



- (51) International Patent Classification:
H01J 49/26 (2006.01)
- (21) International Application Number:
PCT/US2012/024362
- (22) International Filing Date:
8 February 2012 (08.02.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
61/440,463 8 February 2011 (08.02.2011) US
- (71) Applicant (for all designated States except US): **INDIANA UNIVERSITY RESEARCH AND TECHNOLOGY CORPORATION** [US/US]; 351 West 10th Street, Indianapolis, IN 46202 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **ORTOLEVA, Peter, J.** [US/US]; 1105 Brooks Drive, Bloomington, IN 47401 (US). **JARROLD, Martin** [US/US]; 3750 East Bluebird Lane, Bloomington, IN 47405 (US).
- (74) Agent: **ADDISON, Bradford, G.**; Barnes & Thornburg LLP, 11 South Meridian Street, Indianapolis, IN 46204 (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK,

[Continued on next page]

(54) Title: ION MOBILITY SPECTROMETRY AND THE USE OF THE SEQUENTIAL ELIMINATION TECHNIQUE

(57) Abstract: The present disclosure describes a method of acquiring data and analyzing that data. The disclosure describes how useful information can be extracted from a collection of data; for example, how useful information can be extracted from data generated using IM-MS on a large macromolecular assembly. A method for analyzing IM-MS data is described which uses a multiscale approach providing a new paradigm for analyzing IM-MS data where simulations are guided to states consistent with the measured data.

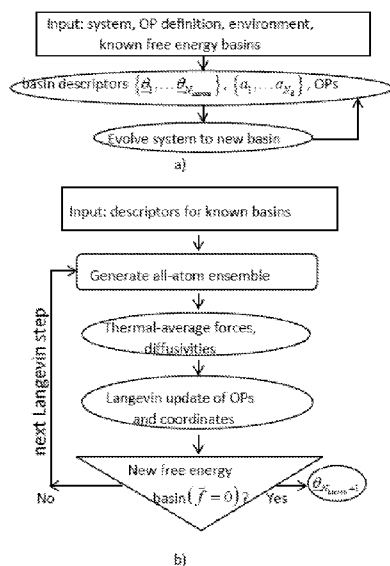


FIG. 6



SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

— of inventorship (Rule 4.17(iv))

Declarations under Rule 4.17:

Published:

— as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

— without international search report and to be republished upon receipt of that report (Rule 48.2(g))

— as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))

ION MOBILITY MASS SPECTROMETRY AND THE USE OF THE SEQUENTIAL ELIMINATION TECHNIQUE

This application claims the benefit of U.S. Provisional Application 61/440,463, filed February 8, 2011.

BACKGROUND

[0001] The present disclosure relates to ion mobility mass spectrometry (IM-MS) and the analysis of supramolecular assemblies. The present disclosure further relates to the collection and analysis of ion mobility mass spectrometric data and the analysis thereof for the determination of the structure of supramolecular assemblies of biological molecules.

[0002] It has long been recognized that many important functions of a cell are performed by supramolecular assemblies. However, it is only recently that the scale and complexity of the protein-protein interactions occurring within an organism have been fully appreciated [1-3]. It is now thought that most protein functions are carried out through multicomponent complexes and (in most cases) the assembly, function, and regulation of these complexes remains unknown. For those that are known, their large physical size makes them difficult to characterize from a structural point of view, and little is known about the spatial arrangement of their components.

[0003] Ion mobility mass spectrometry (IM-MS) has shown promise for analyzing low resolution structures of supramolecular assemblies [4-6]. For example, Robinson and collaborators used IM-MS to investigate the subunit architecture of the human eukaryotic initiation factor 3, for which there is no high resolution structure [7]. Heck and collaborators used IM-MS to probe the T=3 and T=4 capsids of Hepatitis B and found two structures with mobilities that differ by 4% for T=3 [8]. In related work, Ashcroft and collaborators recently used IM-MS to investigate structures of the oligomeric species formed in early stages of fibril assembly, and suggested that the oligomers adopt elongated geometries [9].

SUMMARY

[0004] Ion-mobility mass spectrometry (IM-MS) may be used to generate mass spectral data quickly using relatively small sample sizes. The present disclosure describes a method of acquiring mass spectral data using IM-MS and the interpretation of that data.

[0005] In one aspect, the present disclosure addresses the issue of how useful information can be extracted from a collection of data. In an illustrative embodiment, the present disclosure addresses the issue of how useful information can be extracted from data generated using IM-MS on a large macromolecular assembly. In illustrative embodiments, a method for analyzing IM-MS data is described which uses a multiscale approach. In particular, disclosed is a new paradigm for analyzing IM-MS data where simulations are guided to states consistent with the measured data.

[0006] In further illustrative embodiments, a method for analyzing IM-MS data comprises an ion mobility sequential elimination technique (IMSET). IMSET is a new paradigm for analyzing IM-MS data. In one embodiment, IMSET automatically seeks free-energy basins which (1) have structure compatible with the observed cross section; and (2) were not discovered in an earlier step in the sequential basin discovery process. One aspect of IMSET is that it saves an enormous amount of time that would have been spent randomly searching regions of assembly configuration space not relevant to the experimental measurements. While the present disclosure specifically relates to data obtained through ion-mobility mass spectrometry, the method of applying a sequential elimination technique to analytical data has applications in transforming data generated through other characterization methods (e.g., chemical labeling [45], nanofluidics [46] and AFM).

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] So the manner in which the features of the present disclosure can be understood in detail, the accompanying figures provide additional information with respect to an exemplary embodiment.

[0008] FIG. 1 is a schematic showing a loop in which order parameters characterizing nanoscale features affect the probability of atomistic configurations which, in turn, determine the forces driving order parameter dynamics;

[0009] FIG. 2(A-B) are ribbon structure depictions of a simulation showing the *Satellite Tobacco Mosaic Virus* RNA structural transition (A) at 0 ns and (B) at 55 ns, showing the speed-up over an ensemble of conventional MD is 6-11 fold;

[0010] FIG. 3(A-B) are ribbon structure depictions of a simulation showing the structural transition in *Cowpea Chlorotic Mosaic Virus*, wherein (A) shows the initial T=3 symmetric state; and (B) shows the symmetric swollen state after 18 ns of simulation time; although the transition starts from, and ends in a T=3 state, the transition proceeds through states of lower symmetry; because of the symmetry breaking this transition cannot be simulated using symmetry-constrained methods;

[0011] FIG. 4(A-B) are ribbon structure depiction of a simulation showing disassembly/collapse of a capsid-like structure consisting of 60 copies of the *Human Papilloma Virus L1 protein*; wherein (A) shows the initial state and (B) shows the collapsed state after 100 ns; this instability has been observed in experiments and results from selected L1-helix truncations; and

[0012] FIG. 5 is a graphical representation showing a ribbon structure depiction and a graphical representation of moment of inertia eigenvalues inset within a graph showing energy vs. iteration number for a preliminary IMSET simulation for Lactoferrin showing the rise and fall of the average potential energy accompanying transition from closed to open state; also showing the increase in the molecular descriptor (notably the major eigenvalue of the moment of inertia tensor reflecting predominant expansion along the X-axis); showing that as the method is refined as described herein, the transition plot yields an accurate barrier height.

[0013] FIG. 6 shows workflow of the DMS.BD algorithm that enables traversal of FE barriers and discovery of new basins. (a) Input includes initial all-atom structure in solvent, definition of basis functions and OPs, size of the Langevin timestep, update frequency for the reference structure, and conditions in the host medium. Discovery of a new FE basin starts with establishing descriptors and values of OPs at the bottom of known basins, and choosing width parameters for the Δ^- factors. (b) Flowchart for evolution to new basin via guided Langevin dynamics. For detailed explanation of each step see first paragraph of second subsection in Supporting Information.

[0014] FIG. 7 presents the following types of human lactoferrin structures: crystallographic (a), (b); discovered basin bottoms (c), (e) and (f), (h); and transition point (d), (g). These are (a) closed 1LFG X-ray structure, (a') 1LFG MD used to start DMS simulation; (b) open 1LFH; (c) at the bottom of basin 1; (d) transition point along the basin 1→2 pathway; (e) arbitrarily chosen Langevin timestep from basin 2 (called “descent 2” structure, see Appendix D); (f) bottom of basin 2; (g) basin 2→3 transition point (starting at the “descent 2” structure); (h) bottom of basin 3. All these structures are of lowest potential energy among those in the ensemble consistent with the instantaneous OP values. A transition point between basins is taken to be at the Langevin timestep for which the potential energy goes through a maximum. Such points are close to transition regions where FE force changes sign (see paragraph 6 in Validation for Human Lactoferrin, below).

[0015] FIG. 8 shows DMS.BD Langevin timecourses for descriptors showing distinct differences between the basins. (a) Eigenvalue 1, (b) eigenvalue 2, and (c) eigenvalue 3 of the moment of inertia tensor for human lactoferrin. Basin 1—2 transition (bottom); control simulation launched from basin 2 (top in (a) and middle in (b)-(c)); basin 2-3 transition proving robustness of the basin discovery method (middle in (a) and top in (b)-(c)) (Appendix D). The eigenvalues remain fairly constant at the bottom of basins (Figure 14), and change during inter-basin transitions. Eigenvalues 2 and 3 behave similarly during basin 1—2 transition; however, this similarity is lost in the next transition. If one starts in basin 1, only modest changes in descriptors are observed. In contrast, when precluding basins 1 and 2 in going to basin 3, descriptors change much more implying greater extent of lobe opening.

-5-

[0016] In FIG. 9 thermal-average forces $\bar{f}_k$ are shown along a transition path from basin 1 to 2: (a) forces of maximum magnitude showing clear inter-basin transition pattern ($k = 100X = \text{top line at } 30, k = 010Y = \text{middle line at } 30, k = 001Z = \text{bottom line at } 30$) (b) second-highest amplitude ones oscillating around zero ($k = 001X = \text{top line at } 8, k = 001Y = \text{middle line at } 8, k = 100Z = \text{bottom line at } 8$). The figure suggests that the FE minimizing tendency of lactoferrin in basin 1 makes it contract in z-direction and expand in x-direction (implied by the negative f_{001Z} and positive f_{100X}). During barrier crossing, the sign of these forces changes to positive along z and negative along x direction. This leads to further expansion of lactoferrin along z direction, and hence lobe opening in basin 2. (c) Contrasting behavior of thermal-average f_{100X} (top line at 30) and biasing f_{100X}^b (bottom line at 30) forces. f_{100X}^b is maximum near the bottom of basin and gradually decreases as the system escapes from the FE minimum as indicated by the increase in f_{100X} . Insert shows that f_{100X} computed from MD simulations are small and random when sampling structures in basin 1 and, therefore, do not drive transitions between basins.

[0017] FIG. 10 shows energy timecourse of lowest potential energy structures of human lactoferrin generated from constant OP ensembles during discovery of FE basins. Line styles and simulations are the same as in Figure 8 (lowest line at 80 and 110 is basin 2).

[0018] FIG. 11 shows potential energy timecourse of lactoferrin from MD sampling simulations starting at structures near the bottom of discovered basins. (top) Basin 1, potential energy minimum of basin 1 was achieved at time instance $t = 9.940$ ns; (below top and overlapping line below) "descent 2" structure from basin 2, with potential energy minimum at $t = 3.686$ ns; (above bottom and overlapping line above) lowest-energy basin 2 structure, with potential energy minimum at $t = 3.62$ ns; (bottom) basin 3, with potential energy minimum at $t = 2.733$ ns. Distinct energy bands indicate that each MD trajectory is confined to a given basin.

[0019] FIG. 12 shows residue-by-residue RMSD relative to closed-lobe diferric X-ray structure 1LFG averaged over 25 residue window: (bottom at 250) diferric basin 1, (2d from bottom at 250) pseudodiferric basin 2, (3d from bottom at 250) open-lobe apo-like basin 3, and

(top) the apolactoferric structure 1LFH. This analysis implies that the lactoferrin gradually opens during DMS.BD simulation exploring a range of states from diferric to apolactoferric character.

[0020] FIG. 13 shows Ramachandran plots for discovered lactoferrin structures: (darker) bottom of basin 1, (lightest) top of the barrier for basin 1—*2 transition, and (lighter) bottom of basin 2. Most residues have dihedral angles within theoretically favored regions [60°] shown as background shadows, which validates that the configurations obtained from DMS.BD simulation retain key secondary structure. Distribution of qnv angles is distinct for the three structures. Arrows indicate large changes in dihedral angles for hinge residues: (horizontal) THR90, (vertical) VAL250.

[0021] FIG. 14 shows descriptors and their ranges in each of the three discovered FE basins are shown below as an additional evidence of crossing the free energy barriers and resulting distinction between basin characteristics. In a) at 1 ns, the top curve is basin 3; the next curve is basin 2, the next curve is "descent 2", and the bottom curve is basin 1. In b) at 1 ns, the top curve is basin 3; and the other curves overlap with "descent 2" on top, basin 2 in the middle and basin 1 on the bottom. In b) at —3.5 ns, the top curve is basin 3; the next curve is "descent 2"; and the basin 2 and basin 1 curves overlap. In b) from —5.5 ns to 10 ns, the order of curves from the top is basin 3, "descent 2", basin 2, basin 1, with various overlaps.

DETAILED DESCRIPTION

[0022] *Cross section Calculations*

[0023] The mobility is a measure of how rapidly an ion (or macroion) travels through an inert buffer gas under the influence of a weak electric field. The low field mobility provides a value for the average collision cross section, and structural information has been obtained by comparing the measured cross sections with those calculated for trial geometries. To be rigorous, the cross section Ω referred to above is really an orientationally averaged collision integral which is calculated by averaging the momentum transfer cross section over the relative velocity and collision geometry [10]:

$$\Omega = \frac{1}{4\pi^2} \int_0^{2\pi} d\theta \int_0^\pi d\phi \sin\phi \int_0^{2\pi} d\gamma \frac{\pi}{8} \left(\frac{\mu}{k_B T} \right)^3 \int_0^\infty dg g^5 e^{-\mu g^2 / 2k_B T} \int_0^\infty db 2b (1 - \cos\chi) \quad (1)$$

[0024] In the preceding equation, θ , φ , and γ define the collision geometry (the orientation of the ion with respect to the incoming buffer gas atom), g is the relative velocity, μ is the reduced mass, and b is the impact parameter. The last integral in Eqn. (1) is the momentum transfer cross section which depends on the scattering angle, χ , the angle between the ion and buffer gas atom trajectories before and after a collision. The most rigorous approach for determining the cross section of a polyatomic assembly is to propagate trajectories via the complete many-atom potential. Then, the collision integral is obtained by averaging over relative velocity and collision geometry as in Eqn. (1).

[0025] There are currently three main methods used to calculate orientationally averaged collision integrals for comparison with the measured values:

[0026] 1) The Projection Approximation (PA) [11-12]. Here, the momentum transfer cross section is approximated by the two-dimensional projection (i.e., the shadow) of the ion. This approach ignores the details of the scattering process (i.e., scattering angles are not calculated) and neglects long-range interactions.

[0027] 2) The Exact Hard Spheres Scattering (EHSS) model [13]. In this approach, all the atoms are replaced by hard spheres and the scattering angles are determined from hard sphere scattering. The long-range interactions between the ion and buffer gas atom are ignored.

[0028] 3) The Trajectory Method (TM) [14]. Here, the interactions between the ion and the buffer gas atom are approximated by an effective potential consisting of a sum of two-body Lennard-Jones interactions and charge-induced dipole interactions. The scattering angles are determined from trajectory calculations.

[0029] All three methods employ Monte Carlo (MC) techniques to perform the integrations over the orientation and impact parameter. In the case of the projection approximation, the MC integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates. The cross section is obtained from the fraction of trajectories that strike the ion, averaged over orientation. The EHSS model uses a similar approach, except that the scattering angle is calculated and used to determine the collision integral. With the EHSS model the buffer gas atom can undergo more than one collision

with the ion. Such multiple scattering events are more common for large ions, particularly those with concave surfaces. The implementation of the trajectory method follows a similar recipe to the EHSS model. Multiple collisions also occur with the TM model.

[0030] As with all MC integration methods, the accuracy of the final result depends on the number of trials that are performed. In the present case, 10^6 - 10^7 trajectories are required for the collision integral to converge to the accuracy required for ion mobility studies (i.e., an uncertainty of less than 1%). Algorithms for the PA and the EHSS model have recently been optimized [15]. Even with the unoptimized code, converged PA and EHSS cross sections can be calculated for large protein complexes in a couple of hours. However, the TM method is a different matter. It is much more computationally expensive to implement, and calculating the TM cross section even for peptides is a challenge.

[0031] Careful studies for small systems [14-16] show that the projection approximation tends to underestimate the collision integral while the EHSS model tends to overestimate it. For small systems, the PA and EHSS cross sections tightly bound the true cross section, but as the system becomes larger the cross sections obtained from these two models diverge. For example, for GroEL (PDB ID: 1XCK), the PA cross section ($21,303 \text{ \AA}^2$) and the EHSS cross section ($29,033 \text{ \AA}^2$) differ by 36%. That the cross section estimated from experiments ($24,384 \text{ \AA}^2$) [17] lies between these two extremes, provides little useful information. It is not possible to tell whether the complex has retained its solution phase conformation, collapsed, or unfolded. If the PA model provides an accurate value for the cross section, then the measured cross section is larger than the value obtained from the PDB file. This suggests that the GroEL complex expands in the gas phase. On the other hand, if the EHSS model gives the correct cross section then the measured cross section is smaller than the value deduced from the PDB file, suggesting that the complex becomes more compact in the gas phase.

[0032] The difference between the PA and EHSS cross sections results from multiple scattering which contributes to the EHSS cross sections but not the PA ones. For objects with entirely convex surfaces (where multiple scatterings cannot occur) [13] the two models predict the same cross section. Opportunities for multiple scattering phenomena increase with ion size (which is why the EHSS and PA cross sections diverge with increasing size). However, it is

known from studies of small systems that the EHSS model overestimates the contribution from multiple scattering, probably because the hard sphere surface is too rough. On the other hand, the PA completely ignores multiple scattering and so it certainly underestimates this effect. The true cross section must lie between these two extremes, but in order to know where (i.e., to determine a reliable and accurate value for the cross section) it is necessary to perform calculations using the trajectory method.

[0033] The trajectory method provides the most accurate values for the collision cross sections. It correctly incorporates multiple scattering and employs a realistic potential that includes long-range attractive interactions as well as short-range repulsive ones. However, it is computationally expensive, and its use has been restricted to small systems. To make meaningful quantitative comparisons with measured cross sections, the trajectory method must be extended to much larger systems. We describe how this goal will be achieved in Sect. III.A.

[0034] *Locating Low Free-Energy Geometries with Cross sections that Match Experiment*

[0035] As described above, structural information has been obtained from ion mobility measurements by calculating cross sections for trial geometries, and then comparing the calculated values to measured cross sections. For small systems, the trial geometries have usually been obtained from MD or MC simulations [18]. However, with MD in particular, the trajectories often become trapped in a small region of the available phase space because there is significant activation barriers associated with substantial structural changes. Raising the temperature can help to surmount activation barriers, however, raising the temperature also makes it more difficult to access low entropy configurations. MC methods suffer from the same problems. The use of replica exchange [19] and genetic algorithms (GA) [20] has shown promise for small systems. However, for the large systems of interest here, the computational times for major structural changes become long, making MD, MC and GA impractical. Thus, the methods that have been used for small systems (with only limited success in some cases) cannot be used for the large protein complexes of interest here. One aspect of the present disclosure is that this computer time limitation can be solved using a multiscale approach [21-37]. Multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes

coupled across scales in space and time. According to one aspect of the present application, rapidly fluctuating atomic configurations yield the free-energy driving forces mediating slow structural transitions in macromolecules. This framework is the basis of the macromolecular simulation methodology to be used in this project.

