma_l V_l^T \quad (14)$$

is a singular value decomposition (SVD) of $B_l^* F_l$ (step 323). Thus, A_l can be estimated by the following least squares estimator (step 325):

$$\hat{A}_{l,LS} = F_l V_l U_l^T \quad (15)$$

[0060] Hereafter, we drop the subscript l for convenience, since the procedure can be applied to each l separately.

[0061] As a natural generalization, one may wonder whether the estimator $2\hat{A}_{LS} - B$ (i.e., an estimator utilizing twicing) is unbiased for $(N,D) \neq (1,1)$. We assume that A is sampled from the model $A = FV$, wherein $F F^* = C$ and V is a random orthogonal matrix (or a random unitary matrix) sampled from the uniform distribution with Haar measure over the orthogonal group when A is a real-valued matrix or the unitary group (when A is a complex-valued matrix). This probabilistic model is reasonable for equation (1), because when $A A^*$ is given, F is known and V is an unknown orthogonal or unitary matrix, that is, we have no prior information about V . In addition, we assume the B is a matrix close to A such that $A - B$ is fixed. An unbiased estimator of A which is an affine transformation of $\hat{A}_{LS}$ is determined as follows.

[0062] When $N=D$, assuming that the spectral decomposition of C is given by $C=U \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_D)U^*$, then using our probabilistic model we have

$$\mathbb{E}[A - \hat{A}_{LS}] = UTU^*(A - B) + o(\|A - B\|_F) \quad (16)$$

where T is a diagonal matrix with i^{th} diagonal entry given by

$$T_{ii} \begin{cases} \frac{1}{D} \left[-\frac{1}{2} + \sum_{1 \leq j \leq D} \frac{\lambda_i^2}{\lambda_i^2 + \lambda_j^2} \right] & \text{when } A, C \in \mathbb{R}^{D \times D} \\ \frac{1}{D} \sum_{1 \leq j \leq D} \frac{\lambda_i^2}{\lambda_i^2 + \lambda_j^2} & \text{when } A, C \in \mathbb{C}^{D \times D} \end{cases}$$

and $f(X) = o(\|X\|_F)$ means that $\limsup_{\|X\|_F \rightarrow 0} f(X)/\|X\|_F \rightarrow 0$.

[0063] From equation (16), we have $(-I + UTU^*)(A - B) = B - \mathbb{E}[\hat{A}_{LS}] + o(\|A - B\|_F)$ and an asymptotically consistent estimator of A is given by

$$\hat{A}_{AT} = B - (I - UTU^*)^{-1}(B - \hat{A}_{LS}) = B + UWU^*(\hat{A}_{LS} - B) \quad (17)$$

where $W = (I - T)^{-1}$. When $A, B \in \mathbb{C}^{1 \times 1}$, the matrices reduce to scalars: $U = 1$, $T = 1/2$, $W = 2$, and $\hat{A}_{AT} = B + 2(\hat{A}_{LS} - B)$.

[0064] From equation (16), it follows that

$$[A] = \mathbb{E}[\hat{A}_{LS}] + UTU^*(A - B) + o(\|A - B\|_F) \quad (18)$$

[0065] A could be approximated in the right-hand-side of equation (18) by $\hat{A}_{LS}$, which leads to a new estimator $\hat{A}_T^{(1)} = \hat{A}_{LS} + UTU^*(\hat{A}_{LS} - B)$. In fact, there exists a family of estimators by approximating A recursively in the right-hand-side of equation (18) by $\hat{A}_T^{(t-1)}$ with $\hat{A}_T^{(0)} = \hat{A}_{LS}$:

$$\hat{A}_T^{(t)} = \hat{A}_{LS} + UTU^*(\hat{A}_T^{(t-1)} - B) \quad (19)$$

This family of estimators can be written explicitly as

$$\hat{A}_T^{(t)} = B + U(I + T + T^2 + \dots + T^t)U^*(\hat{A}_{LS} - B) \quad (20)$$

Using $W = (I - T)^{-1} = \sum_{i=0}^{\infty} T^i$, we have that $\hat{A}_T^{(t)} = \hat{A}_{AT}$ as $t \rightarrow \infty$. In general, this family of estimators has smaller variance than $\hat{A}_{AT}$, but larger bias since they are not unbiased (see, e.g., Figs. 4A and 4B).

[0066] If $N > D$, then the column space A is the same as the column space of C . Let P be the projector of size $N \times D$ to this column space, then we have $A = PP^*A$. As a result, to find an unbiased estimator of A , it is sufficient to find an unbiased estimator of P^*A , which is a square matrix. Since P^*A is close to P^*B and $(P^*A)(P^*A)^* = P^*CP^*$ is known, an unbiased estimator of P^*A can be obtained through equation (17), with B replaced by P^*B and C replaced by P^*CP^* . In summary, an unbiased estimator of A can be obtained in two steps: First, find $\hat{A}_{AT}^{(0)}$, an unbiased estimator of P^*A , by applying equation (17), with B replaced by P^*B and C replaced by P^*CP^* . Second, an unbiased estimator of A is obtained by $\hat{A}_{AT} = P\hat{A}_{AT}^{(0)}$.

[0067] If $N < D$, we use the following heuristic estimator. Let P be a matrix of size $D \times N$ that is the projector to the row space of B , and assuming that $\hat{A}$, the estimator of A , has the same row space as B , then $\hat{A} = \hat{A}PP^*$, and it is sufficient to find $\hat{A}P$, and estimator of AP . With $(AP)(AP)^* = AA^* = C$ known and the fact that AP is close to BP , we may use the estimator of equation (17). In summary, we use the following procedure: First, find $\hat{A}_{AT}^{(0)}$, an estimator of AP , by applying the estimator of equation (17), with B replaced by BP and, second, an estimator of A is obtained by $\hat{A}_{AT} = \hat{A}_{AT}^{(0)}P^*$.

[0068] We remark that for $N < D$, there is no theoretical guarantee to show that it is an unbiased estimator, unlike the setting $N \geq D$. However, the assumption that $\hat{A}$ has the same column space as B is reasonable and the proposed estimator performs well in practice.

[0069] The autocorrelation matrices C_l are derived from the covariance matrix Σ of the 2D Fourier transformed projection images through the equation:

$$C_l(|k_1|, |k_2|) = 2\pi(2l+1) \int_0^\pi \Sigma(|k_1|, |k_2|, \psi) P_l(\cos \psi) \sin \psi d\psi \quad (21)$$

where ψ is the angle between the vectors k_1 and k_2 in the x-y plane. We estimate the covariance matrix Σ of the underlying 2D Fourier transformed clean projection images. This estimation method provides a more accurate covariance compared to a classical sample covariance matrix. First, it corrects for the CTF. Second, it performs eigenvalue shrinkage, which is critical for high dimensional statistical estimation problems. Third, it exploits the block diagonal structure of the covariance matrix in a steerable basis, a property that follows from the fact that any experimental image is just as likely to appear in different in-place rotations. A steerable basis consists of outer products of radial functions (such as Bessel functions) and Fourier angular modes. Each block along the diagonal corresponds to a

different angular frequency. Moreover, the special block diagonal structure facilitates fast computation of the covariance matrix.

[0070] Since the autocorrelation matrix C_l estimated from projection images can have a rank exceeding $2l + 1$, we first find its best rank $2l + 1$ approximation via singular value decomposition, before computing its Cholesky decomposition. In the case of symmetric molecules, we use the appropriate rank as dictated by classical representation theory of $SO(3)$ (less than $2l + 1$).

[0071] Fig. 3C illustrates a method for estimating the 3D structure of a target molecule using anisotropic twicing as described above (particularly in reference to equations (16) through (21)). Given a known structure B of a homologous molecule and a covariance matrix C of the target molecule, the system is configured to find a matrix $F \in \mathbb{C}^{N \times D}$ such that $C = F F^*$ (step 331). Next, the system calculates $B^* F$ and its singular value decomposition $B^* F = U_0 \Sigma_0 V_0^*$ (step 333). The system then calculates the OE-LS estimator as described in Fig. 3B (step 335). If $N=D$ (step 337), then the anisotropic twicing estimator is defined as $\hat{A}_{AT} = B + U W U^* (\hat{A}_{LS} - B)$ (step 339). If $N > D$ (step 341), then the anisotropic twicing estimator is defined as $\hat{A}_{AT} = P \hat{A}_{AT}^{(0)}$ assuming that P is a projectors of size $N \times D$ to the D -dimensional subspace spanned by the columns of C (step 343). If $N < D$, Then the anisotropic twicing estimator is defined as $\hat{A}_{AT} = \hat{A}_{AT}^{(0)} P^*$ where P is a projector of size $D \times N$ to the N -dimensional subspace in $\mathbb{R}^D$ spanned by the rows of B (step 345).

[0072] For any parameter θ , the performance of its estimator $\theta^\wedge$ can be measured in terms of its mean squared error (MSE), $\mathbb{E} \|\theta - \theta^\wedge\|^2$. The MSE of any estimator can be decomposed into its bias and variance:

$$\text{MSE} = [\|\theta - \theta^\wedge\|^2] = \|\text{bias}\|^2 + \text{var} \quad (22)$$

where

$$\text{Bias} = [\theta^\wedge] - \theta \quad (23)$$

and

$$\text{Var} = [\|\theta^\wedge - \mathbb{E}[\theta^\wedge]\|^2] \quad (24)$$

[0073] Unbiased estimators are often not optimal in terms of MSE, but they can be valuable for being unbiased. In the examples below describe a numerical experiment starting

with a fixed $F \in \mathbb{R}^{10 \times 10}$ and an unknown matrix $A = FO$ where O is a random orthogonal matrix. We are given a known similar matrix B such that $A = B + E$. The goal is to estimate A given B and F . Fig. 4A shows a comparison of the bias and Fig. 4B shows a comparison of the root mean squared error (RMSE) of different estimators averaged over 10,000 runs of the numerical experiment: the anisotropic twicing estimator, the twicing estimator, the least squares estimator, and estimators from the family of estimators for some values of t . The figures demonstrate that the anisotropic twicing estimator is asymptotically unbiased at the cost of higher MSE.

[0074] Figs. 5A, 5B, 5C, and 5D illustrate another numeric example using synthetic datasets generated from an artificial molecule. Fig. 5A illustrates the target molecule made up of ellipsoids with the density set to 1 inside the ellipsoids and set to zero outside. The target molecule illustrated in Fig. 5A includes a small ellipsoid 501 that is not present in the homologous molecule. This represents the small perturbation E . When the Fourier volume $B + E$ is expanded in the truncated spherical Bessel basis described above, the average relative perturbation $\|E_l\|/\|A_l\|$ for the first few coefficients in the truncated spherical harmonic expansion for $l = 1, \dots, 10$ is 8%.

[0075] Next we generated 10,000 projection images from the volume of the artificial molecule of Fig. 5A. We then employed OE to reconstruct the volume A (as illustrated in Fig. 5A) from a known homologous structure B (similar to the volume in Fig. 5A, but without the additional ellipsoid 501). Fig. 5B shows the reconstructed volume using a least squares estimator. Fig. 5C shows the reconstructed volume using a twicing estimator. Fig. 5D shows the reconstructed volume using an anisotropic twicing estimator. We note that, while all three estimators are able to recover the additional subunit 501, the anisotropic twicing estimator of Fig. 5D best recovers the unknown subunit to its correct relative magnitude. The relative error in the region of the unknown subunit 501 is 59% with least squares, 31% with twicing, and 19% with anisotropic twicing.

[0076] Figs. 6A through 6G illustrate another example using a synthetic dataset generated from a TRPV1 molecule (with imposed C_4 rotational symmetry) in complex with DkTx and RTX (4). This volume is available on EMDB as EMDB-8117. The actual 3D structure of the target molecule is illustrated in Fig. 6A and the homologous molecule (i.e., the “known” structure) is illustrated in Fig. 6B. The small additional subunit E is visible in Fig. 6A as an extension over the top of the molecule. This represents the small perturbation E (i.e., the

difference between the target molecule structure A and the known homologous molecule structure B). Fig. 6C illustrates the “ground truth” representation of the structure of the target molecule A while Fig. 6D illustrates the formulaic representation of the homologous structure B . A total of 26,000 projection images were generated from A (i.e., the volume as illustrated in Fig. 6C), the effects of both CTF and additive white Gaussian noise were added, and then OE techniques were used to reconstruct the volume A . Fig. 6E shows the reconstruction of the target molecule structure using a least squares estimator, Fig. 6F shows the reconstruction using a twicing estimator, and Fig. 6G shows the reconstruction using an anisotropic twicing estimator. The C_4 symmetry was taken into account in the autocorrelation analysis by including in equation (2) only symmetry-invariant spherical harmonics Y_l^m for which $m=0 \bmod 4$.

[0077] Again, all three estimators are able to recover the additional subunit E , but the anisotropic twicing estimator best recovers the unknown subunit to its correct relative magnitude. The relative error in the unknown subunit is 43% with least squares, 56% with twicing, and 30% with anisotropic twicing.

[0078] In the example of Figs. 6A through 6G, OE techniques were applied to experimental data of the TRPV1 molecule in complex with DkTx and RTX, determined in lipid nanodisc, available on the public database Electron Microscopy Pilot Image Archive (EMPIAR) as EMPIAR-10059, and the 3D reconstruction is available on the electron microscopy data bank (EMDB) as EMDB-8117. The dataset consists of 73,000 motion corrected, picked particle images (which were used for the reconstruction in EMDB-8117 (Fig. 6A)) of size 192×192 with a pixel size of $1.3\text{\AA}$. The 3D structure of TRPV1 alone was used as the “homologous structure.” This is available on EMBD as EMDB-8118. The two structures differ only by a small DkTx and RTX subunit at the top, which can be seen in Fig. 6A.

[0079] Since the noise in the experimental images is colored while our covariance estimation procedure requires white noise, we first preprocess the raw images in order to “whiten” the noise. We estimate the power spectrum of noise using the corner pixels of all images. The images are then whitened using the estimated noise power spectrum. In the context of our mathematical model, the volume EMDB-8117 of TRPV1 with DkTx and RTX is the unknown volume A and the volume EMDB-8118 of TRPV1 alone is the known, similar

volume B . We use OE to estimate A given B and the raw, noisy projection images of A from the experimental dataset.