[0036] There is an extensive literature on the use of dimensionality reduction techniques to accelerate macromolecular simulations. These include principal component analysis and normal modes to identify collective behaviors [38-40], curvilinear coordinates to characterize macromolecular folding and coiling [33], models wherein a peptide or nucleotide is represented by a bead which interacts with others via a phenomenological force [41], and shape-based coarse-grained models [42-44]. Coarse-grained and lattice models use "pseudo-atoms" to represent groups of atoms. These require careful calibration of new parameters in addition to the originally developed and calibrated MD parameters, increasing the level of approximations in these simulations. Likewise, simulations of processes on long timescales (beyond about 1 microsecond) may incur numerical inaccuracies arising from machine or algorithmic precision. These systematic errors may cause the results to deviate from observed experimental values. These approaches have lead to important insights but, in the present context, suffer from one or more of the following difficulties: (1) the characteristic variables are not slowly varying in time and thus cannot serve as a basis for a multiscale approach; (2) macromolecular twisting and branching are not accounted for; (3) inelasticity and frictional interactions are neglected; (4) recalibration is required for most new application; and (5) multiple minima cannot be probed.

[0037] The use of multiscale methods will allow us to extend the strategy used to analyze IM-MS data for small systems to large supramolecular assemblies. However, the strategy of identifying trial geometries using simulations, calculating cross sections, and then comparing them to measured values is inefficient; furthermore it may miss the relevant structure altogether. Here we propose an entirely new approach; we will develop a search algorithm that directly identifies conformations that have cross sections which agree with measured ones.

[0038] In illustrative embodiments, the software package MOBCAL is used to calculate collision cross sections for IM-MS measurements. In one embodiment, MOBCAL incorporates all three methods mentioned above (PA, EHSS, and TM). It was written by us in the late 1990s

and it has been available for free from our website (nano.chem.indiana.edu) for the last decade. We support the program by helping people install it and use it, and by answering general questions. There are numerous MOBCAL users around the world.

[0039] In illustrative embodiments, MOBCAL may be upgraded so that IM-MS measurements for large assemblies can be interpreted reliably. Without this, the measurements may be misinterpreted. With the availability of the faster and more reliable models to calculate the cross sections (provided herein) it will be possible to extract more information from the IM-MS measurements (i.e., interpret smaller differences between measured and calculated cross sections). The strategies that will be used to accelerate cross section calculations have not been used in the present context before. Details are provided Sect III. A.

[0040] *Locating Low Free-Energy Geometries with Cross sections that Match Experiment*

[0041] The present disclosure particularly relates to the use of multiscale methods to interpret IM-MS data. Illustratively, this multiscale method may be used for the analysis of the macromolecular complexes of interest here because MD, MC, and GA simulations are too slow for the large assemblies currently being studied by IM-MS methods. Accordingly, one aspect of the present disclosure is that the method of interpreting IM-MS data does not rely on MD, MC or GA simulations.

[0042] In illustrative embodiments, the method comprises an IMSET (Ion Mobility Sequential Elimination Technique) approach that provides a new paradigm for analyzing IM-MS data. In one embodiment, IMSET automatically seeks free-energy basins which (1) have structure compatible with the observed cross section; and (2) were not discovered in an earlier step in the sequential basin discovery process. In another embodiment, this method saves substantial time that would have been spent randomly searching regions of assembly configuration space not relevant to the experimental measurements. IMSET has applications in other characterization methods as well (e.g., chemical labeling [45], nanofluidics [46] and AFM).

[0043] While not limited to a particular theory, it is of note that there have been major advances in the theory of macromolecular folding and in sampling configurations in free-energy basins [47-51]. IMSET combines elements of (1) our multiscale analysis [25,28,30-32,35-36,52-

58]; (2) the notion of a stepwise procedure that precludes evolution into basins of attraction identified in earlier steps in the computation; (3) an order parameter method for simplifying the free-energy landscape to eliminate thermally irrelevant basins of attraction [54,56-59]; and (4) a highly optimized algorithm for computing cross sections from ensembles of atomistic structures. Any free-energy landscape exploration approach requires a degree of coarse-graining of the original N -atom potential. Our order parameters provide a natural and general way to achieve this [36,54,60].

[0044] In illustrative embodiments, the method of analyzing comprising multiscale theory starts with a set of order parameters Φ characterizing the overall structure of a macromolecular assembly. In one embodiment, changes in Φ describes the coherent motion of many atoms simultaneously. This implies that they evolve much more slowly than the fluctuations of individual atoms. This timescale separation enables one the derivation of Langevin-type equations for stochastic order parameter dynamics [21-23,25-28,31-32,34-35,53-54,61]. By coevolving the order parameters with the quasi-equilibrium probability density implied by the theory, one embodiment of the method captures the coupling of processes across scales in space and time that we believe to be the hallmark of nanosystem dynamics (FIG. 1) [30,62]. For example, FIG. 1 is a schematic showing a loop in which order parameters characterizing nanoscale features affect the probability of atomistic configurations which, in turn, determine the forces driving order parameter dynamics.

[0045] The types of order parameters we have used via this approach include structural variables [24,28,37,53-54,58,60], curvilinear coordinates [33], scaled atomic positions [26,29], density-like field variables [27,32,35-36,61], and major subcomponent conformations [30]. These systems can involve a discrete or quasi-continuous set of time scales [63].

[0046] Order parameter dynamics is driven by the gradient of the Hemholtz free-energy F with respect to Φ , i.e., by thermal-average forces. However, $F = E - TS$ where the entropy S depends on information available to the observer, and E is the system energy. Entropy always incorporates our knowledge of the system (e.g., we know the system is in a given volume of space). Thus, in our approach we incorporate knowledge of free-energy basins of attraction already identified in previous steps of a sequential basin discovery process. The algorithm

suggested by the above allows one to generate a search for free-energy minima without constructing the entire landscape. This is significant because it is not practical to map out the global landscape for large assemblies. The algorithm we develop remains local in character, proceeding via the order parameter dynamics equations (Section III.B) in a manner that increasingly limits configuration space as more and more basins are automatically discovered sequentially.

[0047] One aspect of making the above basin discovery scheme practical is an efficient algorithm for generating an ensemble of atomistic configurations at fixed values of order parameters, and for efficiently simulating many-atom system via evolutionary order parameters. One aspect of the present disclosure is a unique algorithm of this type. Illustratively, it has been implemented as the *SimNanoWorld*TM software [57-58]. Examples of *SimNanoWorld*TM simulations of large systems are provided in FIG. 2 through FIG. 4.

[0048] For example, FIG. 2(A-B) are ribbon structure depictions of a simulation showing the *Satellite Tobacco Mosaic Virus* RNA structural transition (A) at 0 ns and (B) at 55 ns, showing the speed-up over an ensemble of conventional MD is 6-11 fold. FIG. 3(A-B) are ribbon structure depictions of a simulation showing the structural transition in *Cowpea Chlorotic Mosaic Virus*, wherein (A) shows the initial T=3 symmetric state; and (B) shows the symmetric swollen state after 18 ns of simulation time; although the transition starts from, and ends in a T=3 state, the transition proceeds through states of lower symmetry; because of the symmetry breaking this transition cannot be simulated using symmetry-constrained methods. FIG. 4(A-B) are ribbon structure depiction of a simulation showing disassembly/collapse of a capsid-like structure consisting of 60 copies of the *Human Papilloma Virus L1 protein*; wherein (A) shows the initial state and (B) shows the collapsed state after 100 ns; this instability has been observed in experiments and results from selected L1-helix truncations.

[0049] Another aspect of the present disclosure is that we have disclosed a manner in which an advanced version of this software which integrates information on basins that have already been identified in previous steps of our IMSET algorithm may be created. The proposed algorithm for stepwise free-energy basin discovery is presented in Sect. III.B along with a detailed roadmap for implementation and demonstration. To place the IMSET approach in

perspective, consider the following. The use of experimental data in the statistical mechanical analysis of many-atom systems dates back to the notions of Gibbs. Gibbs set forth the ensemble method which implies that the probability of atomistically resolved states depends on conditions (i.e., constraints) to which the system is subjected (e.g., isothermal versus insulated, isobaric versus isovolumetric). The modern expression of this concept is the entropy maximization principle of Jaynes [64], where the probability of atom-resolved configurations is that which maximizes entropy constrained by observed information. Illustratively, we disclose an adaption of the entropy maximization principle to the nanosciences via multiscale analysis [24-25,27-36,53-55, 58-59,61-63,65-80]. To lowest order in the multiscale theory, the probability of atomistic configurations (i.e., the positions and momenta of N atoms) has the form ρW , where ρ is the conditional probability for atomistic configurations given that the order parameters have specified values, while W is the time-dependent probability of the state of the order parameters characterizing nanometer-scale features of a macromolecular system. W evolves slowly relative to the timescale of individual atomic vibrations/collisions. For most systems of interest, W satisfies a Smoluchowski equation. It has been shown that solving this equation is equivalent to simulating an ensemble of order parameter timecourses generated via a set of Langevin equations [28].

[0050] Another aspect of the present disclosure is that the experimental data influences the Langevin evolution, i.e., drives the dynamics along trajectories consistent with our knowledge of the system. In the present context, it drives the trajectory towards structures with the measured cross section(s). This technique is distinct from Elber's milestone method, where milestones are defined on a reaction-pathway and simulations are performed about each of the milestones [81]. First-passage time distributions for reaching neighboring milestones are calculated and the segments are joined together to yield a long-time trajectory. This method cannot be applied when the reaction path is unknown. In contrast, our approach does not require knowledge of the reaction path. We accomplish this by automatically sampling constant order parameter ensembles to calculate thermal-average forces and diffusion factors, and self-consistent Langevin noise to construct the long-time trajectory of the system, providing both the state of the order parameters and the quasi-equilibrium distribution of atomic configurations

consistent with it. We automatically probe multiple timecourses by our constant OP ensembles of atomistic configurations [28,37,53,58]. In this sense, our methodology provides all the information of ensemble MD [57] but carries out the calculation with an increase of several orders of magnitude in computational efficiency.

[0051] Yet another aspect of the present disclosure is that the method of analysis includes using our software for (1) calculating the cross section for a given all-atom nanostructure will be optimized, and (2) similarly for our *SimNanoWorld*TM software. Subsequently the new algorithm for calculating cross sections with *SimNanoWorld*TM is integrated within the method and our sequential elimination algorithm to create the IMSET software for delineating the full set of free-energy minimizing structures of a macromolecular assembly that are consistent with observed cross section data. We will then demonstrate IMSET via a number of macromolecular assemblies, assess accuracy and speed, make necessary upgrades, and make IMSET available to experimental groups.

[0052] Using the described approach, we will create the next generation of our MOBCAL program (MOBCAL2) to enable the trajectory method (TM) calculated cross sections for large macromolecular assemblies. To accomplish this, we will reduce the computer time to calculate each trajectory. The most time-consuming part of the calculation is evaluation of the effective potential. In MOBCAL, the effective potential is presently evaluated by summing Lennard-Jones and ion-induced dipole interactions from all atoms in the assembly. The computer time required to sum these contributions will be shortened substantially if, rather than summing over all atoms, we bundle groups of atoms and assign an effective potential to the group. This can only be implemented for groups of atoms that are well-removed from the buffer gas atom undergoing the collision. When the group is near the buffer gas atom, its contribution to the potential will be determined as before by summing over contributions from individual atoms. Thus, the overall potential will be given by the sum of contributions from individual atoms within a distance d_s of the buffer gas atom, plus the sum of the contributions from groups of atoms beyond d_s .

[0053] We follow a trajectory to illustrate how this will work, starting from when the buffer gas atom is far away from the assembly. At large distances between the buffer gas atom

and the assembly, the effective potential is obtained by summing over the group contributions. As the distance between the buffer gas and the assembly decreases, the groups nearest to the buffer gas atom transition, and their contribution to the effective potential will be obtained by summing contributions from individual atoms. After the collision as the buffer gas atom moves away, the reverse process occurs and the effective potential at larger distances will be obtained by summing over the group contributions.

[0054] Another aspect of the present disclosure is that the overall effective potential changes smoothly as the atoms transition between being treated individually or as a group; otherwise the trajectory will be discontinuous and energy will not be conserved. To this end, we will gradually make the transition using a switching function like

$$E(d) = \frac{E_{GROUP}(d)}{1 + \exp\left(\frac{d_s - d}{S}\right)} + \frac{\sum E_{ATOM}(r)}{1 + \exp\left(\frac{d - d_s}{S}\right)} \quad (2),$$

[0055] where $E_{GROUP}(d)$ is the energy of the group of atoms at a distance d from the buffer gas atom, $\sum E_{ATOM}(r)$ is the energy of the group of atoms from summing over the individual atomic contributions, r is the distance between the individual atoms and the buffer gas atom, d_s is the distance where the transition occurs, and S is a parameter that controls how gradually the transition occurs. Both d_s and S will be selected to minimize the range of distances where the potential must be calculated using all atoms, while preserving accuracy.

[0056] Yet another aspect of the present disclosure is the incorporation of coarse-graining concepts. For disclosure related to implementation in other systems, reference is made to the references [42-44]. Within the present implementation, we will group atoms in the assembly using several methods, and will determine which approach optimizes the balance between efficiency and accuracy. In one approach they will be grouped by residue. In this way, it will be possible to establish an average set of group potentials for each amino acid. For residues that are far from the buffer gas atom, we will investigate an even coarser graining, where groups of residues are bundled together and assigned an overall potential. In this case, residues that are close in the three-dimensional structure will be grouped, not necessarily residues that are close in the sequence. This second tier grouping will be less general, but it will save a lot of computer

time in selected cases. For example, for a virus, all the residues from a capsid protein on the opposite side of the virus from the buffer gas atom can be bundled together to give a potential for the whole protein. We will carry out an extensive trial and error procedure to find optimal coarse and fine-graining strategies and the ranges over which they will be used. Balancing accuracy and speed will be the main objective of this facet of the project. In addition to these improvements, we will parallelize the code and configure it to run trajectories simultaneously on clusters of CPUs.

[0057] Despite improvements in the speed of TM cross section calculations, this process will still consume considerable computer resources. In order to accomplish our main goal, automated discovery of free-energy minimizing structure consistent with a given measured cross section, we need to perform a large number of cross section calculations. Thus, we will require further acceleration in the cross section calculations. To accomplish this goal, we will develop an empirical model that uses the PA and EHSS cross sections to predict the TM cross sections. We have developed a similar model for peptides [82]. In that work, we used the expression

$$\Omega_{CCS} = a + b\Omega_{EHSS} + c\Omega_{EHSS}^2 + d(\Gamma - 1) + ez \quad (3)$$

[0058] where Ω_{EHSS} is the EHSS cross section, Γ is an asymmetry parameter that measures how distorted the geometry is from spherical, and z is the charge. While this model is not expected to work for large assemblies, we are confident that an empirical relationship can be derived for large complexes to provide an accurate value of the cross section. For structures with cross sections close to the measured value we expect to repeat the cross section calculations with the TM method.

[0059] *Locating Low Free-Energy Geometries with Cross sections that Match Experiments*

[0060] In one embodiment, the method includes the implementation of descriptors. In one embodiment, IMSET will implement a set of molecular descriptors Ψ that characterize the geometry of a macromolecular assembly regardless of its position or orientation. For example, these can be computed from the atomic configurations provided by *SimNanoWorld*TM. Illustratively, these descriptors are used as a computational device, and therefore experimental data on them is not required as input to IMSET. In the proposed implementation of IMSET, we

-18-

will use mass and charge descriptors. However, other descriptors will be introduced if further information is needed to drive simulated conformations away from free-energy minimizing structures already known from earlier stages of an IMSET calculation. We will use total mass and the three eigenvalues of the mass-moment tensor $M_{\alpha\alpha'}$,

$$M_{\alpha\alpha'} = \sum_{i=1}^N m_i s_{i\alpha} s_{i\alpha'}, \quad (4)$$

[0061] for components $s_{i\alpha}$ of the relative position $\vec{s}_i$ along the $\alpha = x, y, z$ Cartesian directions, and mass m_i of atom i ($i = 1, 2, \dots, N$), for the N -atom assembly. We will also use the total charge, the length of the dipole moment vector and the three eigenvalues of the charge quadrupole moment $Q_{\alpha\alpha'}$ (defined in analogy with Eqn. (4) except with masses replaced by charges). With this set of nine descriptors Ψ molecular spectral peaks will be related to structure. A larger set of descriptors related to higher moments can also be used if further information is needed to resolve geometries.

[0062] *Modified Thermal Forces*

[0063] As already implemented in *SimNanoWorld*TM, we introduce a set of order parameters (OPs) Φ that characterize overall features of a macromolecular assembly and which are shown to evolve slowly [28,37,53, 58,60]. Using multiscale analysis, we showed that Φ satisfies Langevin equations that are the basis of the schematic dynamic of FIG. 1 [28,37,53,58]. Algorithms for computing all factors (diffusions and thermal-average forces) in these equations are implemented in *SimNanoWorld*TM. The thermal-average forces driving Φ dynamics are the gradients of free energy with respect to Φ ; they are constructed by our ensemble method [28,37,53,58].

[0064] Here, we will modify this procedure to realize IMSET. Our development starts with the entropy S that takes the modified form

$$S = -k_B \int d\Gamma^* \Delta^{(+)}(\Phi - \Phi^*) \Delta^{(-)}(\underline{\Psi}, \Psi^*) \rho^* \ln \rho^*. \quad (5)$$

[0065] Here k_B is Boltzmann's constant, ρ is the conditional probability of the N -atom configuration Γ given particular values of the OPs Φ , Γ^* is the configuration being integrated over, and $*$ indicates evaluation at Γ^* . The factor $\Delta^{(+)}$ is used in the entropy calculation to only

include configurations consistent with the given values of OPs Φ . The $\Delta^{(-)}$ factor is included to guide the simulations away from the free-energy minima already identified in an earlier stage of an IMSET simulation; for each of these identified basins, there is a set of associated descriptors and $\underline{\Psi}$ is the set of descriptors from the free-energy minimized structures at the bottom of each basins.

[0066] Maximizing S with respect to ρ constrained by average energy (the isothermal condition) and normalization yields $\rho(\Gamma; \Phi; \Psi)$. Given ρ , we construct the thermal-average forces driving Φ dynamics [28, 37, 53, 58]. The result of our deductive multiscale analysis and the construction of constraint quasi-equilibrium probability density ρ , is a Langevin equation for evolution of the OPs; evolution is guided so as to escape/avoid free-energy basins associated with previously-identified free-energy minimized structures. With this, Φ is driven via Langevin evolution to a free-energy minimum that is distinct from those already discovered in previous steps of the IMSET procedure. This algorithm has already been tested; Fig. 5 shows preliminary IMSET simulation wherein the macromolecule starts in one basin and rises up and over the barrier to the next one. For a traditional MD simulation, the barrier would have not been identified or crossed.

[0067] Thus far, we have not mentioned the cross section data Ω . This will be accounted for in IMSET via the method developed earlier [28]. To accomplish this, we add a factor $\Delta(\Omega^{obs} - \Omega^*)$ in the above expression for S :

$$S = -k_B \int d\Gamma^* \Delta^{(+)}(\Phi - \Phi^*) \Delta^{(-)}(\underline{\Psi}, \Psi^*) \Delta(\Omega^{obs} - \Omega^*) \rho^* \ln \rho^*. \quad (6)$$

[0068] The factor Δ is a Gaussian-like function, Ω^{obs} is the observed cross section and Ω^* is related to the all-atom configuration Γ^* via Eqn. (1). This factor guides the simulations to configurations consistent with experimentally observed cross section data.