[0080] The basis assumption in Kam's theory is that the distribution of viewing angles is uniform. This assumption is difficult to satisfy in practice, since molecules in the sample can often have preference for certain orientations due to their shape and mass distribution. The viewing angle distribution in EMPIAR-10059 is non-uniform as illustrated in Fig. 7A. As a robustness test, we attempt 3D reconstruction with (i) all images, such that the viewing angle distribution is non-uniform (i.e., Fig. 7A) and (ii) by sampling images such that the viewing angle distribution of the images is approximately uniform (as shown in Fig. 7B). For the sampling procedure, we chose 10,000 points at random from the uniform distribution on the sphere and classify each image into these 10,000 bins based on the point closest to it. We discard bins that have no images, and for the remaining bins we pick a maximum of 3 points per bin. We use the selected images (slightly less than 30,000) for reconstruction with roughly uniformly distributed viewing angles.

[0081] Figs. 8A, 8B, and 8C illustrate the 3D reconstructions of the TRPV1 in complex with DkTx and RTX (EMPIAR-10059) using least squares, twicing, and anisotropic twicing, respectively, based on the images selected by sampling to impose approximately uniform viewing angle distribution. Figs. 9A, 9B, and 9C illustrate the 3D reconstructions of the TRPV1 in complex with DkTx and RTX (EMPIAR-10059) using least squares, twicing, and anisotropic twicing, respectively, based on all 73,000 images such that the viewing angle distribution is non-uniform. Fig. 10A shows the FCR curve for the reconstruction of the entire molecule obtained by OE using the least squares, twicing, and anisotropic twicing estimators corresponding to Figs. 8A, 8B, and 8C. Fig. 10B shows the FCR curve for only the unknown subunit.

[0082] The methods described in the examples above were implemented in the UNIX environment on a computer with 60 cores running at 2.3 GHz with total RAM of 1.5 TB. Using 20 cores, the total time taken for preprocessing the 2D images (whitening, background normalization, etc.) and computing the covariance matrix was 1400 seconds. Calculating the autocorrelation matrices using equation (21) involved some numerical integration which took 790 seconds, but for a fixed c and R (satisfied for datasets of roughly similar size and quality) these can be precomputed. Computing the basis functions and calculating the coefficient matrices of the homologous structure took 30 seconds and recovering the 3D

streutreu by applying the appropriate estimator (AT, twicing, or LS) and computing the volume from the estimated coefficients took 10 seconds.

[0083] The orthogonal retrieval problem in SPR is akin to the phase retrieval problem in x-ray crystallography. In crystallography, the measured diffraction patterns contain information about the modulus of the 3D Fourier transform of the structure but the phase information is missing and needs to be obtained by other means. In crystallography, the particle's orientations are known but the phase of the Fourier coefficients is missing, while in cryo-EM, the projection images contain phase information but the orientations of the particles are missing. Kam's autocorrelation analysis for SPR leads to an orthogonal retrieval problem which is analogous to the phase retrieval problem in crystallography. The phase retrieval problem is perhaps more challenging than the orthogonal matrix retrieval problem in cryo-EM. In crystallography, each Fourier coefficient is missing its phase, while in cryo-EM only a single orthogonal matrix is missing per several radial components. For each l , the unknown coefficient matrix A_l is of size $S_l \times (2l+1)$, corresponding to $(2l+1)$ radial functions. Each A_l is to be obtained from C_l , which is a positive semidefinite matrix of size $S_l \times S_l$ and rank at most $2l+1$. For $S_l > 2l+1$, instead of estimating $S_l(2l+1)$ coefficients, we only need to estimate an orthogonal matrix in $O(2l+1)$ which allows $l(2l+1)$ degrees of freedom. Therefore, there are $(S_l-l)(2l+1)$ fewer parameters to be estimated.

[0084] In some implementations, OE requires a sufficient number of 2D images to estimate the covariance matrix to the desired level of accuracy, so it has a greater chance of success for homology modeling from noisy images than other ab-initio methods such as those based on commons line, which fail at high noise levels.

[0085] In the examples described above, various implementations provide, among other things, a general magnitude correction scheme for a class of phase-retrieval problems, in particular, for orthogonal extension in cryo-EM. In some implementations, the magnitude correction scheme includes a variation of "twicing" referred to herein as anisotropic twicing. An asymptotically unbiased estimator is describes and the examples demonstrate 3D homology modeling using OE with synthetic and experimental datasets. These and similar techniques can be used to provide models and to initialize refinement directly from experimental images without performing class averaging and orientation estimation in cryo-EM and XFEL.

[0086] While anisotropic twicing outperforms least squares and twicing for synthetic data, in the examples described above, the three estimation methods have similar performance for experimental data. One possible explanation is that the underlying assumption made by all estimation methods that C_l are noiseless as implied by imposing the constraint $C_l = A_l A_l^*$ is violated more severely for certain types of experimental data. Specifically, the C_l matrices are derived from the 2D covariance matrix of the images and estimation errors are the result of noise in the images, the finite number of images available, non-uniformity of viewing directions, and imperfect estimation of individual image noise power spectrum, contrast transfer function, and centering. These effects are likely to be more pronounced in some types of experimental data compared to synthetic data. As a result, the error in estimating the C_l matrices from experimental data is larger. The error in the estimated C_l can be taken into consideration by replacing the constrained least squares problem of equation (1) with the regularized least squares problem

$$\min_A \|A - B\|_F^2 + \lambda \|C - AA^*\|_F^2 \quad (25)$$

where $\lambda > 0$ is a regularization parameter that would depend on the spherical harmonic order l .

[0087] Thus, the invention provides, among other things, methods and systems for generating a reconstruction of a 3D structure of an object (such as, for example, a target molecule) based on a low order moment of 2D image data for the object and prior knowledge of a similar, homologous 3D structure. Various features and advantages of the invention are set forth in the following claims.

CLAIMS

What is claimed is:

1. A method for digitally reconstructing an unknown 3D structure of a target molecule, the method comprising:
 - receiving, by an electronic processor, a plurality of 2D images of the target molecule captured by an imaging system;
 - calculating, by the electronic processor, an estimated low order moment of the unknown 3D structure of the target molecule based on the plurality of 2D images;
 - identifying a homologous molecule having a known 3D structure;
 - determining, by the electronic processor, at least one expansion coefficient of the known 3D structure of the homologous molecule;
 - calculating, by the electronic processor, at least one estimated expansion coefficient for the unknown 3D structure based at least in part on the estimated low order moment of the unknown 3D structure of the target molecule and the at least one expansion coefficient of the known 3D structure of the homologous molecule; and
 - generating, by the electronic processor, an estimated 3D reconstruction of the target molecule based on the at least one estimated expansion coefficient for the unknown 3D structure of the target molecule.
2. The method of claim 1, further comprising outputting on a display screen a visual representation of the estimated 3D reconstruction of the target molecule.
3. The method of claim 1, further comprising storing the estimated 3D reconstruction of the target molecule to a non-transitory computer-readable memory.
4. The method of claim 1, wherein calculating the estimated low order moment of the unknown 3D structure of the target molecule based on the plurality of 2D images includes calculating a covariance matrix from the plurality of 2D images.
5. The method of claim 4, further comprising calculating at least one autocorrelation matrix based on the covariance matrix, and
 - wherein calculating the at least one estimated expansion coefficient for the unknown 3D structure based at least in part on the estimated low order moment of the

unknown 3D structure of the target molecule and the at least one expansion coefficient of the known 3D structure of the homologous molecule includes calculating a set of expansion coefficients for the unknown 3D structure based at least in part on the at least one autocorrelation matrix.

6. The method of claim 4, further comprising capturing a plurality of noisy two-dimensional projection images of the target molecule using a cryo-electron microscopy system, and wherein calculating the covariance matrix includes calculating an estimated covariance matrix indicative of a plurality of clean two-dimensional projection images of the target molecule based on the plurality of noisy two-dimensional projection images.