[0069] *Technical Details on Thermal-Average Forces and Diffusion Factors*

[0070] In the multiscale formulation underlying *SimNanoWorld*TM, the Langevin dynamics of the OPs evolves the system to a free-energy minimum. The Langevin dynamics takes the form

-20-

$$d\Phi / dt = Df + \xi . \quad (7)$$

[0071] Here, ξ is a set of random forces whose statistics are constrained by D ; D is a matrix of diffusion factors and f is the set of thermal-average forces. We proceed in analogy with the method for constructing D and f presented earlier [24,28,58,65]. The method starts with the relationship between the positions of the N atoms relative to the CM ($\bar{s}_i$ ($i = 1, 2, \dots, N$)) and the OPs Φ [53-54]:

$$\bar{s}_i = \sum_k \bar{\Phi}_k U_{ki} + \bar{\sigma}_i . \quad (8)$$

[0072] Here, the U_{ki} are basis functions, the $\bar{\Phi}$ are vector OPs, and the $\bar{\sigma}_i$ are “random” displacements over-and-above the coherent motion due to the dynamic generated by the OPs [60]. D and f are obtained by constructing an ensemble of atomic configurations for fixed $\bar{\Phi}_k$ ($k = 1, \dots, k_{\max}$ for a description in terms of $k_{\max}$ vector order parameters) [28]. Our technique for constructing the residual displacements $\bar{\sigma}_i$ avoids unphysical, high-energy configurations [57-58]. Since the OPs only vary appreciably over timescales much longer than that of atomic collisions/vibrations, we use a hybrid sampling technique where multiple short MD runs are used to enrich an ensemble by generating the configurations for given OPs [58]. The instantaneous OP forces for the set of configurations are then averaged to obtain the thermal-average forces. The $\Delta^{(+)}$ factors are automatically incorporated in our Monte Carlo integration used to calculate the thermal-average forces f by our method of constructing the ensemble of atomic configurations. The $\bar{f}_k$ are computed from

$$\bar{f}_k = \frac{1}{\beta Q} \frac{\partial}{\partial \bar{\Phi}_k} \int d\Gamma^* \Delta^{(+)}(\Phi - \Phi^*) \Delta^{(-)}(\Psi - \Psi^*) \exp(-\beta H^*) \quad (9)$$

for system energy H and for $\beta = 1/k_B T$. $Q(\Phi, \beta)$ is the partition function associated with ρ .

Bringing the $\partial/\partial \bar{\Phi}_k$ within the integral in combination with the chain rule yields an expression for the total thermal-average force. The contribution to the thermal-average force from the gradient of the $\Delta^{(-)}$ factor, $\bar{f}_k^{(-)}$, is given by

-21-

$$\beta Q \bar{f}_k^{(-)} = \int d\Gamma^* \exp(-\beta H^*) \sum_{i=1}^N \frac{\partial}{\partial \bar{r}_i} \{ \Delta^{(+)} \Delta^{(-)} \} U_{ki} . \quad (10)$$

With this, the thermal-average force is equal to that already computed in *SimNanoWorld*TM (except with the $\Delta^{(-)}$ weighting) [58] plus a term $\bar{f}_k^{(-)}$ arising from the gradient of $\Delta^{(-)}$. The $\bar{f}_k^{(-)}$ forces guide the Langevin evolution of the OPs away from the free-energy minimizing structure already discovered in earlier steps in an IMSET simulation. This procedure will readily be implemented by modifying existing modules in *SimNanoWorld*TM, and has already been used to obtain the results of FIG. 5. FIG. 5 is a graphical representation showing a ribbon structure depiction and a graphical representation of moment of inertia eigenvalues inset within a graph showing energy vs. iteration number for a preliminary IMSET simulation for Lactoferrin showing the rise and fall of the average potential energy accompanying transition from closed to open state; also showing the increase in the molecular descriptor (notably the major eigenvalue of the moment of inertia tensor reflecting predominant expansion along the X-axis); showing that as the method is refined as described herein, the transition plot yields an accurate barrier height.

[0073] *Implement Multiple Sets of Order Parameters for Macromolecular Assemblies*

[0074] Simulating macromolecular assemblies such as those noted in Significance presents a challenge for traditional MD methods, and similarly for the use of a single set of OPs via Eqn. (8). Neighboring macromolecules in an assembly may move in different ways during the course of a structural transition, or disassembly/assembly. To address this, we will introduce a set of OPs for each macromolecule involved in the macromolecular assembly. The sets used will be similar to those presented earlier for a single macromolecule [60]. We will introduce sets $\Phi^{(m)}$ of order parameters for each of the M macromolecules in the assembly ($m=1, \dots, M$). We start with the hypothesis that the N -atom probability density ρ depends on the N -atom positions and momenta Γ both directly and, via $\Phi = \{\Phi^{(1)}, \dots, \Phi^{(M)}\}$, indirectly. With this, our multiscale approach yields

$$\frac{d\bar{\Phi}_k^{(m)}}{dt} = \sum_{m'=1}^M \bar{\bar{D}}_{kk'}^{(m,m')} \bar{f}_{k'}^{(m')} + \bar{\xi}_k^{(m)}, \quad (11)$$

the multiple subsystem generalization of Eqn. (7) [30]. These equations will be solved numerically using methods implemented earlier in *SimNanoWorld*TM [27,57-58].

[0075] The multiscale approach in *SimNanoWorld*TM has already provided great simulation efficiency over traditional MD for large protein assemblies [57]. However, in this project we will adapt *SimNanoWorld*TM for use on GPU platforms. In preliminary benchmarking we have found this to yield a factor 2-6 in speed over non-GPU parallelization of *SimNanoWorld*TM.

[0076] *Integration of the Latest Interatomic Force Fields*

[0077] As noted above, conventional MD is used in *SimNanoWorld*TM to enhance ensembles and construct diffusion factors (FIG. 1). However, interatomic force fields present difficulties for MD. The current status of force field development is as follows. Traditionally, MD simulations [83] have been used to study protein dynamics in solution, and are mostly limited to the nanosecond regime. Simulations validating MD results at the sub microsecond timescale against experiments are now extensively performed [84-91]. Recent algorithmic developments and hardware advances [92-93] have made microsecond timescale simulations accessible for biomolecular dynamics with atomic resolution via MD simulations. Thus, with the microsecond timescale reachable in experiment and computation it becomes possible to validate simulated dynamics with experiments. The accuracy of the results depends on a variety of factors including choice of the force field, and treatment of electrostatics (e.g. Particle-mesh-Ewald approaches outperform cut-off and reaction-field approaches). Hydrogen bonds in several systems are not well-described by a given force field. For example, hydrogen bonds in beta sheets are best described by CHARMM force fields, whereas AMBER force fields are best suited for sugars and nucleic acids such as DNA or RNA [94]. For an accurate force field, a good match between experiments and simulations is attained if the timescales of both are similar. However, it still turns out that the best fit to many experimental quantities that probe protein dynamics on the microsecond timescale is reached by multiple many-nanosecond MD simulations rather than a single one for the majority of current force fields. Longer trajectories are necessary to probe processes with long correlation times, but these also entail an accumulated

probability of sampling conformational states that are nonnative. They are only weakly populated in the physical system, but are hard to escape during the timescales available through conventional MD. Thus, ensemble MD methods are most physically relevant in this regard. To address the need for this type of simulations, in this project we use the multiscale approach which is both highly efficient and corresponds to ensemble MD [37, 58]. Furthermore, our implementation of the multiscale approach integrates the latest advances in the force fields.

[0078] Currently, it is difficult to determine whether the force fields are at fault and need to be corrected to destabilize non-native conformational states mentioned above, or whether much longer simulation times are required. However, a combination of improved force fields and longer time simulations may be used to resolve this issue. Thus cross section calculation and our free-energy basin exploration approach, proposed in this project will benefit further development of force fields. Currently, there are major efforts going on in the development of force fields. For example, a general CHARMM parameter force field for drug-like molecules (CGenFF) was introduced in 2009 [95]. This is well-suited for variety of chemical groups, ligands and drug-like molecules. This is a major step towards making traditional protein-simulation based packages more suitable for simulation of molecules other than protein [95]. In particular, this option avoids the need for generating new parameters or recalibrating those of small molecules or ligands. The family of CHARMM force field includes two polarizable fields (a) the fluctuating charge (FQ) model [96-97] and (b) the dispersion oscillator model [98-99]. However, as stated above, these force fields still need improvement to fit many experiments for a wide range of timescales and conditions.

[0079] The parameterization of MD force fields is thus an evolving field of research that tries to generate better parameters addressing a variety of molecular structures and fits a range of experimental values. These often consist of fitting the experimental values to the geometrical models, calculating high-level quantum mechanical calculations of subunits and matching the resulting parameters for terms such as atomic partial charges, bonds, and angles empirically, by matching the behavior of the model to appropriate experimental data or all-atom simulations. With the ongoing efforts, force fields will improve in the near future. In general, all major developments are being made available in new distributions of the force fields. Moreover,

the open source NAMD software package (many facets of which are used in our *SimNanoWorld*TM approach) is continuously being adapted to the latest CHARMM-based force field releases, keeping the present project in pace with the unfolding advances [100].

[0080] *Tests and Benchmarks*

[0081] The IMSET search algorithm may be tested against MD simulations. In addition, both IMSET and MOBCAL2 may be tested against experimental data individually. The integrated IMSET software may then be tested against experimental data. There is a wealth of information available in the literature for this purpose. We have selected the following test cases because they raise interesting scientific questions while offering a range of different system sizes on which to test the theoretical methods developed here:

[0082] *Trp RNA-Binding Attenuation Protein*

[0083] The trp RNA-binding attenuation protein (TRAP) is a 91 kDa complex of eleven subunits that are arranged in a ring structure [101]. It regulates the expression of the tryptophan biosynthetic genes of several bacilli by binding single-stranded RNA. TRAP is the first protein complex to be investigated by IM-MS methods [4]. Robinson and collaborators reported cross sections for the +19 to +22 charge states of the apo-complex; for the +20 to +23 charge states for TRAP bound to eleven tryptophans, and for the +20 and +21 charge states with the TRAP bound to eleven tryptophans and a 53 base segment of RNA. They concluded that the apo complex retains the ring structure at least for the +19 charge state. However, the higher charge states have smaller cross sections, suggesting a collapse of the quaternary structure. It is not clear why the higher charge states should lead to more compact structures. Because at the time these measurements were made it was not possible to perform accurate cross section calculations for a 91 kDa complex, Robinson and collaborators compared the experimental results to cross sections calculated for a model system where each subunit is treated as a sphere. This approach probably underestimates of the effects of multiple scattering because the spheres lack the roughness of a real protein surface.

[0084] Its small size, and the fact that it has a distinctive ring structure in the condensed phase, makes TRAP a good test system to benchmark the new computational methods developed in this project. Furthermore, the experimental results of Robinson and collaborators show a number of interesting features begging for a more detailed analysis. We will start with the crystal structure and calculate cross sections via the trajectory method and compare these with measured values. It is important to know how closely the gas phase structure corresponds to the crystal structure if ion mobility is to be a useful technique for determining the low resolution structure for supramolecular assemblies. For example, if there is a systematic expansion of contraction on going into the gas phase this needs to be accounted for. For small proteins it appears that there is a small contraction [102].

[0085] The next step will be to perform MD simulations for the complex to see if the structures from the simulations more closely match the experimental cross sections. These studies will be performed for a range of charge states in an effort to determine the origin of the quaternary structure collapse that apparently occurs for the higher charge states. If the simulations reproduce the collapse, they should provide insight into the cause. Because the TRAP complex is relatively small both MD simulations and IMSET simulations can be performed and compared. And so next we will perform IMSET simulations using the experimental cross sections as input to compare the structures identified by this method to those found by MD simulations. Similar studies will be performed with the complex with bound tryptophans and with a bound segment of RNA. In solution, addition of the RNA stabilizes the ring structure, and this stabilizing effect was evident in the cross section measurements. We will explore this in the simulations.

[0086] *GroEL*

[0087] The chaperone protein complex GroEL is a dual-ringed tetradecamer with a mass of 860 kDa. Heck and collaborators were the first to perform ion mobility measurements for GroEL [17] and as noted above they reported cross sections for this complex. They also argued that the GroEL retained its barrel structure in the gas phase because the addition of a substrate protein to GroEL did not cause the cross section to increase significantly (implying that

the protein must be sequestered inside the complex). Again our first step will be to perform accurate cross section calculations starting with the crystal structure to see how closely the gas phase structure conforms to the crystal structure. We will then perform IMSET simulations to investigate any structural changes that may occur on making the transition to the gas phase.

[0088] Robinson and collaborators [103] recently reported that the cross section measured for GroEL depends on the buffer (volatile versus involatile) from which the ions are electrosprayed. They suggested that the involatile buffer helps to retain the solution phase structure, and prevents a transition to a more compact gas phase conformation. Again, the analysis of these results is presently limited by the accuracy of the cross section calculations (the measurements are compared to PA cross sections which almost certainly underestimate the cross section). Our understanding of the final stages of macromolecular ion formation from an electrospray droplet is poor and this problem provides a good opportunity to expand our knowledge of this important question [104]. We will perform IMSET studies of the structures of the GroEL complex both with and without salt to investigate the influence of the salt on the gas phase structure. These studies should help to identify the optimum conditions for retaining the solution phase conformation when electrospraying large complexes. This is clearly important in applications using IM-MS to investigate solution phase conformations.

[0089] *Hepatitis B Virus Capsids*

[0090] Heck and coworkers recently reported IM-MS studies of Hepatitis B Virus (HBV) capsids with masses of 3 and 4 MDa for T=3 and T=4 (with 180 and 240 capsid proteins, respectively [8]). For T=3, they found two conformations for each charge state with cross sections that differ by around 4%. Furthermore, the cross sections increase significantly with charge state, suggesting that a structural transition (perhaps swelling) is driven by the charge increase. They found that the smaller conformation dissociates more readily. These results are important and fascinating. However, the authors were unable to provide an explanation for their observations because they lacked calculated structures as well as a way to make connections between these structures and measured cross sections. The developments described here will make these connections possible. We recently reported multiscale simulations of virus structural

transitions [37,53,58]. We will perform similar studies for HBV and use IMSET to find structures that match the measured cross sections. These simulations will be performed as a function of charge to investigate the role of capsid charge in driving the swelling transitions.

[0091] There is a spectrum of characterization technologies which could benefit from the approach proposed here. Progress to achieving this has been slowed due to the lack of an efficient simulation platform and modules that can integrate predictors of this data (e.g., the cross section reconstruction software as in Sect. III.A). In light of the extensive efforts in chemical labeling [45, 105], neutron-inelastic-scattering [106-109], microfluidics [110-111], and atomic force microscopy (AFM) [112-113], such software could have a major impact on research in the life sciences. Thus we plan follow-on projects based on the integration of *SimNanoWorld*TM with modules for these nanocharacterization technologies. In our earlier development [28], we presented a general theoretical approach wherein various types of nanocharacterization data can be integrated into the *SimNanoWorld*TM multiscale framework much like that of Sect. III.B for cross section data.

[0092] There are a plurality of advantages of the present disclosure arising from the various features of the apparatus and methods described herein. It will be noted that alternative embodiments of the apparatus and methods of the present disclosure may not include all of the features described yet still benefit from at least some of the advantages of such features. Those of ordinary skill in the art may readily devise their own implementations of an apparatus and method that incorporate one or more of the features of the present disclosure and fall within the spirit and scope of the present disclosure.

[0093] While the invention is susceptible to various modifications and alternative forms, specific embodiments will herein be described in detail. It should be understood, however, that there is no intent to limit the invention to the particular forms described, but on the contrary, the intention is to cover all modifications, equivalents, and alternatives falling within the spirit and scope of the invention.

[0094] References in the specification to "one embodiment", "an embodiment", "an illustrative embodiment", etc., indicate that the embodiment described may include a particular feature, structure, or characteristic, but every embodiment may not necessarily include the

particular feature, structure, or characteristic. Moreover, such phrases are not necessarily referring to the same embodiment. Further, when a particular feature, structure, or characteristic is described in connection with an embodiment, it is submitted that it is within the knowledge of one skilled in the art to affect such feature, structure, or characteristic in connection with other embodiments whether or not explicitly described.

[0095] Embodiments of the invention are further described by the following enumerated clauses:

[0096] 1. A method of acquiring mass spectral data using ion mobility mass spectrometry and interpretation of said data comprising using an ion mobility sequential elimination technique, wherein

using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

[0097] 2. The method of clause 1, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

[0098] 3. The method of clause 2, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

[0099] 4. The method of clause 1, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

[00100] 5. The method of clause 4, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

[00101] 6. The method of clause 5, wherein buffer gas atoms can undergo more than one collision with the ion.

[00102] 7. The method of clause 1, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

[00103] 8. The method of clause 1, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

[00104] 9. The method of any of any of clauses 1-8 wherein the data is generated from a supramolecular assembly.

[00105] 10. The method of clause 9 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

[00106] 11. A method of analyzing ion mobility mass spectrometry mass spectral data comprising using an ion mobility sequential elimination technique, wherein
using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

[00107] 12. The method of clause 11, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

[00108] 13. The method of clause 12, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

[00109] 14. The method of clause 11, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

[00110] 15. The method of clause 14, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

[00111] 16. The method of clause 15, wherein buffer gas atoms can undergo more than one collision with the ion.

[00112] 17. The method of clause 11, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

[00113] 18. The method of clause 11, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

[00114] 19. The method of any of any of clauses 11-18 wherein the data is generated from a supramolecular assembly.

[00115] 20. The method of clause 19 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

[00116] 21. The method of clause 10 or 20 wherein the supramolecular assembly of biological molecules is a virus or portion thereof.

[00117] 22. The method of clause 10 or 20 wherein the supramolecular assembly of biological molecules is a charged biomolecule.

[00118] 23. A method of determining the molecular conformational state of a supramolecular assembly comprising

acquiring mass spectral data from said supramolecular assembly using ion mobility mass spectrometry and analyzing the acquired data using an ion mobility sequential elimination technique, wherein

using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

[00119] 24. The method of clause 23, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

[00120] 25. The method of clause 24, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

[00121] 26. The method of clause 23, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

[00122] 27. The method of clause 26, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

[00123] 28. The method of clause 27, wherein buffer gas atoms can undergo more than one collision with the ion.

[00124] 29. The method of clause 23, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

[00125] 30. The method of clause 23, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

[00126] 31. The method of clause 23 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

[00127] 32. The method of clause 31 wherein the supramolecular assembly of biological molecules is a virus or portion thereof.

[00128] 33. The method of clause 31 wherein the supramolecular assembly of biological molecules is a charged biomolecule.

[00129] 34. A method of determining the molecular conformational state of a supramolecular assembly of any of the descriptions herein comprising

acquiring mass spectral data from said supramolecular assembly using ion mobility mass spectrometry and analyzing the acquired data using a FE basin sequential elimination technique comprising

an all-atom underlying formulation and continuous (and not discrete) configuration space; coarse-grained structural variables; a multiscale methodology to derive Langevin equations for the OPs and algorithms for computing all factors in these equations from an interatomic force field; an FE basin discovery method using modification of FE driving thermal-average forces for OP evolution that integrates prior knowledge of known FE minimizing structures to guide the evolution to yet-unknown ones; an efficient, calibration-free multiscale simulation methodology on which to build the methodical search algorithm, and which is flexible enough to incorporate experimental data of a range of resolutions;

wherein the FE basin sequential elimination technique does not require prior knowledge of the reaction path, nor the final or initial structure.

[00130] 35. A method of determining the molecular conformational state of a supramolecular assembly of any of the descriptions herein comprising acquiring mass spectral data from said supramolecular assembly using ion mobility mass spectrometry and analyzing the acquired data using an ion mobility sequential elimination technique, according to any of the techniques described herein.

[00131] The references cited herein are incorporated by reference in their entirety for disclosure that is related to the present disclosure.

[00132] The invention herein may be described further as follows:

[00133] An approach for the automated discovery of low free energy states of macromolecular systems is presented. The method does not involve delineating the entire free energy landscape but proceeds in a sequential free energy minimizing state discovery, i.e., it first discovers one low free energy state and then automatically seeks a distinct neighboring one. These states and the associated ensembles of atomistic configurations are characterized by coarse-grained variables capturing the large-scale structure of the system. A key facet of our approach is the identification of such coarse-grained variables. Evolution of these variables is governed by Langevin dynamics driven by thermal-average forces and mediated by diffusivities, both of which are constructed by an ensemble of short molecular dynamics runs. In the present approach, the thermal-average forces are modified to account for the entropy changes following from our knowledge of the free energy basins already discovered. Such forces guide the system away from the known free energy minima, over free energy barriers, and to a new one. The theory is demonstrated for lactoferrin, known to have multiple energy-minimizing structures. The approach is validated using experimental structures and traditional molecular dynamics. The method can be generalized to enable the interpretation of nanocharacterization data (e.g., ion mobility – mass spectrometry, atomic force microscopy, chemical labeling, and nanopore measurements).

[00134] Multiple macromolecular conformational states are observed in numerous experimental studies. However, it is often difficult to directly extract the 3D structures of a

biomolecule from such data. Coarse-grained information on macromolecular assemblies is obtained in ion mobility – mass spectrometry experiments, [1'] AFM, [2'] chemical labeling, [3'] and nanopore measurements. [4'] Such experimental data leave much ambiguity regarding detailed structure. These data provide only a few parameters while many configurational variables are required to capture secondary structure of a large macromolecular system. Here, an information theory based method for the sequential discovery of conformational states of a macromolecular system as free energy minimizing structures is presented.

[00135] The dynamics of macromolecular systems involves the coupling of processes across multiple time and space scales. This multiscale character of macromolecular systems presents a challenge for identifying their numerous structural states. Here, this is addressed using a deductive multiscale approach [5'] as the basis of a structure discovery method.

[00136] The structural states usually of interest for macromolecular systems are those which minimize the free energy (FE). Each such state represents an ensemble of all-atom structures. The set of such structures, to which the nearby structures evolve spontaneously, is denoted an FE basin. However, due to thermal fluctuation, there are no all-atom structures that initiate trajectories which subsequently always reside in the basin; although high energy barriers may trap these trajectories within a basin for an exceedingly long time. Here, a framework for characterizing an FE basin and associated ensemble of all-atom configurations is presented. The ensemble for a given basin is required to compute associated average quantities that mediate the evolution of its all-atom configurations.

[00137] As one is typically interested in analyzing the kinetics of the transitions between FE minimizing structures, i.e., FE basins, a dynamical theory of the quantities characterizing the ensembles of all-atom configurations is needed. Here, the set of variables used to characterize such dynamical ensembles are denoted order parameters (OPs) Φ . Implicit in the above discussion is that there is a timescale separation between the dynamics of Φ and the individual all-atom states. Thus, while the system rapidly visits many all-atom configurations within basin, the character of the ensemble, as tracked by Φ , is slowly varying. Therefore, a dynamical theory of OPs allows tracking the ensemble as the system evolves from one basin to another.

[00138] Interest in the biomolecular structure-function relationship has led to the development of theoretical structure discovery methods inspired by quickly growing sequencing data. However, progress in obtaining experimental structures has been much slower. Theoretical and computational methods for discovering the structure of macromolecular systems have recently been reviewed [6'] and include the following: combinatorial methods [7',8'] for finding global minimum energy conformations; all-atom structure reconstruction methods [9'] which start with experimental C α trace and rely on combinatorial side-chain optimization and standard molecular dynamics (MD) energy minimization; approaches which employ specific energy functions, [10'] simulated annealing, [11'] and mean-field optimization; [12'] global optimization approaches for the structure prediction and FE calculations of solvated peptides; [13'] Monte Carlo method [14'] in conjunction with simulated annealing; [15',16'] genetic optimization algorithms; [17',18'] rigid-cluster elastic network interpolation technique [19'] and normal mode analysis [20'] for generating feasible transition pathways between known macromolecular conformations; enhanced rare-event sampling techniques such as transition path sampling [21'] and metadynamics [22',23'] (including molecular dynamics flexible fitting [24'] for resolving low resolution cryo-EM structures).

[00139] These approaches provide insights into global protein structure optimization, transition paths and FE landscapes. However, there is still room for improvement of theoretical structure determination strategies in light of one or more of the following. (1) They are usually limited to small systems and biologically short times. (2) Most implementations use simplified potential models. (3) Coarse-grained models require extensive recalibration for each new application and governing dynamics equations must be postulated. Calibration is often limited to a small set of equilibrium states, even when used to describe non-equilibrium processes. Extensive data is used for calibration including system-specific information such as the area and volume accessible to the solvent, the sparse sets of NMR data, or subsystem structural information integrated via bioinformatics methods. (4) No criterion is provided for determining the completeness of the set of coarse-grained variables used to characterize the FE landscape. (5) All-atom interactions and states are not used or obtained. (6) Only unsolvated structures are addressed. (7) Only a single FE minimizing structure is provided. (8) Potential energy

minimizing structures, and not FE minimizing ones, are provided. (9) Long computational times are needed to simulate transitions between energy basins. (10) Guide a system through a path that is not necessarily natural for the molecular physics. (11) The history of configurations generated in a Monte Carlo sequence is not always accounted for when mapping the energy landscape. (12) No guidelines are provided for optimization, although performance may be very sensitive to the specific implementation, e.g., as in genetic optimization algorithm. (13) The simulation may be trapped in local, but not global, minima. (14) Knowledge of initial and/or final states may be required.

[00140] The objective of the present invention is to overcome most of these difficulties using the following: an all-atom underlying formulation and continuous (and not discrete) configuration space; coarse-grained structural variables; a multiscale methodology to derive Langevin equations for the OPs and algorithms for computing all factors in these equations from an interatomic force field; an FE basin discovery method using modification of FE driving thermal-average forces for OP evolution that integrates prior knowledge of known FE minimizing structures to guide the evolution to yet-unknown ones; an efficient, calibration-free multiscale simulation methodology on which to build the methodical search algorithm, and which is flexible enough to incorporate experimental data of a range of resolutions; the FE basin sequential elimination technique introduced here does not require prior knowledge of the reaction path, nor the final or initial structure.

[00141] A FE basin is defined as an ensemble of all-atom configurations consistent with a set of OPs that minimize the FE, i.e., for which thermal-average forces vanish. As the multiscale simulations progress via Langevin timesteps, it is often necessary to modify the definition of the OPs, [25'] so that other variables (denoted descriptors here) are also used to characterize an FE basin. These descriptors are defined directly in terms of all-atom configurations and are typically directly measurable characteristics (e.g., moments of inertia, or electrical dipole and quadrupole moments). Thermal-average OP forces modified by using the descriptors characterizing known structures are introduced to guide a multiscale simulation away from the known basins to a new one.

-37-

[00142] The multiscale formalism for simulating macromolecular assemblies is reviewed below. The thermal-average forces arising in multiscale analysis are modified such that they drive macromolecular systems to new FE basins, enabling a sequential basin discovery algorithm; details on implementation are provided below. Validation is presented below and conclusions are drawn below.

Brief Review of the Deductive Multiscale Approach

[00143] A natural framework for casting an FE basin discovery algorithm is deductive multiscale analysis. [5',26',27'] Let $\underline{r} \equiv \{\bar{r}_1, \bar{r}_2, \dots, \bar{r}_N\}$ denote the all-atom configuration of the structure of interest. In the approach adopted here, OPs $\underline{\Phi} \equiv \{\bar{\Phi}_1, \bar{\Phi}_2, \dots, \bar{\Phi}_{N_{OP}}\}$ describe overall structure of the system. As earlier, [26',28',29'] the starting point of the analysis is the $\underline{\Phi}$ - $\underline{r}$ relationship: [29']

$$\bar{r}_i = \sum_{k=1}^{N_{OP}} U_{ki} \bar{\Phi}_k + \bar{\sigma}_i. \quad (1')$$

The residuals $\bar{\sigma}_i$ are introduced to address the truncation of the k -sum in (1') resulting from taking a relatively small number of OPs, $N_{OP} \ll N$. With this, the k -sum generates the continuous deformation of the N -atom assembly via changes in $\underline{\Phi}$, while the $\bar{\sigma}_i$ account for more random individual atomic motions. [30'] A reference structure $\underline{r}^\circ$ is used to construct the basis functions U_{ki} as earlier. [25'] Using mass-weighted orthogonalized [31'] U_{ki} , [26'] one obtains

$$\bar{\Phi}_k = \frac{1}{\mu_k} \sum_{i=1}^N m_i U_{ki} \bar{r}_i, \mu_k = \sum_{i=1}^N m_i U_{ki}^2, \quad (2')$$

m_i being the mass of atom i . This formulation has the only difference from the one in ref 26 wherein the μ_k were directly embedded in U_{ki} . This more explicit formulation suggests that the $\bar{\Phi}_k$ are generalized CM variables and the μ_k are associated masses. For example, if U_{ki} is independent of i , then the related OP is the center of mass. Other $\bar{\Phi}_k$ characterize finer details of

-38-

the distribution of mass.[32'] Eq. (2') does not provide the reciprocal relation, i.e., does not imply $\underline{r}$ for given $\underline{\Phi}$, since the $\bar{\sigma}_i$ are yet-unspecified. This is expected since a given coarse-grained description ($\underline{\Phi}$ here) corresponds to an ensemble of all-atom configurations, as addressed in more detail below.

[00144] The deductive multiscale approach [25'] starts with the N -atom probability density ρ that depends on the $6N$ atomic coordinates and momenta (denoted Γ). While the N atoms constitute the structure of interest, atoms in the remainder of the system are labeled with $i > N$. The starting point of the multiscale analysis is the ansatz that ρ depends on Γ both directly and, via the OPs, indirectly. When the U_{ki} change slowly as the reference structure $\underline{r}^\circ$ varies, use of the OPs as defined in Eq. (2') introduces a smallness parameter ε in the Liouville equation obtained through the ansatz:

$$\rho = \rho(\Gamma, \underline{\Phi}; t_0, \underline{t}; \varepsilon), \underline{t} \equiv \{t_1, t_2, \dots\}, t_n = \varepsilon^n t. \quad (3')$$

[00145] Since ε is related to the ratio of the mass of a typical atom to that of a subset of the atoms, it is small and thereby enables a perturbation analysis. [26'] The result is a Langevin equation for $\underline{\Phi}$ and the coevolving OP-constrained, quasi-equilibrium probability density $\hat{\rho}$ of all-atom structures: [25',26',30']

$$\hat{\rho} = \exp(-\beta H) / Q, Q(\underline{\Phi}) = \int d\Gamma^* \Delta^+(\underline{\Phi} - \underline{\Phi}^*) \exp(-\beta H^*), \beta = 1/k_B T, \quad (4')$$

where * denotes evaluation at Γ^* over which integration is taken, and H is the Hamiltonian. The Δ^+ factor is the product of N_{OP} narrow Gaussian-like functions introduced to impose an OP-constraint on the ensemble of all-atom configurations.

[00146] With the above, the ensemble of all-atom configurations characterized by $\hat{\rho}$ evolves with $\underline{\Phi}$. In turn, $\underline{\Phi}$ evolves via the following Langevin dynamics:

$$\frac{\partial \Phi_{k\alpha}}{\partial \tau} = \beta \sum_{k'\alpha'} D_{k\alpha k'\alpha'} f_{k'\alpha'} + \xi_{k\alpha}, \tau = \varepsilon^2 t. \quad (5')$$

$\Phi_{k\alpha}$ is the α -th Cartesian component of $\bar{\Phi}_k$, and $\vec{f}_k$ is the thermal-average force given by the phase space average of the corresponding OP force [25',26']

-39-

$$\bar{f}_k(\underline{\Phi}) = \int d\Gamma^* \Delta^+(\underline{\Phi} - \underline{\Phi}^*) \exp(-\beta H^*) \sum_{i=1}^N U_{ki} \bar{F}_i^*, \quad (6')$$

where $\bar{F}_i$ is the net force on atom i . The diffusivity factors $D_{k\alpha k'\alpha'}$ in Eq. (5') are related to correlation functions of OP time derivatives. [5',26'] A random noise term $\bar{\xi}_k$ determines the stochastic part of Langevin evolution and is constructed by requiring the integral of its autocorrelation function to be proportional to the diffusion coefficient $D_{k\alpha k\alpha}$. [26']

[00147] The above multiscale methodology was implemented as the DeductiveMultiscaleSimulator (DMS) software,[5',26'] ' originally as the MD/OPX software, [29',33'] and recently redesigned, optimized, and seamlessly integrated with NAMD via a new Python interface. In the present implementation, the thermal-average forces $\bar{f}_k$ were calculated using Monte Carlo integration. The ensemble of all-atom configurations $\underline{r}$ needed was generated in two steps. First, Eq. (1') was used with statistically chosen $\underline{\sigma}$ to generate a preliminary ensemble of all-atom configurations consistent with instantaneous values of the OPs as they evolve according to the Langevin dynamics (5'). This initial ensemble was enriched via short isothermal MD runs over which $\underline{\Phi}$ does not change appreciably. The method takes advantage of the special properties of the OPs introduced as in Eq. (1'), [25',26',30'] originally cast as a space-warping framework. [28']

Formulation and Implementation of the Sequential Basin Discovery Method

[00148] **A. Discovery Concepts.** The foundation of the sequential elimination method for FE basin discovery is as follows. Free energy is thermal energy minus temperature times entropy. Entropy depends on the available information on the constraints to which the system is subjected (e.g., temperature or specific values of $\underline{\Phi}$). In the sequential elimination approach, this information includes the fact that some FE basins are known and one seeks to discover new ones.

[00149] To implement this basin discovery method, a set of N_d descriptors

$\underline{\eta} \equiv \{\eta_1, \dots, \eta_{N_d}\}$ is used to characterize a basin. While these descriptors characterize overall

system structure as do the OPs, they are not used directly in the multiscale formulation since they may not serve as the basis of the $\underline{r} - \underline{\Phi}$ relation (1').

[00150] In what follows, we develop formulae for these descriptors and show how they can be used to automatically guide the multiscale dynamics (above) to a new basin given the descriptors for the known ones. For simplicity, we present the method for the case when one basin is known and a second one is sought. Generalization for multiple known basins is straightforward (see below).

[00151] The search algorithm we provide combines elements of (1) the multiscale analysis of macromolecular systems; [25',26',30',34'-47'] (2) the notion of a stepwise procedure that precludes evolution into basins of attraction identified in earlier steps in the computation; (3) an OP method for simplifying the FE landscape to eliminate thermally-irrelevant basins of attraction. In addition, (4) an algorithm for accounting for experimentally determined structural information can be incorporated in the search algorithm. [30']

[00152] **B. Implementation of the Basin Discovery Algorithm.** The methodology of Sect III.A for FE basin discovery was implemented by modifying the DMS software; [5',25',26'] this implementation is denoted here DeductiveMultiscaleSimulator.BasinDiscovery (DMS.BD). DMS uses NAMD with the CHARMM force field to perform selected calculations to construct forces and diffusions in the Langevin equations (above). Details on these MD calculations are provided in Supporting Information. DMS was modified by changing the expressions for thermal-average forces (Sect. D, below), and similarly for the diffusion factors via averaging over restricted phase space. The workflow of DMS.BD is shown in Figure 11.

[00153] The entire FE landscape is not calculated since the discovery of adjacent basins does not require it. [48'] Instead, the natural thermal-average forces $\bar{f}_k$ (6') are used to locate basins in the space of OPs or descriptors. The bottom of a basin is defined to be the point in OP space where all natural thermal-average forces vanish.

[00154] DMS enables evolution of OPs along with an ensemble of all-atom configurations constrained by the instantaneous OP values. Such ensembles are constructed using a set of MD runs that capture a timescale much shorter than those of the OPs and are

initialized to a given value of the OPs using higher-order OP-like variables as earlier [26'] and in the above review. Thus, OPs remain essentially constant when the all-atom configurations are sampled using such MD runs. As a result, the state-counting factor Δ^+ appearing in the partition function (11') is accounted for in thermal force and diffusivity calculations. At each Langevin timestep OPs constrain the quasi-equilibrium ensemble of atomic states which, in turn, enables the computation of the thermal-average forces (23') and diffusivities [5',26'] that mediate Langevin OP dynamics (5'). With this, the modified thermal-average forces guide the system to new FE basins as below in section D.

[00155] In DMS.BD, the coevolving quasi-equilibrium ensemble is modified using method of Sect. D. Information on known basins is accounted for in the state-counting factor Δ^+ in the form of the product, with one factor (10') for each basin. In the current implementation of DMS.BD, each of these factors involves several descriptors, as follows. The set $\{a_1, \dots, a_{N_d}\}$ of accompanying exponential width factors (Sect. D) was taken to be identical for all basins.

[00156] **C. Descriptors.** Examples of descriptors that can be used for basin discovery include total mass, charge, length of the dipole moment, and eigenvalues of the moment of inertia or electrical quadrupole moment tensors. Such descriptors have the important property that they are independent of system orientation. In the current implementation, the three eigenvalues of the moment of inertia tensor of the structure relative to its center of mass are chosen. To discriminate between more complex structures and associated FE basins, more descriptors can be used.

[00157] The moment of inertia tensor $\bar{\bar{M}}$ is defined via

$$\bar{\bar{M}} = \sum_{i=1}^N m_i \begin{pmatrix} y_i^2 + z_i^2 & -x_i y_i & -x_i z_i \\ -x_i y_i & x_i^2 + z_i^2 & -y_i z_i \\ -x_i z_i & -y_i z_i & x_i^2 + y_i^2 \end{pmatrix}. \quad (7')$$

Being the eigenvalues of $\bar{\bar{M}}$, molecular descriptors satisfy the cubic equation

$$\begin{aligned} \eta^3 - (M_{xx} + M_{yy} + M_{zz})\eta^2 + (M_{xx}(M_{yy} + M_{zz}) + M_{yy}M_{zz} - M_{xy}^2 - M_{yz}^2 - M_{xz}^2)\eta + \\ M_{yy}(M_{xz}^2 - M_{xx}M_{zz}) + M_{zz}M_{xy}^2 + M_{yz}(M_{xx}M_{yz} - 2M_{xy}M_{xz}) = 0, \end{aligned} \quad (8')$$

whose coefficients are determined by the elements of matrix (7').

[00158] To proceed in a sequential elimination calculation, all-atom configurations which yield descriptors close to those for the known basin are eliminated from the ensemble as follows. First one must specify the descriptors that characterize the known basin. However, a basin includes an ensemble of all-atom configurations. For isothermal systems, this ensemble is generated as earlier [25',26'] using short isothermal MD runs initialized with configurations consistent with instantaneous values of $\underline{\Phi}$ (Sect. II). Out of this ensemble, a most probable structure with the lowest potential energy is chosen to calculate descriptor values characterizing the known basin.

[00159] D. Modification of Thermal-average Forces to Include Known Basin

Information. The starting point for the sequential basin discovery is entropy maximization to determine the quasi-equilibrium probability density $\hat{\rho}$ constrained by the known information. These constraints include the isothermal condition and fixed system volume, as well as the instantaneous values of the OPs at a given stage of the Langevin dynamics. In addition, states that resemble those in the known basin are excluded from the counting of states in the entropy for a sequential elimination computation. With this, the entropy $\hat{S}$ takes the form:

$$\hat{S} = -k_B \int d\Gamma^* \Delta^+ (\underline{\Phi} - \underline{\Phi}^*) \Delta^- (\underline{\theta}_{\text{known}} - \underline{\eta}^*) \hat{\rho}^* \ln \hat{\rho}^*, \quad (9')$$

where an additional factor Δ^- is introduced to discount the known FE basin via the descriptors at its bottom, $\underline{\theta}_{\text{known}} = \{\theta_1, \dots, \theta_{N_d}\}$. Since the descriptors are coarse-grained variables, they can be expressed in terms of a complete set of OPs (Appendix A). Therefore, they are used here to introduce information on the known basin to enable the discovery.

[00160] The factor Δ^- has the character of $1 - \Delta^+$, and therefore excludes configurations in the known basin, i.e., configurations which have descriptors close to $\underline{\theta}_{\text{known}}$. In other words, to give preference to the states that are different from those in the known FE basin, this counting factor is set to one for configurations distinct from the known one and is zero within the known basin. The particular form of Δ^- was chosen as

-43-

$$\Delta^-(\underline{\eta} - \underline{\theta}_{\text{known}}) = 1 - \exp\left(-\sum_{d=1}^{N_d} a_d |\eta_d(\underline{r}) - \theta_{d,\text{known}}|\right). \quad (10')$$

The parameter a_d is proportional to the inverse width of the Gaussian-like exponential function associated with descriptor d in Eq. (10'). The a_d values are chosen to ensure escape from the known basin (see below).