7. The method of claim 1, further comprising expanding 2D image data from the plurality of 2D images in a truncated steerable basis, and wherein calculating the estimated low order moment includes calculating an estimated covariance matrix by applying a steerable principal component analysis to the expanded 2D image data.

8. The method of claim 1, wherein identifying a homologous molecule having the known 3D structure includes identifying the homologous molecule that is similar to the target molecule and is expected to have a similar 3D structure.

9. The method of claim 8, further comprising accessing a previous reconstruction of the known 3D structure of the homologous molecule from a non-transitory computer-readable memory.

10. The method of claim 1, further comprising
calculating, by the electronic processor, a plurality of estimated spherical harmonic expansion coefficients for the unknown 3D structure, wherein the plurality of estimated spherical harmonic expansion coefficients for the unknown 3D structure does not include at least one missing orthogonal matrix that cannot be recovered from the plurality of 2D images of the target molecule because image data of the target molecule in the plurality of 2D images is incomplete, and
wherein calculating the at least one estimated expansion coefficient of the unknown 3D structure includes

estimating the at least one missing orthogonal matrix based at least in part on data indicative of the known 3D structure of the homologous molecule using orthogonal extension, and

recovering a set of estimated spherical harmonic expansion coefficients for the unknown 3D structure of the target molecule based on the plurality of estimated spherical harmonic expansion coefficient for the unknown 3D structure and the estimated at least one missing orthogonal matrix.

11. The method of claim 10, wherein estimating the at least one missing orthogonal matrix using orthogonal extension includes applying an anisotropic estimator.
12. The method of claim 11, wherein applying the anisotropic estimator includes calculating the at least one missing orthogonal matrix $\hat{A}_{AT}$,
 where $\hat{A}_{AT} = B - UWU^*[B - \hat{A}_{LS}]$,
 where B is indicative of the known 3D structure of the homologous molecule,
 where $\hat{A}_{LS}$ is a least squares estimator of the at least one missing orthogonal matrix and is equal to $FV_\theta U_\theta^*$.
13. The method of claim 11, wherein applying the anisotropic estimator includes calculating the at least one missing orthogonal matrix $\hat{A}_{AT}$, where $\hat{A}_{AT} = P \hat{A}_{AT}^{(0)}$, and where P is a projector of size $N \times D$ to the D -dimensional subspace spanned by the columns of the autocorrelation matrix.
14. The method of claim 11, wherein applying the anisotropic estimator includes calculating the at least one missing orthogonal matrix $\hat{A}_{AT}$, where $\hat{A}_{AT} = \hat{A}_{AT}^{(0)} P^*$, and where P^* is a projector of size $D \times N$ to the N -dimensional subspace in $\mathbb{R}^D$ spanned by the rows of a matrix B indicative of the known 3D structure of the homologous molecule.
15. The method of claim 11, wherein applying the anisotropic estimator includes finding a $F \in \mathbb{C}^{N \times D}$ such that $C = FF^*$, wherein $C \in \mathbb{C}^{N \times N}$,
 calculating B^*F and calculating its singular value decomposition $B^*F = U_\theta \Sigma_\theta V_\theta^*$, where $B \in \mathbb{C}^{N \times D}$ and is indicative of the known 3D structure of the homologous molecule,

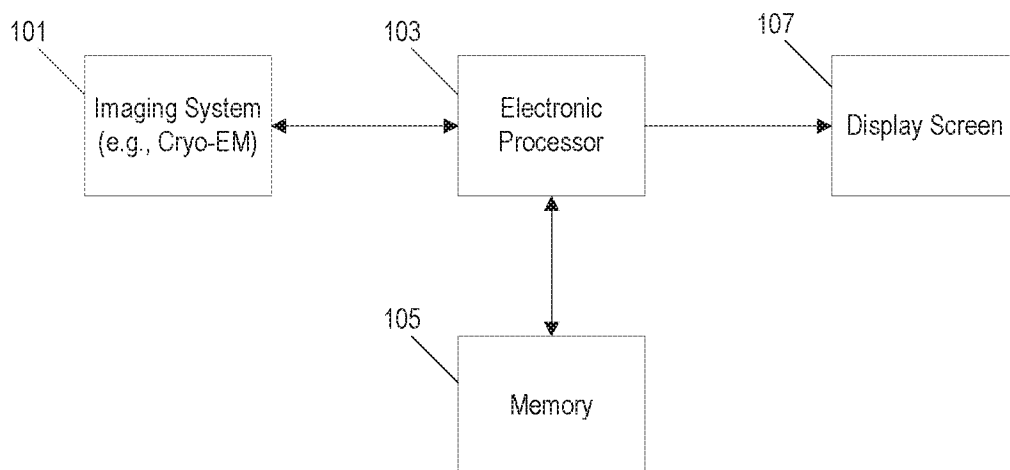
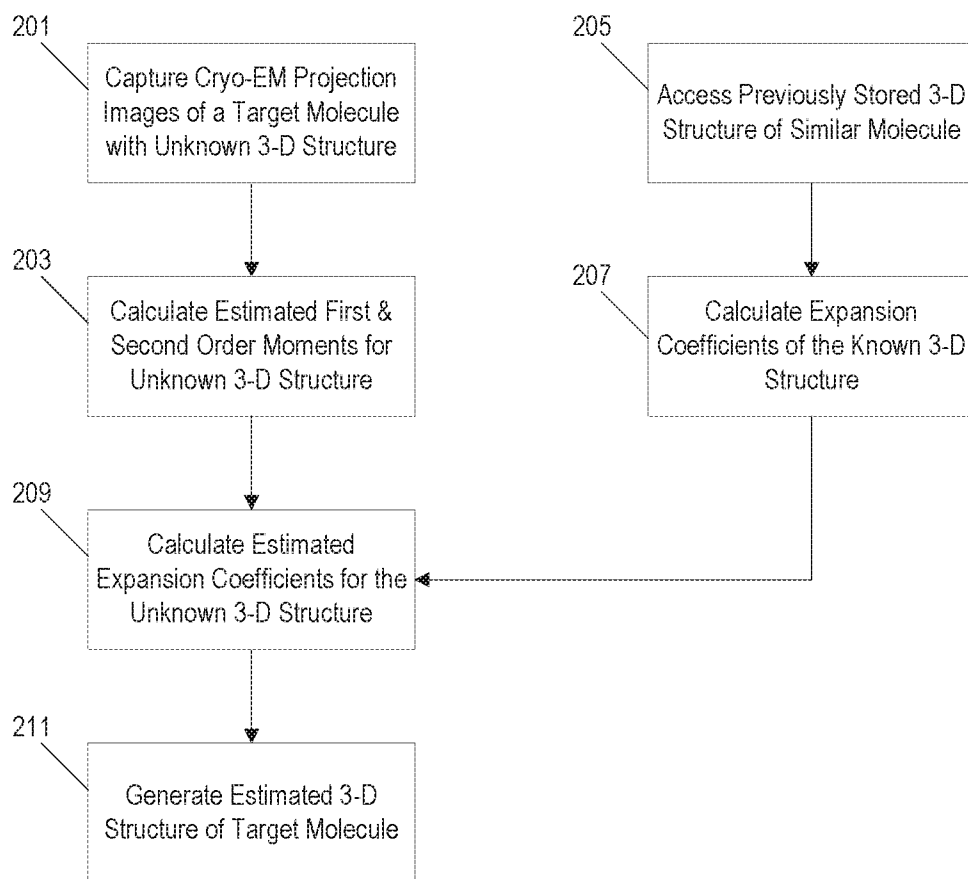
calculating an orthogonal extension least-squares estimator of the at least one missing orthogonal matrix $\hat{A}_{LS}$, where $\hat{A}_{LS}$ is equal to $FV_{\theta}U_{\theta}^*$, and