[00161] Using Eq. (9') and entropy maximization, one arrives at the OP-constrained all-atom probability density $\hat{\rho}$ (4'). The associated partition function Q takes the form

$$Q(\underline{\Phi}; \underline{\theta}_{\text{known}}) = \int d\Gamma^* \Delta^+(\underline{\Phi} - \underline{\Phi}^*) \Delta^-(\underline{\eta}^* - \underline{\theta}_{\text{known}}) \exp(-\beta H^*). \quad (11')$$

This yields the Helmholtz free energy F :

$$F = -\frac{1}{\beta} \ln Q(\underline{\Phi}; \underline{\theta}_{\text{known}}). \quad (12')$$

[00162] By analogy with the developments of the above review, the modified thermal-average forces $\bar{f}_k^m$ are obtained (Appendix B). Then the Langevin equation (5') can be used to evolve the system from the known basin to a new one. This calculation is carried out in the present implementation using methods as earlier [25',26',29'] but with the present modified OP forces described in detail in Appendix B. At the end of each Langevin timestep, the updated OPs are obtained along with the associated ensemble of all-atom configurations. In turn, the latter are used to generate the next Langevin timestep. This Langevin timestepping is stopped when the natural (not modified – see above review) thermal-average forces are negligible, indicating arrival at the bottom of the new FE basin.

[00163] **E. Guided Evolution From the Known to Unknown Basins.** The calculation from a known basin to a new one proceeds as follows. One starts the calculation within the known basin and then evolves the system via Langevin equation (5') with modified thermal-average forces of Eq. (23'). A Langevin evolution course is tracked by the values of potential energy and $\bar{f}_k$ (6'). After a high FE barrier is overcome and the system descends towards the bottom of a new FE basin, the basin discovery simulation is carried on until the thermal-average

forces become negligible. A new basin structure is chosen from the timestep at which $\tilde{f}_k$ are negligible, signifying that minimum free energy was achieved within the new basin.

[00164] Let the biasing thermal-average force $\tilde{f}_k^b$ (**20'**) be the $\bar{\Phi}_k$ -gradient of the FE associated with the modified partition function (**11'**). Specifically, the $\tilde{f}_k^b$ are computed using the derivatives of the state-counting factor (**10'**) with respect to the descriptors,

$$\frac{\partial \Delta^-}{\partial \eta_d} = \Delta^- \sum_{b=1}^{N_{\text{known}}} \frac{\text{sgn}(\eta_d(\underline{r}) - \theta_{d,b}) a_d}{\exp\left(\sum_{d=1}^{N_d} a_d |\eta_d(\underline{r}) - \theta_{d,b}|\right) - 1}. \quad (13')$$

[00165] It is also necessary to make an estimate of the N_d inverse width parameters a_d . In the present implementation, this is accomplished via adjustment to ensure escape from the known basin (see second subsection of the Supporting Information below).

[00166] F. Generalization for Sequential Discovery of Multiple Free Energy Basins.

The case of a single known basin and the discovery of a new one was considered above. This algorithm can readily be generalized to the case of sequential discovery in a step-wise procedure. At each step, the system is guided away from the basins discovered in earlier steps to a new one.

In a given step, the Δ^- factor (**10'**) can be generalized to be a product of similar factors

$\Delta^-(\underline{\eta} - \underline{\theta}_b)$, one for each of the known basins labeled $b = 1, \dots, N_{\text{known}}$. For basin b , the N_d descriptors $\theta_{d,b}$ are accounted for, and the set $\{a_1, \dots, a_{N_d}\}$ of factors in the exponential function (**10'**) are chosen to ensure escape from each of the known basins.

[00167] The Δ^- factors artificially lower the FE of a system as it evolves out of the known basins. When Δ^- are incorporated in a sequential calculation, the system is driven away from all basins discovered in earlier phases of the calculation by the modified thermal-average forces $\tilde{f}_k^m$ (**21'**). Within a given phase of such a calculation, a number of Langevin steps are to be carried out to arrive at the set of OP values at the bottom of a newly discovered basin.

Validation for Human Lactoferrin

[00168] The FE basin discovery method was validated by finding two new FE basins on the FE landscape for human lactoferrin. [49'] A brief summary of observations on this system is as follows. Two crystal X-ray structures are available for this protein: diferric lactoferrin (PDB code 1LFG), and apolactoferrin (PDB code 1LFH). [49']

[00169] To validate the general DMS.BD algorithm of Sect. III.F, an arbitrary structure, and notably the compact closed-lobe diferric conformation 1LFG [19'] with the iron and carbonate ions removed (Figure 7 *a*), was used to start a DMS simulation to find the bottom of a first FE basin. Then DMS.BD was used to simulate traversal of the FE topography and discover a new basin for lactoferrin starting from the diferric basin. A set of descriptors characterizing the closed-lobe lactoferrin structures from the bottom of this first basin was used to guide simulation away from it. Lactoferrin was guided to a second basin with slightly opened structures. Next, a set of descriptors characterizing the second “pseudodiferric” basin was incorporated to guide protein to a third basin. The third basin contains open-lobe structures (Figure 7 *h*) which are similar to the apolactoferrin conformation 1LFH, but are less open than the X-ray structure 1LFH (Figure 7 *b*). Additional details on DMS and DMS.BD simulations are provided in Supporting Information below.

[00170] Implementation of DMS.BD is based on the interplay of Δ^- -modified ($\tilde{f}_k^-$) and biasing ($\tilde{f}_k^b$) components of the modified FE driving forces $\tilde{f}_k^m$ (23'). Inclusion of the Δ^- factor in state counting reduces those FE minimizing forces $\tilde{f}_k^-$ (22') that would have otherwise kept the system within the known basin (i.e., if $\tilde{f}_k^-$ and not $\tilde{f}_k^m$ was used). The biasing forces $\tilde{f}_k^b$, by design, oppose the $\tilde{f}_k^-$ and, by the choice of inverse width parameters a_d in Eq. (10') drive the system away from known basin(s). Once out of a known basin, $\tilde{f}_k^b$ becomes smaller than $\tilde{f}_k^-$ if the structure changes appreciably relative to those characterizing the known basins. Thus, after a barrier is crossed, the $\tilde{f}_k^-$ drive the system towards FE minimizing structure for which descriptors differ from those in the known basin(s). With this, the $\tilde{f}_k^m$ drive the system away from known basins and to new ones.

[00171] To rationalize the above effects, we compute the thermal-average forces $\bar{f}_k$ (6') at each Langevin step of a DMS.BD trajectory. For structures near the bottom of a basin all $\bar{f}_k$ are close to zero. However, stochastic forces ξ (5') force the system to fluctuate about the bottom. If the system is far from the bottom, the $\bar{f}_k$ are appreciable and drive the system to the bottom. With this, the $\bar{f}_k$ provide information on the FE landscape topography along a Langevin evolution path. They indicate the location of FE barriers along the path (i.e., places where the $\bar{f}_k$ vanish). For extensive sampling, integration over these forces yields an estimate of FE barrier height when the $\bar{f}_k$ are integrated (Appendix C).

[00172] For lactoferrin oriented as in Figure 7, lobes open in the xz-plane accompanying the transition from the diferric to the apolactoferrin basin. In particular, OPs Φ_{100X} and Φ_{001Z} track extension-compression along the x and z directions, capturing the structural transition. The closed and open states are also characterized by values of the descriptors (i.e., moment of inertia eigenvalues, Figure 8). Thermal-average forces along the DMS.BD trajectory from the diferric to the pseudodiferric basin are shown in Figure 9. Most forces fluctuate around zero along this trajectory, suggesting that local topography along the guided trajectory has the character of a valley. However, forces f_{100X} and f_{001Z} along the trajectory suggest that a barrier is crossed, i.e., they change from negative to positive as the barrier is traversed, stop growing, and ultimately go to zero as the system approaches the bottom of a new basin (Figure 9). This illustrates that our method explores local topography in the vicinity of high probability pathways, i.e., the FE is minimum along the directions orthogonal to the path.

[00173] In Figure 10, we plot potential energies of the most probable atomic configurations from constant OP ensembles at every Langevin step during transitions between specified basins. The potential energy profile also suggests a barrier crossing. Presence of such barriers suggests that the FE and potential energy landscapes are related, but not identical (Figure 9 versus Figure 10). This is expected because entropy effects at finite temperatures are not reflected in the potential energy profile and, therefore, can shift the location of potential energy

features (minima or transition points) relative to the FE ones. The above transition path and topography are not readily accessible via traditional MD, as follows.

[00174] To confirm that different FE basins were discovered, an ensemble of all-atom configurations in the vicinity of the bottom of each basin was explored using traditional isothermal MD. For a given basin, the MD was initialized with an all-atom state of minimum potential energy and which was consistent with the OPs at the bottom of the basin. Then, 10 ns NAMD runs were performed in order to show that all-atom trajectory starts and ends in the same basin (Figures 11, 9 c). This MD sampling validates our method, i.e., trajectories remain for long times in a given FE basin. That a trajectory remains in a basin is indicated by the fact that the OPs do not change appreciably over their timecourse. In these samplings one does not obtain structures whose set of descriptors (and, therefore, OPs) fall in the domain sampled by MD in any other basin (Figure 14). In addition, the DMS.BD is robust to the choice of initial all-atom structure. The analysis described in Appendix D suggests that DMS.BD can guide a system away from a known basin through the arbitrary choice of initial structure, which does not necessarily characterize the bottom of the basin. In this context, we probe the basin 2 to 3 transition using an arbitrary initial structure denoted “descent 2”.

[00175] We compare our results with those from experimental and previous theoretical observations. Transition of lactoferrin from the diferric to apolactoferrin states is accompanied by changes in the vicinity of residues THR90 and VAL250. These residues act as hinges that facilitate the lobe-opening transition. [50'] We observe substantial differences in the backbone dihedral angles of these residues between the closed state and the discovered slightly open one (Figure 13). In particular, more differences in dihedral angles are observed for residues in the loop region than in the highly structured parts of lactoferrin. This validates that most of the secondary structure is preserved during the transition, as has been suggested by previous theoretical results. [19'] A residue-by-residue RMSD comparison of the backbone C_α atoms between structures from basin 1, 2 and 3 with respect to that of the X-ray structure 1LFG of closed lactoferrin (Figure 12) is performed. To understand differences with the apolactoferrin structure 1LFH, we also plot the RMSD between diferric and apolactoferrin structures. The RMSD gradually increases from basin 1 to 3 indicating lobe opening. The deviations are

significant in the vicinity of residues 90 and 250, indicating that the hinge motion is captured through DMS.BD simulations. Thus, the DMS.BD predicted FE minimizing structures approach the experimentally observed open state 1LFH.

[00176] A methodology for the sequential discovery of FE basins for macromolecular systems is presented. Structural information from known basins is used to escape/avoid them, and thereby enable the discovery of yet-unknown basins. The approach was implemented via our DMS software and validated using two X-ray structures for human lactoferrin. Two new FE basins were discovered. The method has the potential for discovering pathways of transitions between basins, including estimates of FE barriers along the transition paths. Comparison of nanocharacterization data with values calculated for the discovered all-atom states provides an approach for interpretation of such data. One example of nanocharacterization data to which this approach can be applied is collision cross-sections from ion mobility – mass spectroscopy experiments for charged biomolecules.

[00177] The basin discovery algorithm is built on multiscale techniques. The latter provide orders of magnitude increase in the efficiency of simulation for large macromolecular assemblies. [5',26'] These efficiencies make the methodology and the implementation of interest in biophysical studies such as on structural transitions in viruses. [33',51']

[00178] For high temperature, the distribution of likely states within an FE basin is very broad and, therefore, the basin becomes less well defined. In particular, FE barriers that would otherwise sequester all-atom trajectories to lie within the basin are lower, enabling more frequent escape. It was shown here that descriptors chosen at the state of minimum FE in the basin can be used to guide multiscale simulations from known to yet-unknown ones (validation above).

[00179] The present method achieves system evolution and FE landscape exploration via trifold approach. OPs provide the coarse-grained description via an expression that facilitates the construction of the ensemble of all-atom states consistent with the instantaneous OP values. However, as the system departs significantly from an initial reference all-atom structure, the OPs may not provide a viable description. Thus, in our implementation a new all-atom reference configuration and resulting newly defined OPs are established when needed. This implies that

the present OPs do not serve as an appropriate coarse-grained description for mapping the broader FE landscape. In contrast, the system descriptors can serve as the coarse-grained state variables with which to define the landscape since their definition does not involve a reference configuration. However, the descriptors do not provide a convenient way to generate the ensembles of all-atom states needed to construct the thermal forces and diffusion factors mediating the evolution of the coarse-grained state. Thus, the present OPs facilitate ensemble generation and coarse-grained evolution; the descriptors provide a coarse-grained variable for a continuous mapping of the FE landscape despite the changing definition of the OPs. Thus, our method integrates the OPs, the descriptors, and ensembles of all-atom states to enable multiscale simulations across an FE landscape. This is the logic behind our trifold simulation and basin discovery approach.

Appendix A: Relationship Between Descriptors and OPs

[00180] Here we derive relationship between the descriptors $\underline{\eta}$ (8') and the OPs $\underline{\Phi}$. By substituting the $\underline{\Phi}$ - $\underline{r}$ relationship (1') for the atomic coordinates in the moment of inertia tensor (7'), one obtains

$$M_{\alpha\alpha'} = \sum_{i=1}^N m_i \left(\delta_{\alpha\alpha'} \sum_{\alpha''=\{x,y,z\}} r_{i\alpha''}^2 - r_{i\alpha} r_{i\alpha'} \right), r_{i\alpha} = \sum_{k=1}^{N_{OP}} U_{kt} \Phi_{k\alpha} + \sigma_{i\alpha}. \quad (14')$$

The cross-terms involving $\sigma_{i\alpha}$ and $\Phi_{k\alpha}$ are small because the residuals $\underline{\sigma}$ are random variables and the OPs $\underline{\Phi}$ are defined such as to minimize the mass-weighted square residuals. In addition, at the FE bottom the OPs do not change a lot and the re-referencing is applied.

[00181] In view of the quadratic relation (14') for the matrix elements $M_{\alpha\alpha'}(\underline{\Phi})$ and the cubic equation for $\underline{\eta}(\bar{\bar{M}})$ (8') with the polynomial coefficients, the relationship between the descriptors $\underline{\eta}$ and the OPs $\underline{\Phi}$ is binomial. This justifies the use of the descriptors $\underline{\theta}_{\text{known}}$, evaluated using the all-atom states (1') corresponding to the FE minimum values of the OPs, to characterize the basin in Eq. (9') for the DMS-guided FE basin discovery.

Appendix B: Derivation of the Modified Thermal-average Forces

[00182] Here we derive the expression for the thermal-average forces, modified by the state-counting factors Δ^- (10') accounting for the earlier discovered FE basins. Provided the set of known FE basins with associated low-energy states, characterized by $N_b N_d$ descriptor values $\theta_{d,b}$, the Δ^- is calculated for each of the atomic configurations $\underline{r}$ generated by MD sampling at a given Langevin timestep. The associated contribution to the OP forces is composed of the derivatives of Δ^- with respect to the OPs $\bar{\Phi}_k$. The latter can be computed using the derivatives with respect to atomic coordinates, the chain rule, and the $\underline{r} - \underline{\Phi}$ relationship (1'). When deriving new thermal-average forces $\bar{f}_k^m$, one brings the $\partial / \partial \bar{\Phi}_k$ derivative into the integral (11') and uses the property of the Hamiltonian H that it does not depend on OPs explicitly,

$$\beta Q \bar{f}_k^m = \int d\Gamma \exp(-\beta H) \sum_{i=1}^N \frac{\partial}{\partial \bar{r}_i} \{ \Delta^+ \Delta^- \} \frac{\partial \bar{r}_i}{\partial \bar{\Phi}_k}. \quad (15')$$

Assume that $\underline{r}$ can be obtained from an augmented set of OPs (i.e., those including the residual parameters as in Eq. (1')). Then the following approximation holds [29']

$$\frac{\partial \bar{r}_i}{\partial \bar{\Phi}_k} = \sum_{\alpha=\{x,y,z\}} \frac{\partial r_{i\alpha}}{\partial \Phi_{k\alpha}} = \sum_{\alpha} U_{k\alpha i} (r_{i\alpha}^o). \quad (16')$$

Using Eq. (16'), one obtains

$$\beta Q f_{k\alpha}^m = \int d\Gamma \exp(-\beta H) \sum_{i=1}^N U_{k\alpha i} \frac{\partial}{\partial r_{i\alpha}} \{ \Delta^+ \Delta^- \}. \quad (17')$$

With this, the thermal-average force is that obtained earlier (6'), with the extra Δ^- weighting factor (10'), plus a new term $\bar{f}^b$ arising from the following integral

$$\bar{f}^b = \sum_{k=1}^{N_{OP}} \bar{f}_k^b, f_{k\alpha}^b = \frac{1}{\beta Q} \int d\Gamma \Delta^+ \exp(-\beta H) \sum_{i=1}^N U_{k\alpha i} \sum_{d=1}^{N_d} \frac{\partial \Delta^-}{\partial \eta_d} \frac{\partial \eta_d}{\partial r_{i\alpha}}. \quad (18')$$

In view of Eq. (18'), the derivatives of Δ^- with respect to the descriptors $\underline{\eta}$ should achieve their maximum values in the already discovered stable states in order to maximize biasing force $\bar{f}^b$.

-51-

[00183] The derivatives of the descriptors with respect to the atomic coordinates $\partial\eta_d / \partial r_{i\alpha}$ in the new thermal-average force term $\bar{f}_k^b$ in Eq. (18') are to be taken numerically for each of the OP-restricted configurations within same Langevin timestep. These derivatives were calculated using $3N$ independent offsets in the x, y and z coordinates of each atom with subsequent recalculations of η_d . The speed-up in calculating these derivatives was achieved by using the analytical roots of polynomial (8'), as opposed to using the eigenvalue calculation subroutines.

[00184] The thermal-average forces $\bar{f}$ in our earlier approach (Appendix C in Ref. 26') are obtained from the $\frac{\partial \Delta^+}{\partial \bar{r}_i}$ term in Eq. (17'),

$$\exp(-\beta H) \frac{\partial \Delta^+}{\partial r_{i\alpha}} = \frac{\partial}{\partial r_{i\alpha}} \{ \Delta^+ \exp(-\beta H) \} - \Delta^+ \frac{\partial \exp(-\beta H)}{\partial r_{i\alpha}}. \quad (19')$$

Using the property of Δ^+ that it does not depend on spatial coordinates explicitly, but rather via OPs, and employing the divergence theorem, we present the first term in Eq. (10') in a form of full gradient and note that its contribution to the integral (17') is zero. The space derivative of H in the second term of (19') is a negated α -component of the corresponding atomic force, $\bar{F}_i$.

Thus, one obtains

$$\bar{f}^- = \sum_{k=1}^{N_{OP}} \bar{f}_k^-, \quad f_{k\alpha}^- = -\frac{1}{Q} \int d\Gamma \exp(-\beta H) \Delta^+ \Delta^- \sum_{i=1}^N U_{k\alpha i} F_{i\alpha}. \quad (20')$$

Here, the $\bar{f}_k^-$ are thermal-average forces modified by Δ^- factor in the phase space integral (20').

[00185] It was verified that by neglecting the biasing force (18') and using only the term $\bar{f}^-$ (20'), i.e., by simply multiplying the integrand in the expression for thermal-average forces (6') by the antigaussian-like probability function Δ^- of descriptors, one does not provide desired driving force for the system to evolve out of the discovered FE basins. This is because the noise term dominates over the OP forces (Eq. (5')). Attempts to increase the Langevin timestep Δt and narrowing the widths a_d^{-1} of the antigaussian Δ^- did not lead to the increase in OP forces.

[00186] The overall thermal-average force consists of two components: the FE driving forces f_{k_α} modified by Δ^- factor, and the biasing information theory-guiding ones ($f_{k_\alpha}^b$),

$$f_{k_\alpha}^m = f_{k_\alpha}^- + f_{k_\alpha}^b = \frac{1}{Q} \sum_{i=1}^N U_{k_\alpha^i} \int d\Gamma \exp(-\beta H) \Delta^+ \left(-\Delta^- F_{i\alpha} + \frac{1}{\beta} \sum_{d=1}^{N_d} \frac{\partial \Delta^-}{\partial \eta_d} \frac{\partial \eta_d}{\partial r_{i\alpha}} \right). \quad (21')$$

Discussed above in validation, mutually opposing nature of these forces underlies the discovery of basins and associated free energy minimizing structures via DMS.BD.