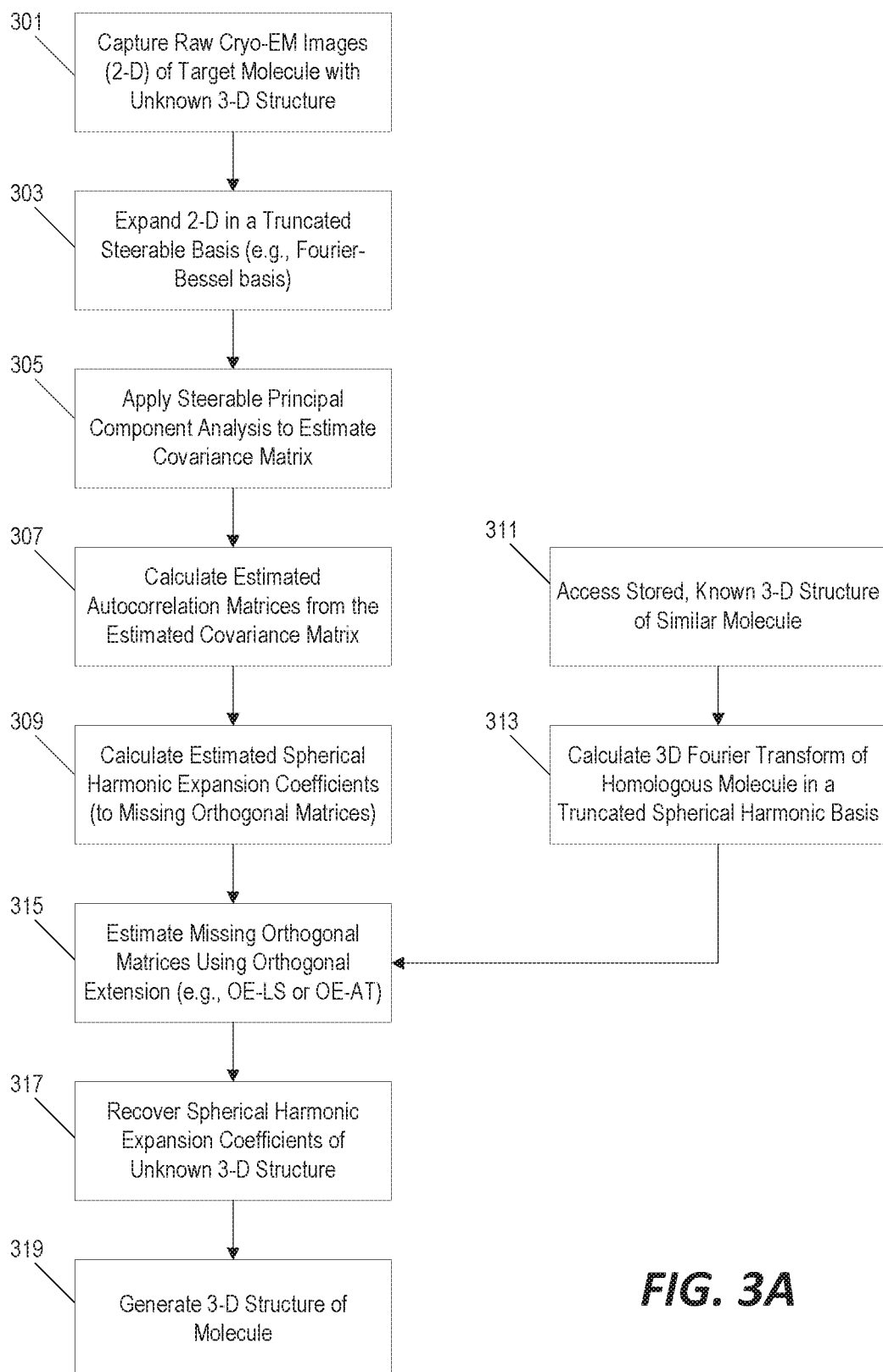
calculating the at least one missing orthogonal matrix $\hat{A}_{AT}$, wherein

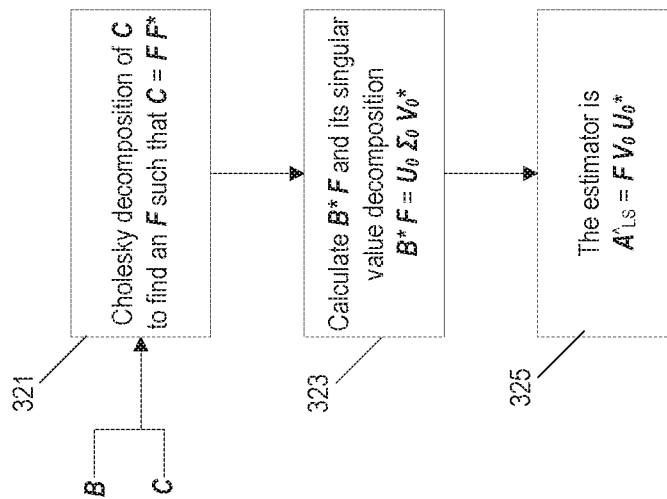
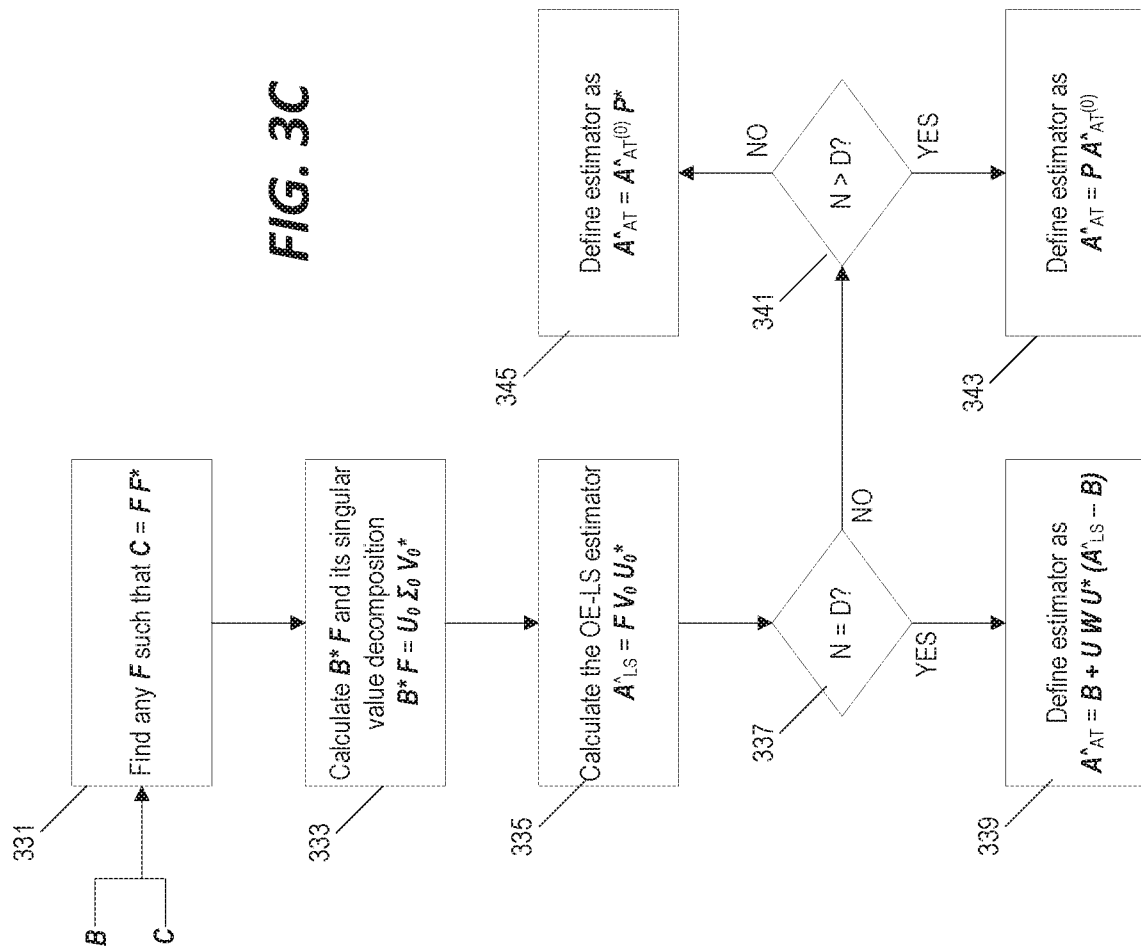
when $N=D$, $\hat{A}_{AT}$ is calculated as $B - UWU^*[B - \hat{A}_{LS}]$,

when $N > D$, $\hat{A}_{AT}$ is calculated as $P\hat{A}_{AT}^{(0)}$, and P is a projector of size $N \times D$ to the D -dimensional subspace spanned by the columns of the autocorrelation matrix, and

when $N < D$, $\hat{A}_{AT}$ is calculated as $\hat{A}_{AT}^{(0)}P^*$, and P^* is a projector of size $D \times N$ to the N -dimensional subspace in $\mathbb{R}^D$ spanned by the rows of a matrix B indicative of the known 3D structure of the homologous molecule

**FIG. 1****FIG. 2**

**FIG. 3A**



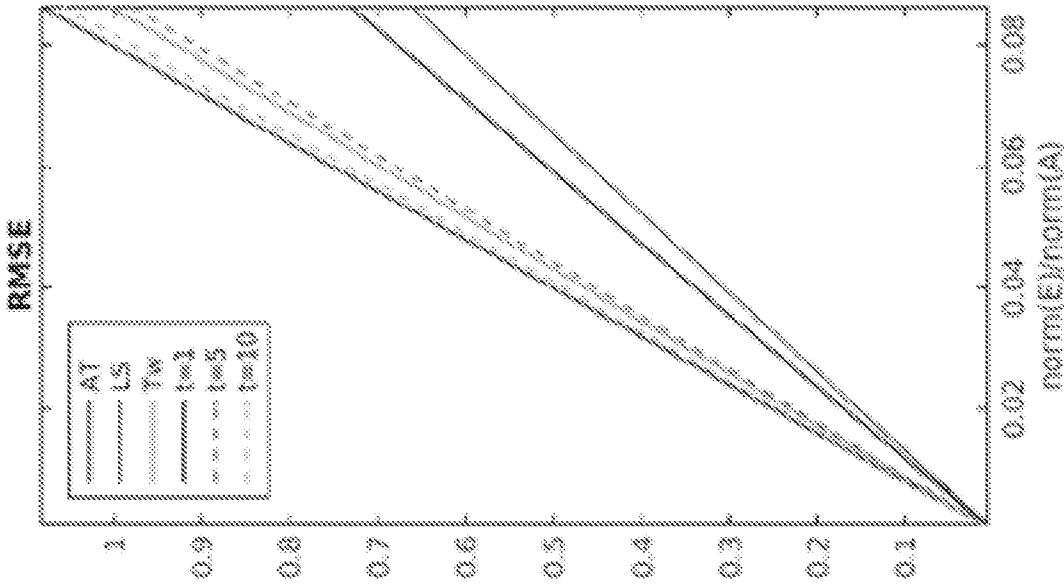


FIG. 4B

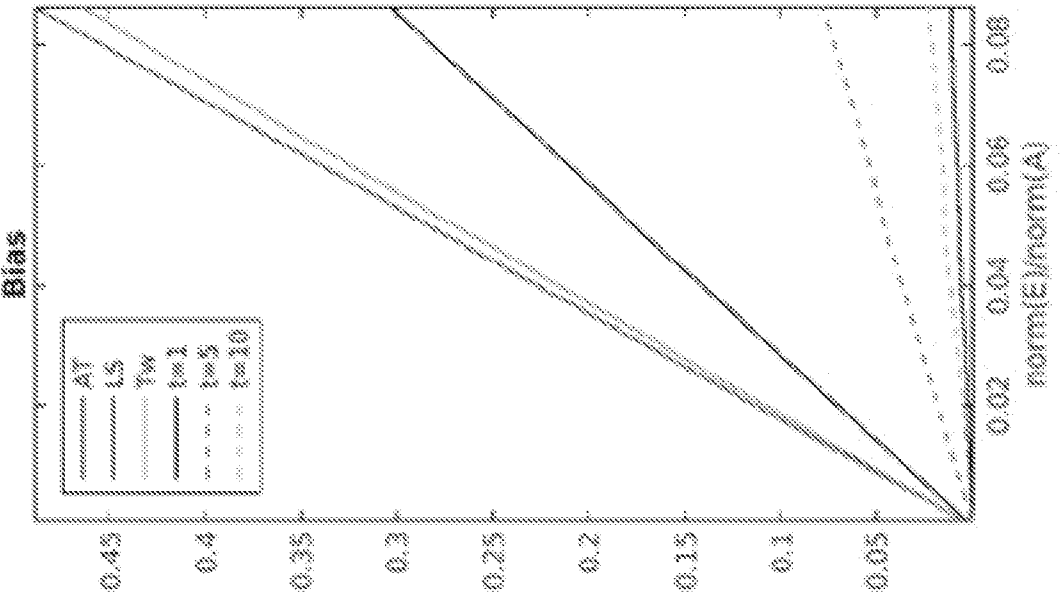


FIG. 4A

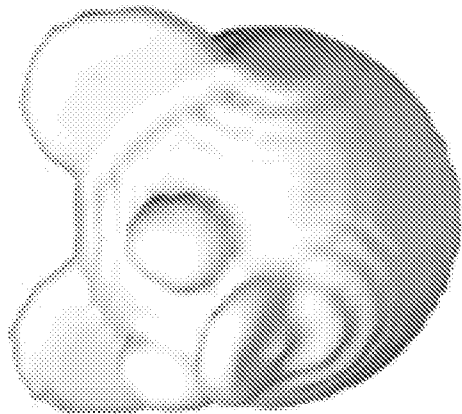


FIG. 5B

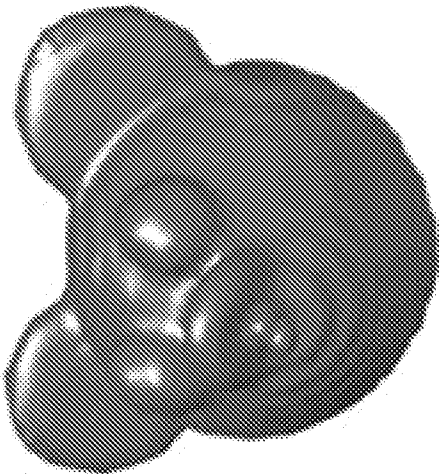


FIG. 5D

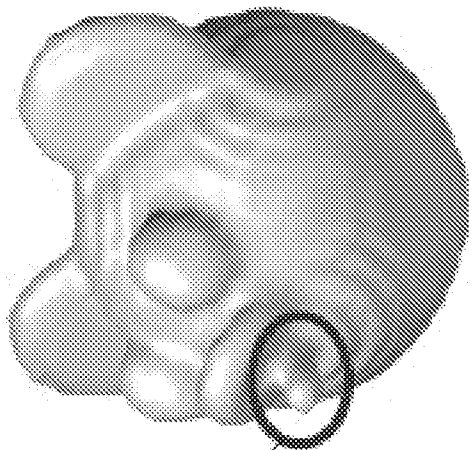


FIG. 5A

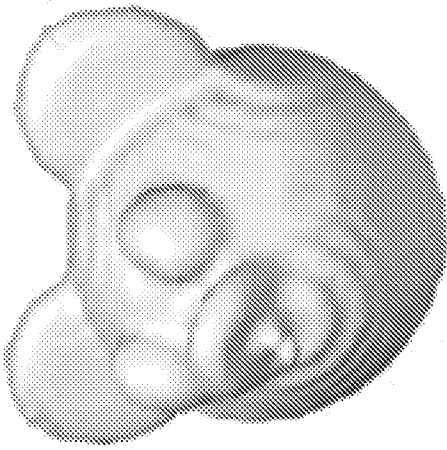


FIG. 5C

501

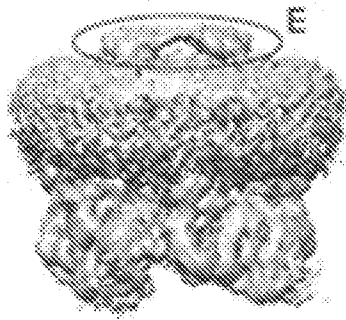


FIG. 6A

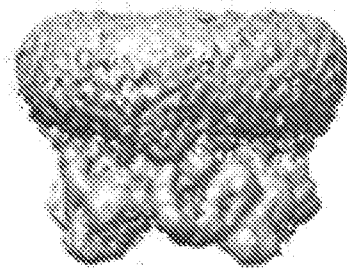


FIG. 6B

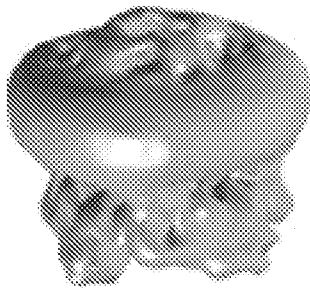