Appendix C: Computing Free Energy Profiles Along Langevin Paths

[00187] The FE can be calculated numerically using the values of thermal-average forces (6') along a Langevin evolution path. Let the FE and the values of OPs at the bottom of a basin be F_{bot} and $\underline{\Phi}_{\text{bot}}$, respectively. Then the FE along a Langevin trajectory can be computed via the path integral

$$F(\underline{\Phi}) = F_{\text{bot}} - \sum_{k=1}^{N_{\text{OP}}} \int_{\underline{\Phi}_{\text{bot}}}^{\underline{\Phi}} d\bar{\Phi}_k \cdot \bar{f}_k. \quad (22')$$

This can be written in discretized form using the values of $\underline{\Phi}$ and $\bar{f}_k$ at each Langevin timestep to arrive at specific numerical algorithm.

Appendix D: Robustness of DMS.BD to Choice of Initial All-atom Structure

[00188] After finding a local FE minimum structure (called “descent 2” structure) at $\tau = 79$ (Figure 7 e), a control DMS.BD simulation was launched starting at “descent 2” structure (Figures 8 and 10). A second simulation with narrowed widths a_d was branched off from the control one at $\tau = 89$ (Figures 8 and 10). Knowledge of the “descent 2” structure (Table S1) was used in constructing Δ^- for these two simulations. The goals of these simulations were to 1) further test if the DMS.BD algorithm is robust enough to not require knowledge of the FE bottom location, but can be started at some arbitrary structure within a basin; and 2) speed up finding the FE basin 3 before even reaching the bottom of the second basin.

[00189] At the same time, the initial (basin 1→2) DMS.BD simulation with only one basin structure known was continued in order to find a lower energy structure representative of

basin 2 (Figures 7 *f*, 9, 10). Overall, this simulation allowed sampling structures from the pseudodiferic basin 2, but did not evolve away from the basin.

[00190] In contrast, the control simulation with added information on basin 2 (in the form of descriptors for “descent 2” structure, but not for the bottom of basin 2) exhibits sharp increase in eigenvalue 1 of the moment of inertia tensor (Figure 8 *a*, *top line*). This implies that structural evolution is speeded up as compared to the DMS.BD simulation which accounts for only one known basin. However, eigenvalues 2 and 3 do not deviate much from those in the latter basin 1→2 run (Figures 8 *b-c*, *unlabeled line*). Narrowing the Gaussian-like exponents in Δ^- factors (**10'**) helps to push the system out of known basins more efficiently. This was shown for basin 2, where the increase of parameters a_d in (**10'**) from $(9.0 \cdot 10^{-8}, 2.6 \cdot 10^{-8}, 4.7 \cdot 10^{-8})$ in the control run to $(9.5 \cdot 10^{-6}, 2.0 \cdot 10^{-6}, 1.9 \cdot 10^{-6})$ in the run branched off at $\tau = 89$ enables successful basin 2→3 DMS.BD transition (Figure 8, *basin 3*). A considerable increase in all three eigenvalues was achieved in the latter run (Figure 8), implying overall expansion of lactoferrin.

Supporting Information

Details on MD Used to Generate Ensembles for Thermal-Average Force and Diffusion Computations and for Basin Sampling

[00191] MD ensemble generation in DMS.BD runs and sampling of the obtained FE basin structures were performed using NAMD [52'] in explicit solvent,[53',54'] which is a natural choice since the core of the DMS package already incorporates NAMD. [25'] Missing hydrogen atoms of the protein were added using the psfgen package of VMD with all-hydrogen topology file version 31 for proteins and nucleic acids.[55'.56'] The water molecules in the crystal structure of the diferic lactoferrin were discarded and the protein was solvated using the rigid three-site TIP3P model for water [57'] in a rectangular box with a minimum 15 Å distance between the box edges and the solute. Eight chloride ions were added to neutralize the positive charges of the protein. Default protonation states at a neutral pH were assigned to all ionic residues. Periodic boundary conditions were applied during the MD simulations. The particle mesh Ewald method [58'] was adopted to treat the long-range electrostatic interactions. All

calculations were carried out with a time constant of 1 fs and a cutoff of 12 Å with switching distance 10 Å. At every Langevin timestep, the structure obtained in the previous step was prepared for MD sampling by potential energy minimization using 1000 steps of the conjugate gradient method, and isothermal isobaric ensemble (NPT) thermal equilibration for 1 ps. Atom coordinates were recorded every 10 fs for calculating the factors in the OP dynamics equation (5'), and every 1 ps during MD sampling of the discovered FE basins. During the MD production phase at 310 °K the Langevin dynamics was used for the temperature control.

[00192] The starting structure for the DMS.BD calculation is prepared either by running traditional MD or multiscale MD (as in above reveiw) as implemented in DMS. [5'] While DMS provides lower energy structure and at a less computational expense, in the present validation of DMS.BD method the initial structure was obtained using a short canonical ensemble (NVT) MD run at constant volume after the above described minimization and thermalization steps. From the configurations sampled in the latter stage, the one with the lowest protein energy was selected as an initial structure. This allowed us to demonstrate the robustness of the sequential elimination algorithm versus a random choice of an initial macromolecular configuration and, in particular, some metastable initial state. Also this shows that knowledge of an initial structure is not needed and it can be easily generated using traditional MD.

Details on DMS.BD Workflow, Parameters, and Discovered States

[00193] The steps in DMS.BD algorithm involved in guiding the macromolecular structure to new basin are as follows. An input macromolecular structure is used as the reference structure $\underline{r}^o$ to construct OPs via Eq. (2'). Langevin timestep starts by generating an OP-constrained ensemble of N_{ens} all-atom structures, each described by a set of atomic coordinates $\underline{r}$. Atomic coordinates $\underline{r}_{\text{OP}}$ are reconstructed from OPs using Eq. (1') and are used to determine the residual displacements $\underline{\sigma}$ as the differences between actual coordinates $\underline{r}$ and $\underline{r}_{\text{OP}}$. A set of atomic forces $\underline{F}$ for each of N_{ens} all-atom structures from MD ensemble, and $k_{\text{max}}^3 N_{\text{ens}}$ OP forces $\tilde{f}_k^m$ (21') and OP velocities are calculated in order to compute the k_{max}^3 thermal-average forces $\langle \tilde{f}_k^m \rangle$ and k_{max}^3 diagonal elements $\bar{\bar{D}}_{kk}$ of the diffusivity tensor in

Eq. (5') with an account for known FE basins. Langevin equation is then solved numerically to update the OPs. The evolved OP-constrained atomic positions $\underline{r}_{\text{OP}}^{\text{new}}$ are then reconstructed from the latter; then the residuals are reintroduced, $\underline{r}^{\text{new}} = \underline{r}_{\text{OP}}^{\text{new}} + \underline{\sigma}$. Finally, the structure is resolvated. The reference structure $\underline{r}^{\circ}$ is updated after some number of Langevin timesteps. Once in a while, verification is performed if a new FE basin is reached.

[00194] Technical details on the performed DMS.BD guided simulations are as follows. To obtain a molecular configuration representative of an initial diferric FE basin, the closed X-ray structure 1LFG was prepared for simulations as described above. The structure with the lowest protein energy (called 1LFG MD structure) was selected from the last MD production-run stage (Figure 2 a'). As Table S1 and Figure 2 a, a' show, the thermalization of 1LFG crystal structure does not alter it considerably. For example, the number of residues which participate in either helical or beta-strand structures decreases from 372 in 1LFG to 352 in 1LFG MD, i.e., by only 5% (Table S1). Energy is not shown for X-ray structure since it was crystallized at a low temperature.

[00195] Next, DMS simulation was started from the 1LFG MD structure. This simulation has led to diferric basin structure (Figure 2 c), as indicated by vanishing thermal-average forces (Figure 4). Among structures compatible with OPs from the bottom of this diferric basin, the most probable all-atom structure with the maximum of Boltzmann weight, i.e., with minimum potential energy is chosen to represent the basin for obtaining the descriptors and illustration purposes (Table S1 and Figure 2 c). The latter structure is close to, but not the same as the FE minimum structure. However, since it implies the same OPs as those of the FE minimum, descriptors obtained from this structure suffice to characterize the diferric basin. Similar procedure is used to compute molecular descriptors for all the other basins shown here.

[00196] The values of the descriptors $\theta_{d,1}, d=1,2,3$ representing the diferric basin (Table S1) are determined by the values of OPs at the bottom of this basin using the relationship developed in Appendix A. The N_d inverse width parameters a_d in Eq. (10') were determined empirically, i.e., by adjusting them until thermal-average forces along the DMS.BD trajectory indicate escape from the basin by having a profile similar to that in Figure 4. However, in the

case of transition between diferric and pseudodiferric basins, the initial guess on a_d (24') was sufficient.

[00197] To make the initial estimate for a_d , the following two values were chosen: relative change of each descriptor η_d after transition from the last discovered basin to a new one, $\zeta_d = \theta_{d, N_{\text{known}}+1} / \theta_{d, N_{\text{known}}} - 1 = 0.15$, and additional state-counting factor (10') for the new basin, $\Delta^-(\eta; \theta_{N_{\text{known}}+1}) = 0.99$. This order of magnitude for ζ_d can be suggested, e.g., by the difference of the descriptors $\theta_{d,1}$ for the diferric basin and $\theta_{d,2}$ for the apolactoferrin structure (third row of data in Table S1). The same values were used for both simulated FE basin transitions to estimate the inverse widths as

$$a_d = -\frac{\ln\left(1 - \Delta^-(\eta; \theta_{N_{\text{known}}+1})\right)}{\zeta_d \left| \theta_{d, N_{\text{known}}} \right|}. \quad (24')$$

As a result, for the initial DMS.BD-guided transition from basin 1 to basin 2 the following estimate was used: $a_1 = 9.0 \cdot 10^{-8}$, $a_2 = 2.6 \cdot 10^{-8}$, $a_3 = 4.7 \cdot 10^{-8}$. Note, however, that DMS.BD does not require knowledge of a target structure or a basin; here the information on them was used only to assess the a_d magnitudes.

[00198] To facilitate comparison of the first two discovered FE basins, the Ramachandran plots [59'] for the representative protein structures (Figure 7 c, f), together with the highest-energy transition structure (Figure 7 d), are overlapped in Figure 13. Structure representative of a transition state was chosen from the OP-constrained ensemble corresponding to zero of the selected thermal-average forces by taking the structure with minimum potential energy (similar to the structures representing basins). Beginning with the basin 2→3 transition point at $\tau = 114$ (Figure 5 g), the structures have two of the three protein lobes more open (Figures 7 g, h) and their distinction from basin 1 structures becomes obvious.

TABLE S1: Characteristics of human lactoferrin structures discovered using DMS and DMS.BD

structure name	energy ^a (kcal/ mol)	eigenvalues ^b (10 ⁷ amu·Å ²)			RMSD (Å) ^c		SS (res.) ^d
					1LFG	1LFH	
1LFG MD → basin 1 → basin 2 DMS.BD simulation							
1LFG X-ray	N/A	2.11699	5.22844	5.92435	0	6.42580	372
1LFG MD ^e	-24686	2.15083	5.35982	5.96525	1.94763	6.49242	352
1LFH X-ray	N/A	2.42499	4.50417	5.38791	6.78823	0	378
basin 1	-25273	2.17045	5.27381	5.91399	2.07272	6.63463	339
basin 1 MD	-25884	2.19232	5.09673	5.68184	2.58986	5.93308	345
transition 1→2	-24367	2.21977	5.36098	6.02694	2.75830	7.04831	335
descent 2	-25602	2.25518	5.40126	6.06159	2.83033	7.22222	335
descent 2 MD	-26175	2.22644	5.51651	6.18090	2.69150	7.35988	332
basin 2	-25700	2.27002	5.46262	6.19471	3.65766	8.24560	325
basin 2 MD	-26383	2.21718	5.34342	6.05574	3.07368	7.89076	332
descent 2 → basin 3 DMS.BD simulation							
transition 2→3	-24113	2.52843	5.69365	6.51387	5.10911	8.84757	226
basin 3	-26490	2.49442	5.83352	6.69610	5.19012	8.75881	268
basin 3 MD	-27776	2.39254	5.92919	6.35707	5.44923	8.44729	271

^a Protein potential energy, ^b three eigenvalues of its moment of inertia tensor, ^c RMSD for backbone atoms relative to initial 1LFG and open 1LFH crystallographic structures, and ^d number of secondary-structure residues. ^e Structures from different FE basins were further optimized using NAMD to be closer to the bottom of corresponding basins (resulting structures have “MD” in their names).

[00199] FIG. 13 shows Ramachandran plots for discovered lactoferrin structures: (darker) bottom of basin 1, (lightest) top of the barrier for basin 1—*2 transition, and (lighter) bottom of basin

2. Most residues have dihedral angles within theoretically favored regions⁶⁰ shown as background shadows, which validates that the configurations obtained from DMS.BD simulation retain key secondary structure. Distribution of qnv angles is distinct for the three structures. Arrows indicate large changes in dihedral angles for hinge residues: (horizontal) THR90, (vertical) VAL250.

Results on MD Sampling of Basins

[00200] FIG. 14 shows descriptors and their ranges in each of the three discovered FE basins are shown below as an additional evidence of crossing the free energy barriers and resulting distinction between basin characteristics. In a) at 1 ns, the top curve is basin 3; the next curve is basin 2, the next curve is "descent 2", and the bottom curve is basin 1. In b) at 1 ns, the top curve is basin 3; and the other curves overlap with "descent 2" on top, basin 2 in the middle and basin 1 on the bottom. In b) at —3.5 ns, the top curve is basin 3; the next curve is "descent 2"; and the basin 2 and basin 1 curves overlap. In b) from —5.5 ns to 10 ns, the order of curves from the top is basin 3, "descent 2", basin 2, basin 1, with various overlaps.

REFERENCES

- (1) Hart, G. T.; Ramani, A. K.; Marcotte, E. M., How complete are current yeast and human protein interaction networks? *Genome Biol.*, 2006. **7**: p. 120-128.
- (2) Stumpf, M. P.; Thorne, T.; de Silva, E.; Stewart, R.; An, H. J.; Lappe, M.; Wiuf, C., Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA*, 2008. **105**: p. 6959-6964.
- (3) Sharon, M., How far can we go with structural mass spectrometry of protein complexes. *J. Am. Soc. Mass Spectrom.*, 2010. **21**: p. 487-500.
- (4) Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V., Evidence for macromolecular protein rings in the absence of bulk water. *Science*, 2005. **310**: p. 1658-1661.
- (5) Uetrecht, C.; Rose, R. J.; van Duijn, E.; Lorenzen, K.; Heck, A. J. R., Ion mobility mass spectrometry of proteins and protein assemblies. *Chem. Soc. Rev.*, 2010. **39**: p. 1633-1655.
- (6) Bernstein, S. L.; Dupuis, N. F.; Lazo, N. D.; Wytenbach, T.; Condrón, M. M.; Bitan, G.; Teplow, D. B.; Shea, J.-E.; Ruotolo, B. T.; Robinson, C. V.; Bowers, M. T., Amyloid- β protein oligomerization and the importance of tetramers and dodecamers in the aetiology of alzheimer's disease. *Nature Chem.*, 2009. **1**: p. 326-331.
- (7) Pukala, L.; Ruotolo, B. T.; Zhou, M.; Politis, A.; Stefanescu, R.; Leary, J. A.; Robinson, C. V., Subunit architecture of multiprotein assemblies determined using restraints from gas phase measurements. *Structure*, 2009. **17**: p. 1235-1243.
- (8) Uetrecht, C.; Versluis, C.; Watts, N. R.; Wingfield, P. T.; Steven, A. C.; Heck, A. J. R., Stability and shape of hepatitis B virus capsid in vacuo. *Angew. Chem. Int. Ed*, 2008. **47**: p. 6247-6251.
- (9) Smith, D. P.; Radford, S. E.; Ashcroft, A. E., Elongated oligomers in β 2-microglobulin amyloid assembly revealed by ion mobility spectrometry-mass spectrometry. *Proc. Natl. Acad. Sci. USA*, 2010. **107**: p. 6794-6798.
- (10) Hirschfelder, J. O.; Curtiss, C. F.; Bird, R. B. *Molecular theory of gases and liquids*. Wiley: New York, 1964.
- (11) Smith, D. P.; Knapman, T. W.; Campuzano, I.; Malham, R. W.; Berryman, J. T.; Radford, S. E.; Ashcroft, A. E., Deciphering drift time measurements from travelling wave ion mobility spectrometry-mass spectrometry studies. *Eur. J. Mass Spectrom.*, 2009. **15**: p. 113-130.
- (12) von Helden, G.; Hsu, M. T.; Gotts, N. G.; Bowers, M. T., Carbon cluster cations with up to 84 atoms: Structures, formation mechanism, and reactivity. *J. Phys. Chem.*, 1993. **97**: p. 8182-8192.

- (13) Shvartsburg, A. A.; Jarrold, M. F., An exact hard spheres scattering model for the mobilities of polyatomic ions. *Chem. Phys. Lett.*, 1996. **261**: p. 86-91.
- (14) Mesleh, M. F.; Hunter, J. M.; Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F., Structural information from ion mobility measurements: Effects of the long range potential. *J. Phys. Chem.*, 1996. **100**: p. 16082-16086.
- (15) Shvartsburg, A. A.; Mashkevich, S. V.; Baker, S. E.; Smith, R. D., Optimization of algorithms for ion mobility calculations. *J. Phys. Chem. A*, 2007. **111**: p. 2002-2010.
- (16) Shvartsburg, A. A.; Schatz, G. C.; Jarrold, M. F., Mobilities of carbon cluster ions: Critical importance of the molecular attractive potential. *J. Chem. Phys.*, 1998. **108**: p. 2416-2423.
- (17) van Duijn, E.; Barendregt, A.; Synowsky, S.; Versluis, C.; Heck, A. J. R., Chaperonin complexes monitored by ion mobility mass spectrometry. *J. Am. Chem. Soc.*, 2009. **131**: p. 1452-1459.
- (18) Fernandez-Lima, F. A.; Wei, H.; Gao, Y. Q.; Russell, D. H., On the structure elucidation using ion mobility spectrometry and molecular dynamics. *J. Phys. Chem. A*, 2009. **113**: p. 8221-8234.
- (19) Baumketner, A.; Bernstein, S. L.; Wytenbach, T.; Bitan, G.; Teplow, D. B.; Bowers, M. T.; Shea, J. E., Amyloid β -protein monomer structure: A computational and experimental study. *Prot. Sci.*, 2006. **15**: p. 420-428.
- (20) Damsbo, M.; Kinnear, B. S.; Hartings, M. R.; Ruhoff, P. T.; Jarrold, M. F.; Ratner, M. A., Application of evolutionary algorithm methods to polypeptide folding: Comparison with experimental results for unsolvated ac-(ala-gly-gly)₅-lysh+. *Proc. Natl. Acad. Sci. USA*, 2004. **101**: p. 7215-7222.
- (21) Bose, S.; Medina-Noyola, M.; Ortoleva, P. J., Third body effects on reactions in liquids. *J. Chem. Phys.*, 1981. **75**: p. 1762-1771.
- (22) Bose, S.; Ortoleva, P. J., Hard-sphere model of chemical-reaction in condensed media. *Phys. Lett. A*, 1979. **69**: p. 367-369.
- (23) Bose, S.; Ortoleva, P. J., Reacting hard-sphere dynamics - Liouville equation for condensed media. *J. Chem. Phys.*, 1979. **70**: p. 3041-3056.
- (24) Cheluvvaraja, S.; Roy, A.; Ortoleva, P., Roadmap for *SimNanoWorld*TM an all-atom nanosystem simulator. *In preparation*, 2010.
- (25) Miao, Y.; Ortoleva, P. J., All-atom multiscaling and new ensembles for dynamical nanoparticles. *J. Chem. Phys.*, 2006. **125**: p. 44901-44908.
- (26) Ortoleva, P. J., Nanoparticle dynamics: A multiscale analysis of the liouville equation. *J. Phys. Chem. B*, 2005. **109**: p. 21258-21266.

- (27) Ortoleva, P. J.; Adhangale, P.; Cheluvvaraja, S.; Fontus, M. W. A.; Shreif, Z., Deriving principles of microbiology by multiscaling laws of molecular physics: Applications in nanomedicine and eneregy. *IEEE Eng. Med. Biol.*, 2009. **28**: p. 70-79.
- (28) Pankavich, S.; Miao, Y.; Ortoleva, J.; Shreif, Z.; Ortoleva, P. J., Stochastic dynamics of bionanosystems: Multiscale analysis and specialized ensembles. *J. Chem. Phys.*, 2008. **128**: p. 234908-234920.
- (29) Pankavich, S.; Ortoleva, P., Multiscale theory of soft matter. *In preparation*, 2010.
- (30) Pankavich, S.; Shreif, Z.; Miao, Y.; Ortoleva, P. J., Self-assembly of nanocomponents into composite structures: Derivation and simulation of Langevin equations. *J. Chem. Phys.*, 2009. **130**: p. 194115-194124.
- (31) Pankavich, S.; Shreif, Z.; Ortoleva, P. J., Multiscaling for classical nanosystems: Derivation of Smoluchowski and Fokker-Planck equations. *Physica A*, 2008. **387**: p. 4053-4069.
- (32) Shreif, Z.; Ortoleva, P. J., Multiscale approach to nanocapsule design. *Technical Proceedings of the 2008 NSTI Nanotechnology Conference and Trade Show*, 2008. **3**: p. 741-744.
- (33) Shreif, Z.; Ortoleva, P. J., Curvilinear all-atom multiscale (CAM) theory of macromolecular dynamics. *J. Stat. Phys.*, 2008. **130**: p. 669-685.
- (34) Shreif, Z.; Ortoleva, P. J., Multiscale derivation of an augmented Smoluchowski equation. *Physica A*, 2009. **388**: p. 593-600.
- (35) Shreif, Z.; Ortoleva, P. J., Computer-aided design of nanocapsules for therapeutic delivery. *Comput. Math. Methods Med.*, 2009. **10**: p. 49-70.
- (36) Shreif, Z.; Pankavich, S.; Ortoleva, P. J., Liquid-crystal transitions: A first-principles multiscale approach. *Phys. Rev. E*, 2009. **80**: p. 031703.
- (37) Ortoleva, P. In *Macroscopic self organization at geological and other first order phase transitions - Non-equilibrium dynamics in chemical systems: Proceedings of the International Symposium*; Vidal, C.; Pacault, A.; Eds. NY: Springer-Verlag: Bordeaux, France, 1984. Vol. p 94-98.
- (38) Chen, C.; Xiao, Y.; Zhang, L., A direct essential dynamics simulation of peptide folding. *Biophys. J.*, 2005. **88**: p. 3276-3285.
- (39) Hayward, S.; Kitao, A.; Berendsen, H. J. C., Model-free methods of analyzing domain motions in protiens from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins: Struct. Funct. Genet.*, 1997. **14**: p. 1767-1717.

- (40) Hayward, S.; Kitao, A.; Go, N., Harmonicity and anharmonicity in protein dynamics: A normal mode analysis and principal component analysis. *Proteins: Struct. Funct. Genet.*, 1995. **23**: p. 177-186.
- (41) Praprotnik, M. S. M.; Delle Site, L.; Kremer, K.; and Clementi C., Adaptive resolution simulation of liquid water *J. Phys.: Condens. Mat.*, 2009. **21**: p. 499801- 499810.
- (42) Arkhipov, A.; Freddolino, P. L.; Schulten, K., Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 2006. p. 1767-1777.
- (43) Zhang, Z.; Pfaendtner, J.; Grafmüller, A.; Voth, G. A., Defining coarse-grained representations of large biomolecules and biomolecular complexes from elastic network models. *Biophys. J.*, 2009. **97**: p. 2327-2337.
- (44) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A., A systematic methodology for defining coarse-grained sites in large biomolecules. *Biophys. J.*, 2008. **95**: p. 5073-5083.
- (45) Beardsley, R. L.; Running, W. E.; Reilly, J. P., Probing the struture of caulobacter crescentus ribosome with chemical labeling and mass spectrometry, *J. Proteome Res.*, 2006. **5**: p. 2935-2946.
- (46) Zhou, K.; Kovarik, M. L.; Jacobson, S. C., Surface-charge induced ion depletion and sample stacking near single nanopores in microfluidic devices. *J. Am. Chem. Soc.*, 2008. **130**: p. 8614-8616.
- (47) Best, R. B.; Chen, Y. G.; Hummer, G., Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure*, 2005. **13**: p. 1755-1763.
- (48) Oliveira, L. C.; Schug, A.; Onuchic, J. N., Geometrical features of the protein folding mechanism are a robust property of the energy landscape: A detailed investigation of several reduced models. *J. Phys. Chem.*, 2008. **112**: p. 6131-6136.
- (49) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comp. Chem.*, 1992. **13**: p. 1011-1021.
- (50) Kim, J.; Keyes, T.; Straub, J. E., Relationship between protein folding thermodynamics and the energy landscape. *Phys. Rev. E*, 2009. **79**: p. 030902-030905(R).
- (51) Okazaki, K.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G., Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci.*, 2006. **103**: p. 11844-11849.
- (52) Iyengar, S. S.; Ortoleva, P. J., Multiscale theory of collective and single-particle modes in quantum nanosystem. *J. Chem. Phys.*, 2008. **128**: 164716-164721.

- (53) Miao, Y.; Ortoleva, P. J., Viral structural transition mechanisms revealed by multiscale molecular dynamics/order parameter extrapolation simulation. *Biopolymers*, 2010. **93**: p. 61-73.
- (54) Miao, Y.; Ortoleva, P. J., Molecular dynamics/order parameter extrapolation (MD/OPX) for bionanosystem simulations. *J. Comput. Chem.*, 2009. **30**: p. 423-437.
- (55) Miao, Y.; Ortoleva, P. J., Viral structural transitions: An all-atom multiscale theory. *J. Chem. Phys.*, 2006. **125**: p. 214901-214911.
- (56) Weitzke, E. L.; Ortoleva, P. J., Simulating cellular dynamics through a coupled transcription, translation, metabolic model. *Comput. Biol. Chem.*, 2003. **27**: p. 469-481.
- (57) Singharoy, A.; Cheluvaraja, S.; Ortoleva, P. J., Order parameters for macromolecules: Application to multiscale simulation. *J. Chem. Phys.*, 2011. **134**: p. 044104-044120.
- (58) Cheluvaraja, S.; Ortoleva, P. J., Thermal nanostructure: An order parameter/multiscale ensemble approach. *J. Chem. Phys.*, 2010. **132**: p. 075102-075110.
- (59) Miao, Y.; Johnson, J. E.; Ortoleva, P. J., All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. *J. Phys. Chem. B*, 2010. **114**: p. 11181-11195.
- (60) Jaqaman, K.; Ortoleva, P. J., New space warping method for the simulation of large-scale macromolecular conformational changes. *J. Comput. Chem.*, 2002. **23**: p. 484-491.
- (61) Shreif, Z.; Adhangale, P.; Cheluvaraja, S.; Perera, R.; Kuhn, R. J.; Ortoleva, P. J., Enveloped viruses understood via multiscale simulation: Computer-aided vaccine design. *Sci. Model. Simul.*, 2008. **15**: p. 363-380.
- (62) Pankavich, S.; Shreif, Z.; Chen, Y.; Ortoleva, P. J., Multiscale theory of finite size bose systems: Implications for collective and single-particle excitations. *Phys. Rev. A*, 2009. **79**: p. 013628-013635.
- (63) Pankavich, S.; Ortoleva, P. J., Mutiscaling for systems with a broad continuum of characteristic lengths and times *J. Math. Phys.*, 2010. **Accepted with revisions**.
- (64) Jaynes, E. T., *Informaiton theory and statistical mechanics*. Benjamin, New York, 1963; p 181.
- (65) Nitzan, A.; Ortoleva, P. J.; Deutch, J.; Ross, J., Fluctuations and transitions at chemical instabilities - analogy to phase-transitions. *J. Chem. Phys.*, 1974. **61**: p. 1056-1074.
- (66) Singharoy, A.; Yesnik, A.; Ortoleva, P. J., Multiscale analytic continuation approach to nanosystem simulation: Applications to virus electrostatics. *J. Chem. Phys.*, 2010. **132**: p. 174112-174126.

- (67) Ortoleva, P. J., Dynamic Pade approximants in theory of periodic and chaotic chemical center waves. *J. Chem. Phys.*, 1978. **69**: p. 300-307.
- (68) Singharoy, A.; Ortoleva, P. J., A hierarchical multiscale approach to the theory of macromolecular assemblies. *In preparation*, 2010.
- (69) Singharoy, A.; Joshi, H.; Cheluvvaraja S.; Brown, D.; Ortoleva, P. J., *Simulating microbial systems: Addressing model uncertainty/incompleteness via multiscaling and entropy methods* Springer Science: New York, 2010.
- (70) Singharoy, A.; Cheluvvaraja, C.; Joshi, H.; McWilliams, K.; Brown, D.; Ortoleva, P. J., Simulating the self-assembly and stability of VLP-type vaccines: Application to human papillomavirus. *In preparation*, 2010.
- (71) Shreif, Z.; Ortoleva, P. J., Multiscale Born-Oppenheimer theory of collective electron-nuclear dynamics in nanosystems. *J. Chem. Theory Comput.*, 2011. **Accepted**.
- (72) Shreif, Z.; Ortoleva, P. J., Scaling behavior of quantum nanosystems: Emergence of quasi-particles, collective modes, and mixed exchange symmetry states. *J. Chem. Phys.*, 2011. **Accepted**.
- (73) Shreif, Z.; Ortoleva, P. J., All-atom/continuum multiscale theory: Application to nanocapsule therapeutic delivery. *In preparation* 2010.
- (74) Shreif, Z.; Joshi, H.; Ortoleva, P., Liposomes and enveloped viruses understood the multiscale way. *In preparation*, 2010.
- (75) Bose, S.; Bose, S.; Ortoleva, P., Dynamic Pade approximants for chemical-center waves. *J. Chem. Phys.*, 1980. **72**: p. 4258-4263.
- (76) Ortoleva, P. J., Multiscale theory of bosonic excitations in fermion nanosystems. *In preparation*, 2010.
- (77) Fan, H.; Perkins, C.; Ortoleva, P., Scaling behavior of electronic excitations in assemblies of molecules with degenerate ground states. *J. Phys. Chem. A*, 2010. **114**: p. 2213-2220.
- (78) Cheluvvaraja, S.; Ortoleva, P. J., Multiple order parameter, multiscale theory of composite nanosystems. *In preparation*, 2010.
- (79) Sayyed-Ahmad, A.; Miao, Y.; Ortoleva, P. J., Poisson-Boltzmann theory for bionanosystems. *Commun. Comput. Phys.*, 2008. **3**: p. 1100-1116.
- (80) Jarymowycz, L.; Ortoleva, P. J., Involatile nanodroplets: An asymptotic analysis. *J. Chem. Phys.*, 2006. **124**: p. 234705-234708.
- (81) Elber, R., A milestoning study of the kinetics of an allosteric transition: Atomically detailed simulations of deoxy scapharca hemoglobin. *Biophys. J.*, 2007. **92**: p. L85-L87.

- (82) Kinnear, B. S.; Kaleta, D. T.; Kohtani, M.; Hudgins, R. R.; Jarrold, M. F., Conformations of unsolvated valine-based peptides. *J. Am. Chem. Soc.*, 2000. **122**: p. 9243-9256.
- (83) van Gunsteren, W. F.; Berendsen., H. J. C., Computer-simulation of molecular-dynamics methodology, applications, and perspectives in chemistry. *Angew. Chem. Int. Ed.*, 1990. **29**: p. 992-1023.
- (84) Cao, Z; Wang, J., A comparative study of two different force fields on structural and thermodynamics character of h1 peptide via molecular dynamics simulations. *J. Biomol. Struct. Dyn.*, 2010 **27**: p. 651-661.
- (85) Feig, M.; Pettitt, B., Structural equilibrium of DNA represented with different force fields. *Biophys. J.*, 1998. **75**: p. 134-149.
- (86) Guvench O.; Jr., MacKerell A. D., Comparison of protein force fields for molecular dynamics simulations. *Methods Mol. Biol.*, 2008. **443**: p. 63-88.
- (87) Norberg J; Nilsson, L., Advances in biomolecular simulations: Methodology and recent applications. *Q. Rev. Biophys.*, 2003. **36**: p. 257-306.
- (88) Lange, O. F.; van der Spoel D.; de Groot, B. L., Scrutinizing molecular mechanics force fields on the microsecond timescale with NMR data. *Biophys. J.*, 2010. **99**: p. 647-655.
- (89) Ricci, C. G.; de Andrade, A. S.; Mottin, M.; Netz P. A., Molecular dynamics of DNA: Comparison of force fields and terminal nucleotide definitions. *J. Phys. Chem. B*, 2010. **114**: p. 9882-9893.
- (90) Rueda, M.; Ferrer-Costa, C.; Meyer, T.; Pérez, A.; Camps, J.; Hospital, A.; Gelpí, J. L.; Orozco, M., A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. USA*, 2007. **104**: p. 796-801.
- (91) Wang, W.; Donini, O.; Reyes, C. M.; Kollman, P. A., Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Ann. Rev. Biophys. Biomol. Struct.*, 2001. **30**: p. 211-243.
- (92) Freddolino, P. L.; Schulten, K., Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophys. J.*, 2009. **97**: p. 2338-2347.
- (93) Klepeis, J. L.; Lindorff-Larsen, K.; Shaw, D. E., Long-time- scale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 2009. **19**: p. 120-127.
- (94) Stortz, C. A.; Johnson, G. P.; French, A. D.; Csonka G. I., Comparison of different force fields for the study of disaccharides. *Carbohydr. Res.*, 2009. **344**: p. 2217-2228.
- (95) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Jr. Mackarell, A.D., Charmm general force field: A force field for

drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.* 2009. **31**: p. 671–690.

(96) Patel, S.; Jr. MacKerell, A.D.; Brooks III, C. L., Charmm fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.*, 2004. **25**: p. 1–15.

(97) Patel, S.; MacKerell, J. A.; Brooks III, C. L., Charmm fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.*, 2004. **25**: p. 1504–1514.

(98) Lamoureux, G; Roux, B., Modeling induced polarization with classical drude oscillators: Theory and molecular dynamics simulation algorithm. *J. Chem. Phys.*, 2003. **119**: p. 3025-3039.

(99) Lamoureux, G.; Harder E; Vorobyov, I.V.; Roux B; Mackarell, A., D., A polarizable model of water for molecular dynamics simulations of biomolecules. *Chem. Phys. Lett.* 2006. **418**: p. 245-249.

(100) Philips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, 2005. **26**: p. 1781-1802.

(101) Antson, A. A.; Dodson, E. J.; Dodson, G.; Greaves, R. B.; Chen, X. P.; Gollnick, P., Structure of the TRP rna-binding attenuation protein, trap, bound to rna. *Nature.*, 1999. **401**: p. 235-242.

(102) Shelimov, K.B; Clemmer, D.E.; Jarrold, M.F., Protein Structure in Vacuo: The Gas Phase Conformations of BPTI and Cytochrome c. *J. Am. Chem. Soc.*, 1997. **119**: p. 2240-2248.

(103) Freeke, J.; Robinson, C. V., Ruotolo, B. T., Residual counter ions stabilize a large protein complex in the gas phase. *Int. J. Mass Spectrom* 2010. **In press**.

(104) Verkerk, P. K.; Verkerk, U. H., Electrospray: From ions in solution to ions in the gas phase, what we know now. *Mass Spectrom. Rev.*, 2009. **28**: p. 898-917.

(105) Lauber, M. A.; Reilly, J. P., A novel amidinating cross-linker for facilitating analyses of protein structures and interactions. *Anal. Chem.*, 2010. **In press**.

(106) Cusack, S.; Smith, J. C.; Finney, J. L.; Tidor, B.; Karplus, M., Inelastic neutron scattering analysis of picosecond internal protein dynamics: Comparison of harmonic theory with experiment. *J. Mol. Biol.*, 1988. **202**: p. 903-908.

(107) Lamy, A.; Smith, J. C.; Yunoki, J.; Parker, S. F.; Kataoka, M., High-resolution vibrational inelastic neutron scattering: A new spectroscopic tool for globular proteins. *J. Am. Chem. Soc.*, 1997. **119**: p. 9268-9273.

- (108) Nelkin, M.; Ortoleva, P. J., Collective modes in liquids and neutron scattering. In *Neutron inelastic scattering*; International Atomic Energy Agency: Vienna, 1968; Vol. 1, p. 535-544.
- (109) Ortoleva, P. J.; Nelkin, M., Fluctuations of single-particle distribution function in classical fluids. *Phys. Rev.*, 1969. **181**: p. 429-31.
- (110) Zhou, K.; Li, L.; Tan, Z.; Zlotnick, A.; Jacobson, S. C., Characterization of hepatitis B virus capsids by resistive-pulse sensing. *J. Am. Chem. Soc.*, 2011. Article ASAP
- (111) Kovarik, M. L.; Brown, P. J. B.; Kysela, D. T.; Berne, C.; Kinsella, A. C.; Brun, Y. V.; Jacobson, S. C., Microchannel-nanopore device for bacterial chemotaxis assays. *Anal. Chem.*, 2010. **82**: p. 9357-9364.
- (112) Spletzer, M.; Raman, A.; Reifenberger, R., Elastometric sensing using higher flexural eigenmodes of microcantilevers. *Appl. Phys. Lett.*, 2007. **91**: p. 184103-184105.
- (113) Hu, S. Q.; Raman, A., Analytical formulas and scaling laws for peak interaction forces in dynamic atomic force microscopy. *Appl. Phys. Lett.*, 2007. **91**: p. 123106-123108.
- (1') Ruotolo, B. T.; Giles, K.; Campuzano, I.; Sandercock, A. M.; Bateman, R. H.; Robinson, C. V. *Science* **2005**, *310*, 1658-1661.
- (2') Binnig, G.; Quate, C. F.; Gerber, C. *Phys. Rev. Lett.* **1986**, *56*, 930.
- (3') Beardsley, R. L., Running, W. E., Reilly, J. P. *J. Proteome Res.* **2006**, *5*, 2935-2946.
- (4') Keyser, U. F.; Koeleman, B. N.; van Dorp, S.; Krapf, D.; Smeets, R. M. M.; Lemay, S. G.; Dekker, N. H.; Dekker, C. *Nat. Phys.* **2006**, *2*, 473-477.
- (5') Joshi, H.; Singharoy, A. B.; Sereda, Y. V.; Cheluvraja, S. C.; Ortoleva, P. J. *Prog. Biophys. Mol. Biol.* **2011**, *107*, 200-217.
- (6') Rangwala, H.; Karypis, G. Introduction to Protein Structure Prediction. In *Introduction to Protein Structure Prediction*; John Wiley & Sons, Inc., 2010; pp 1-13.
- (7') Desmet, J.; Maeyer, M. D.; Hazes, B.; Lasters, I. *Nature* **1992**, *356*, 539-542.
- (8') Georgiev, I.; Donald, B. R. *Bioinformatics* **2007**, *23*, i185-i194.
- (9') Heath, A. P.; Kavraki, L. E.; Clementi, C. *Proteins: Struct. Funct. Bioinf.* **2007**, *68*, 646-661.
- (10') Bower, M.; Cohen, F.; Dunbrack, R. *J. Mol. Biol.* **1997**, *267*, 1268-1282.
- (11') Lee, C.; Subbiah, S. *J. Mol. Biol.* **1991**, *217*, 373-388.
- (12') Lee, C. *J. Mol. Biol.* **1994**, *236*, 918-939.
- (13') Floudas, C. A. *Biotechn. and Bioeng.* **2007**, *97*, 207-213.
- (14') Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087-1092.
- (15') Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. *Science* **1983**, *220*, 671-680.
- (16') Lee, C.; Levitt, M. *Nature* **1991**, *352*, 448-451.
- (17') Holland, J. H. *Adaptation in natural and artificial systems*; The University of Michigan Press: Ann Harbor, MI, 1975.