FIG. 6C

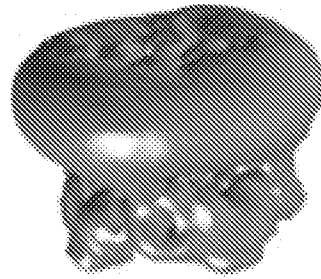


FIG. 6D

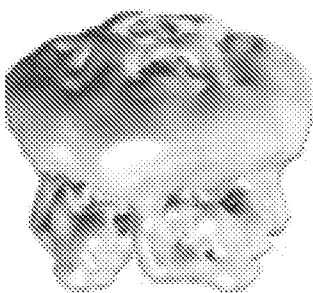


FIG. 6E

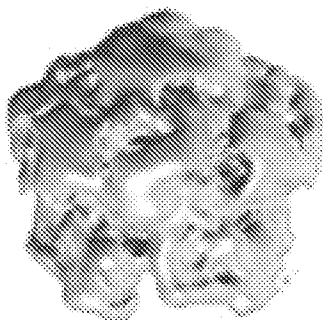


FIG. 6F

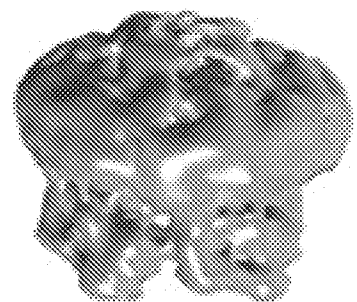


FIG. 6G

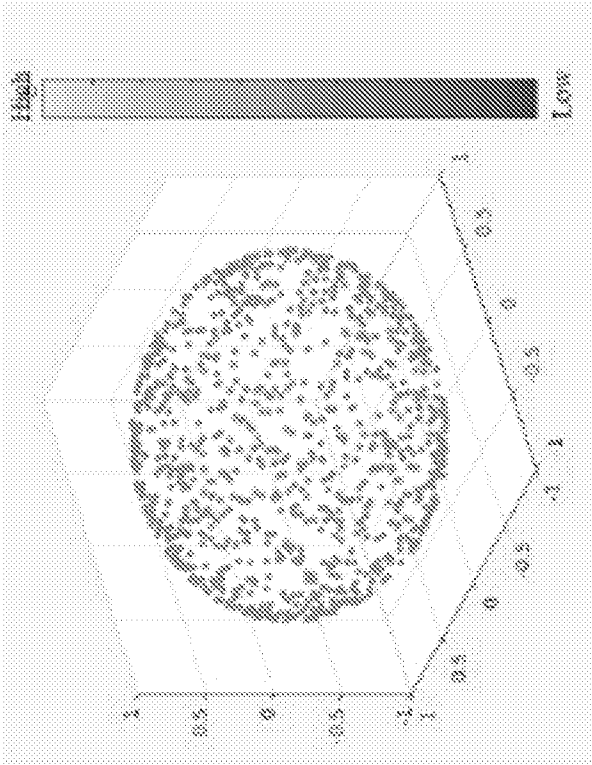


FIG. 7B

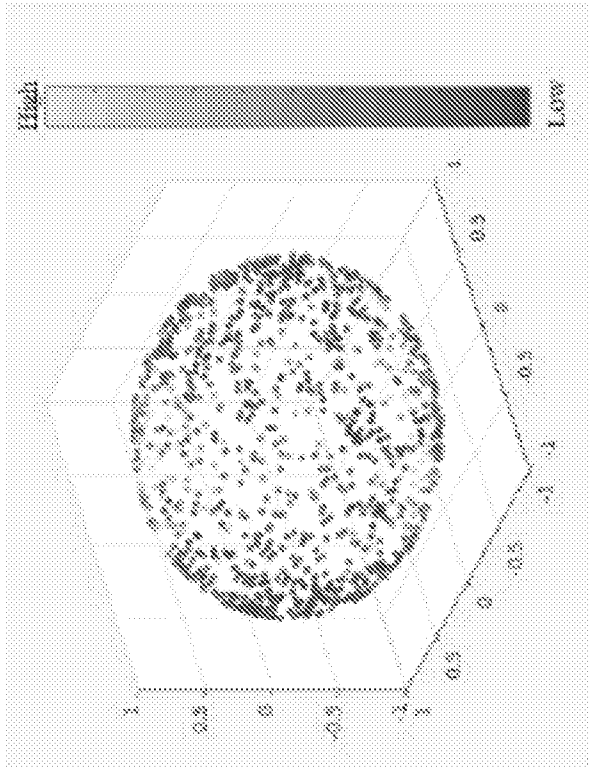


FIG. 7A

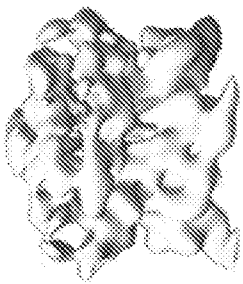


FIG. 8A

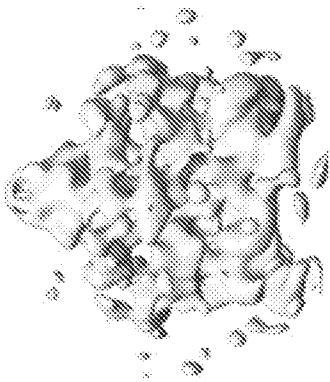


FIG. 8B

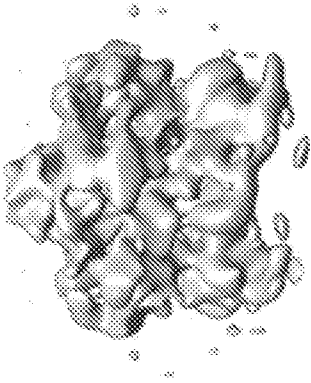


FIG. 8C

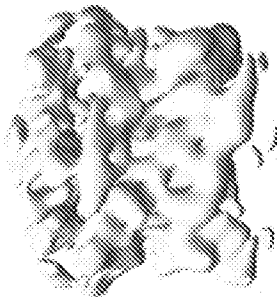


FIG. 9A

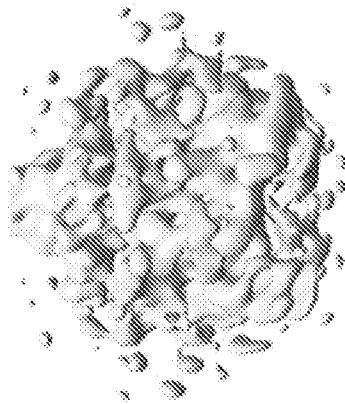


FIG. 9B

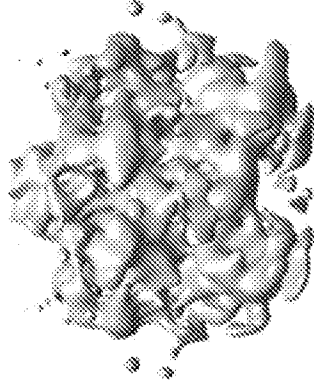
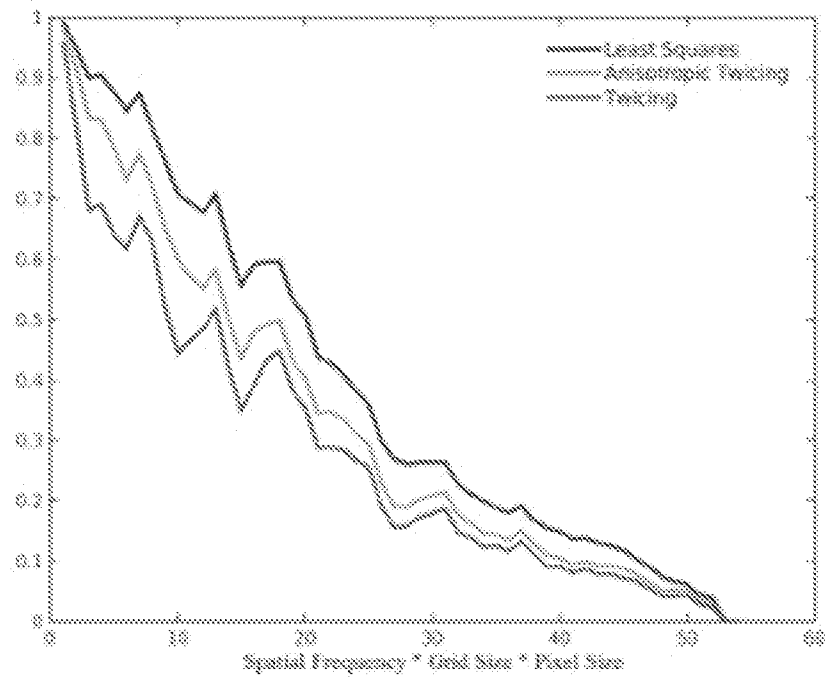
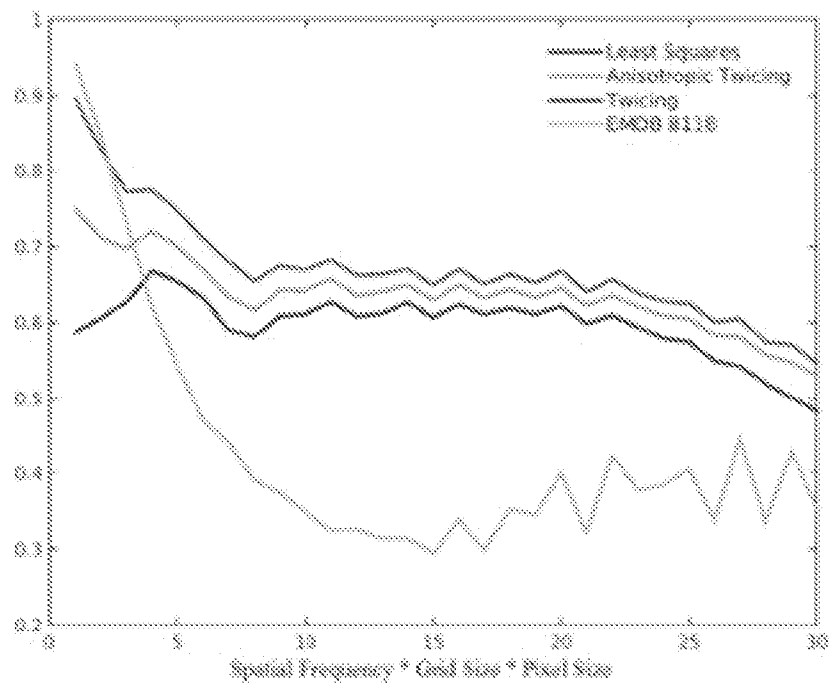


FIG. 9C

**FIG. 10A****FIG. 10B**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 18/29152

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - C07H 21/00; C07K 14/00; C12N 15/74; C12N 5/10 (2018.01)

CPC - C07K 14/001; C07K 14/24; C07K 14/36; C07K 14/47; C07K 14/7151; C07K 19/00; G01N 33/6803; H01J 37/20; H01J 37/222; H01J 37/261; H01J 37/295; C07K 2319/00; H01J 2237/201; H01J 2237/2802

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2017/0103161 A1 (The governing council of the university of Toronto) 13 April 2017 (13.04.2017) entire document especially abstract; para [0003]; figure 1-4	1-15
A	Spahn et al., 'Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM Spahn', Current Opinion in Structural Biology, Volume 19, Issue 5, October 2009, Pages 623-631	1-15
A	US 2010/0010946 A1 (De Winter et al.) 14 January 2010 (14.01.2010) entire document	1-15

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

12 December 2018

Date of mailing of the international search report

31 DEC 2018

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents

P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-8300

Authorized officer:

Lee W. Young

PCT Helpdesk: 571-272-4300

PCT OSP: 571-272-1114