- (18') Unger, R. The Genetic Algorithm Approach to Protein Structure Prediction. In *Applications of Evolutionary Computation in Chemistry*; Johnston, R. L., Ed.; Springer Berlin / Heidelberg, 2004; Vol. 110; pp 2697-2699.
- (19') Kim, M. K.; Jernigan, R. L.; Chirikjian, G. S. *Biophys. J.* **2005**, 89, 43-55.
- (20') Marques, O.; Sanejouand, Y.-H. *Proteins: Struct. Funct. Bioinf.* **1995**, 23, 557-560.
- (21') Dellago, C.; Bolhuis, P. Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events. In *Advanced Computer Simulation Approaches for Soft Matter Sciences III*; Holm, C., Kremer, K., Eds.; Springer Berlin / Heidelberg, 2009; Vol. 221; pp 167-233.
- (22') Laio, A.; Parrinello, M. *PNAS* **2002**, 99, 12562-12566.
- (23') Barducci, A.; Bonomi, M.; Parrinello, M. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, 1, 826-843.
- (24') Trabuco, L. G.; Villa, E.; Schreiner, E.; Harrison, C. B.; Schulten, K. *Methods* **2009**, 49, 174-180.
- (25') Cheluvaraja, S.; Ortoleva, P. J. *J. Chem. Phys.* **2010**, 132, 075102.
- (26') Singharoy, A.; Cheluvaraja, S.; Ortoleva, P. J. *J. Chem. Phys.* **2011**, 134, 044104.
- (27') Singharoy, A.; Joshi, H.; Cheluvaraja, S.; Brown, D.; Ortoleva, P. J. Simulating Microbial Systems: Addressing Model uncertainty/incompleteness via Multiscaling and Entropy methods In *Microbial Systems Biology: Methods and Protocols*; Navid, A., Ed.; Springer Science: New York, 2010; Vol. 881.
- (28') Jaqaman, K.; Ortoleva, P. J. *J. Comput. Chem.* **2002**, 23, 484-491.
- (29') Miao, Y.; Ortoleva, P. J. *J. Comput. Chem.* **2009**, 30, 423-437.
- (30') Pankavich, S.; Miao, Y.; Ortoleva, J.; Shreif, Z.; Ortoleva, P. J. *J. Chem. Phys.* **2008**, 128, 234908-234920.
- (31') Schmidt, E. *Math. Ann.* **1907**, 63, 433-476.
- (32') Singharoy, A.; Sereda, Y. V.; Ortoleva, P. J. *submitted to J. Chem. Theor. Comp.* **2011**.
- (33') Miao, Y.; Ortoleva, P. J. *Biopolymers* **2010**, 93, 61-73.
- (34') Shreif, Z.; Ortoleva, P. J. *Technical Proceedings of the 2008 NSTI Nanotechnology Conference and Trade Show* **2008**, 3, 741-744.
- (35') Shreif, Z.; Ortoleva, P. J. *Comput. Math. Methods Med.* **2009**, 10, 49-70.
- (36') Ortoleva, P. J.; Adhangale, P.; Cheluvaraja, S.; Fontus, M. W. A.; Shreif, Z. *IEEE Eng. Med. Biol.* **2009**, 28, 70-79.
- (37') Pankavich, S.; Shreif, Z.; Ortoleva, P. J. *Physica A* **2008**, 387, 4053-4069.
- (38') Shreif, Z.; Ortoleva, P. J. *Physica A* **2009**, 388, 593-600.
- (39') Ortoleva, P. J. *J. Phys. Chem. B* **2005**, 109, 21258-21266.
- (40') Miao, Y.; Ortoleva, P. J. *J. Chem. Phys.* **2006**, 125, 44901-44908.
- (41') Bose, S.; Ortoleva, P. J. *J. Chem. Phys.* **1979**, 70, 3041-3056.
- (42') Bose, S.; Ortoleva, P. J. *Phys. Lett. A* **1979**, 69, 367-369.
- (43') Bose, S.; Medina-Noyola, M.; Ortoleva, P. J. *J. Chem. Phys.* **1981**, 75, 1762-1771.
- (44') Shreif, Z.; Ortoleva, P. J. *Stat. Phys.* **2008**, 130, 669-685.

- (45') Pankavich, S.; Shreif, Z.; Miao, Y.; Ortoleva, P. J. *J. Chem. Phys.* **2009**, *130*, 194115-194124.
- (46') Shreif, Z.; Pankavich, S.; Ortoleva, P. J. *Phys. Rev. E* **2009**, *80*, 031703.
- (47') Pankavich, S.; Ortoleva, P. J. *Math. Phys.* **2010**, *51*, 063303-063316.
- (48') Li, Z.; Scheraga, H. A. *PNAS* **1987**, *84*, 6611-6615.
- (49') Norris, G. E.; Anderson, B. F.; Baker, E. N. *Acta Crystallographica Sect. B* **1991**, *47*, 998-1004.
- (50') Gerstein, M.; Anderson, B. F.; Norris, G. E.; Baker, E. N.; Lesk, A. M.; Chothia, C. *J. Molec. Biol.* **1993**, *234*, 357-372.
- (51') Miao, Y.; Johnson, J. E.; Ortoleva, P. J. *J. Phys. Chem. B* **2010**, *114*, 11181-11195.
- (52') Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781-1802.
- (53') Kumar, R.; Skinner, J. L. *J. Phys. Chem. B* **2008**, *112*, 8311-8318.
- (54') Nerenberg, P. S.; Head-Gordon, T. *J. Chem. Theor. Comput.* **2011**, *7*, 1220-1230.
- (55') Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400-1415.
- (56') MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.
- (57') Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926(910 pages).
- (58') Darden, T.; York, D.; Pedersen, L. *Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems*; AIP, 1993; Vol. 98.
- (59') Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J. Mol. Biol.* **1963**, *7*, 95-99.
- (60') Lovell, S. C.; Davis, I. W.; Arendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins: Struct. Funct. Bioinf.* **2003**, *50*, 437-450.

CLAIMS

1. A method of acquiring mass spectral data using ion mobility mass spectrometry and interpretation of said data comprising using an ion mobility sequential elimination technique, wherein

using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

2. The method of claim 1, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

3. The method of claim 2, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

4. The method of claim 1, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

-71-

5. The method of claim 4, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

6. The method of claim 5, wherein buffer gas atoms can undergo more than one collision with the ion.

7. The method of claim 1, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

8. The method of claim 1, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

9. The method of any of any of claims 1-8 wherein the data is generated from a supramolecular assembly.

10. The method of claim 9 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

11. A method of analyzing ion mobility mass spectrometry mass spectral data comprising using an ion mobility sequential elimination technique, wherein using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

12. The method of claim 11, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

13. The method of claim 12, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

14. The method of claim 11, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

15. The method of claim 14, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

16. The method of claim 15, wherein buffer gas atoms can undergo more than one collision with the ion.

17. The method of claim 11, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body

Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

18. The method of claim 11, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

19. The method of any of any of claims 11-18 wherein the data is generated from a supramolecular assembly.

20. The method of claim 19 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

21. The method of claim 10 or 20 wherein the supramolecular assembly of biological molecules is a virus or portion thereof.

22. The method of claim 10 or 20 wherein the supramolecular assembly of biological molecules is a charged biomolecule.

23. A method of determining the molecular conformational state of a supramolecular assembly comprising

acquiring mass spectral data from said supramolecular assembly using ion mobility mass spectrometry and analyzing the acquired data using an ion mobility sequential elimination technique, wherein

using the ion mobility sequential elimination technique includes obtaining the ion's cross section from the fraction of trajectories that strike the ion, averaged over orientation and

the ion mobility sequential elimination technique uses a multiscale analysis, wherein the multiscale analysis is a suite of concepts and mathematical techniques used to simultaneously account for processes coupled across scales in space and time.

24. The method of claim 23, wherein using a ion mobility sequential elimination technique includes using a projection approximation, wherein a momentum transfer cross section is approximated by a two-dimensional projection of the ion and ignores the details of scattering processes and neglects long-range interactions.

25. The method of claim 24, wherein Monte Carlo integrations are implemented by randomly selecting the orientation of the ion, drawing a two-dimensional box around it, and then firing buffer gas atoms at the area inside the box, randomly selecting the x- and y-coordinates.

26. The method of claim 23, wherein using the ion mobility sequential elimination technique includes using an exact hard spheres scattering model, wherein all atoms are replaced by hard spheres and scattering angles are determined from hard sphere scattering, wherein long-range interactions between the ion and buffer gas atom are ignored.

27. The method of claim 26, wherein using the ion mobility sequential elimination technique further comprises using the exact hard spheres scattering model to calculate and use the collision integral.

28. The method of claim 27, wherein buffer gas atoms can undergo more than one collision with the ion.

29. The method of claim 23, wherein using the ion mobility sequential elimination technique includes using a trajectory method, wherein interactions between the ion and buffer gas atoms are approximated by an effective potential consisting of a sum of two-body

Lennard-Jones interactions and charge-induced dipole interactions, wherein scattering angles are determined from trajectory calculations.

30. The method of claim 23, wherein an amount of time for interpretation is reduced by not randomly searching regions of assembly configuration space not relevant to experimental measurements.

31. The method of claim 23 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

32. The method of claim 31 wherein the supramolecular assembly of biological molecules is a virus or portion thereof.

33. The method of claim 31 wherein the supramolecular assembly of biological molecules is a charged biomolecule.

34. A method of determining the molecular conformational state of a supramolecular assembly comprising

acquiring mass spectral data from said supramolecular assembly using ion mobility mass spectrometry and analyzing the acquired data using a free energy (FE) basin sequential elimination technique comprising:

an all-atom underlying formulation and continuous (and not discrete) configuration space; coarse-grained structural variables; a multiscale methodology to derive Langevin equations for the order parameters (OPs) and algorithms for computing all factors in these equations from an interatomic force field; an FE basin discovery method using modification of FE driving thermal-average forces for OP evolution that integrates prior knowledge of known FE minimizing structures to guide the evolution to yet-unknown ones; an efficient, calibration-free multiscale simulation methodology on which to build the methodical

search algorithm, and which is flexible enough to incorporate experimental data of a range of resolutions;

wherein the FE basin sequential elimination technique does not require prior knowledge of the reaction path, nor the final or initial structure.

35. The method of claim 34 wherein the supramolecular assembly is a supramolecular assembly of biological molecules.

36. The method of claim 34 wherein the supramolecular assembly of biological molecules is a virus or portion thereof.

37. The method of claim 34 wherein the supramolecular assembly of biological molecules is a charged biomolecule.

1/14

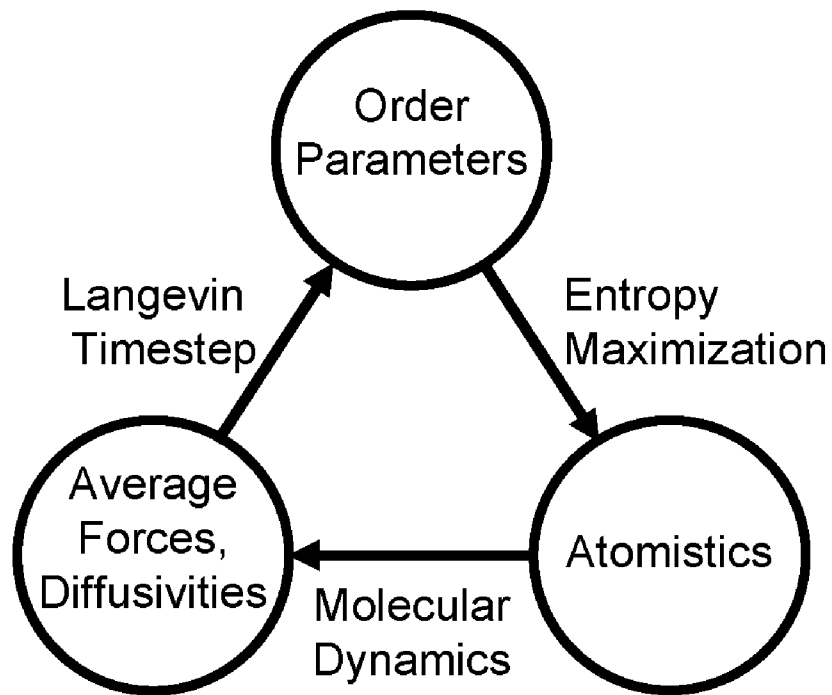


FIG. 1

2/14

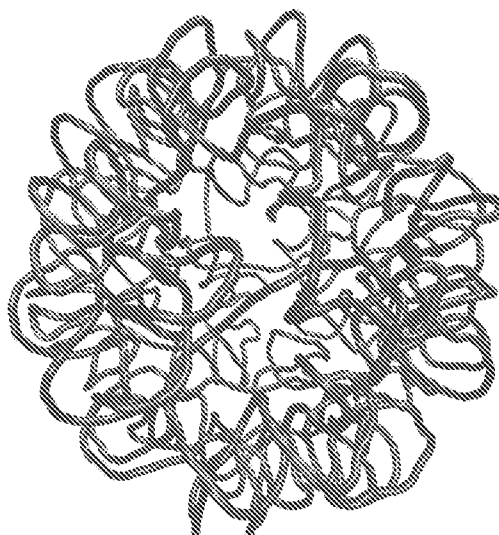


FIG. 2A

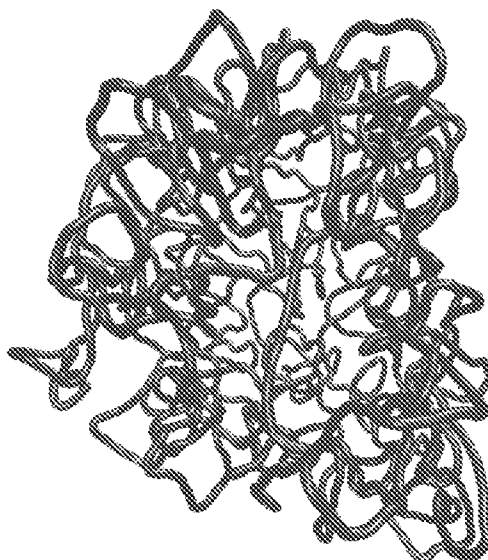


FIG. 2B

3/14

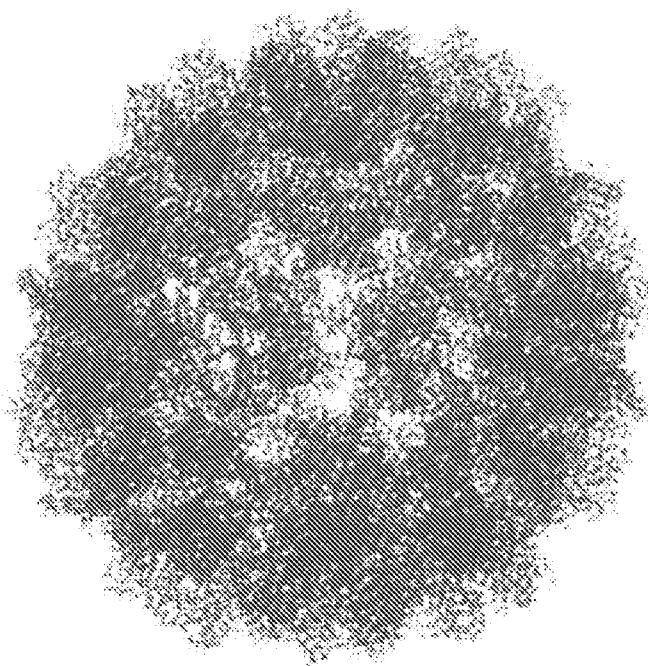


Fig. 3A

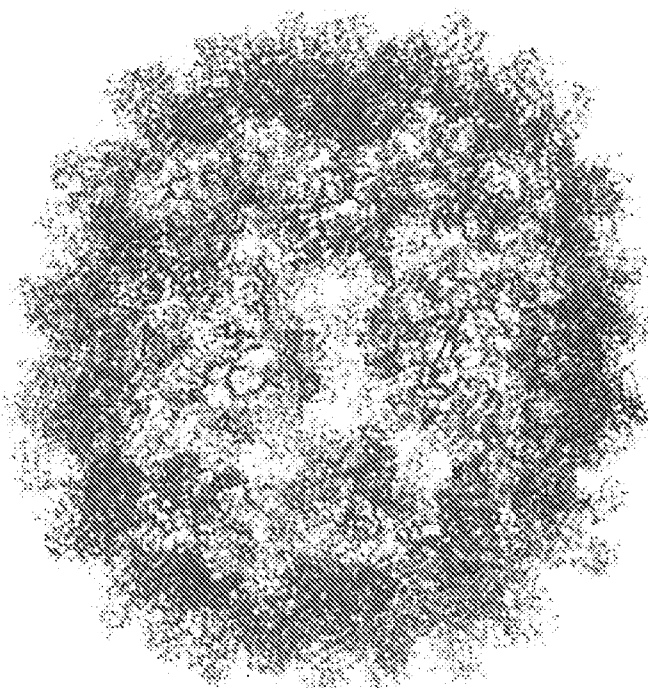


Fig. 3B

4/14

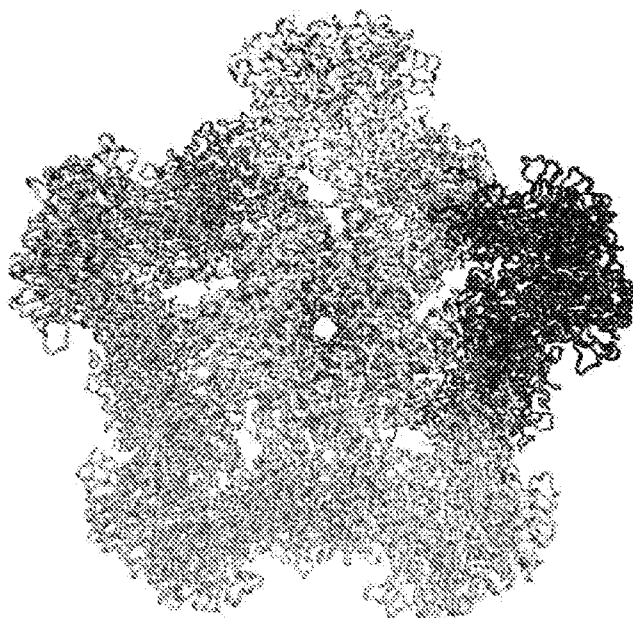


FIG. 4A

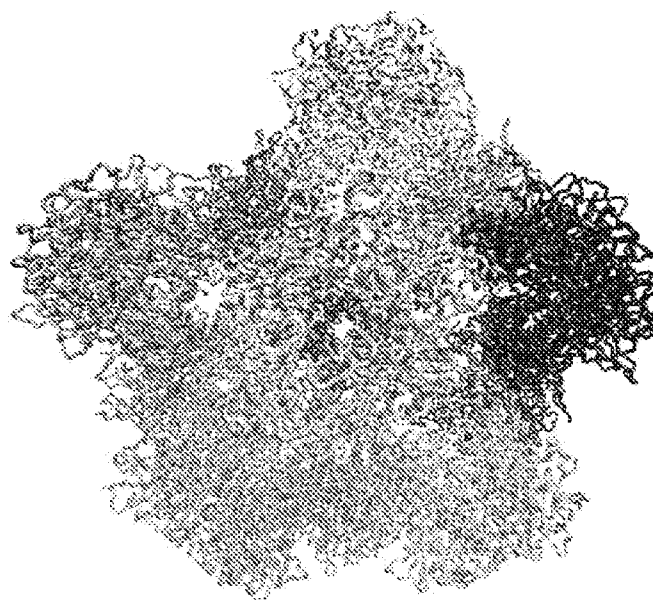


Fig. 4B

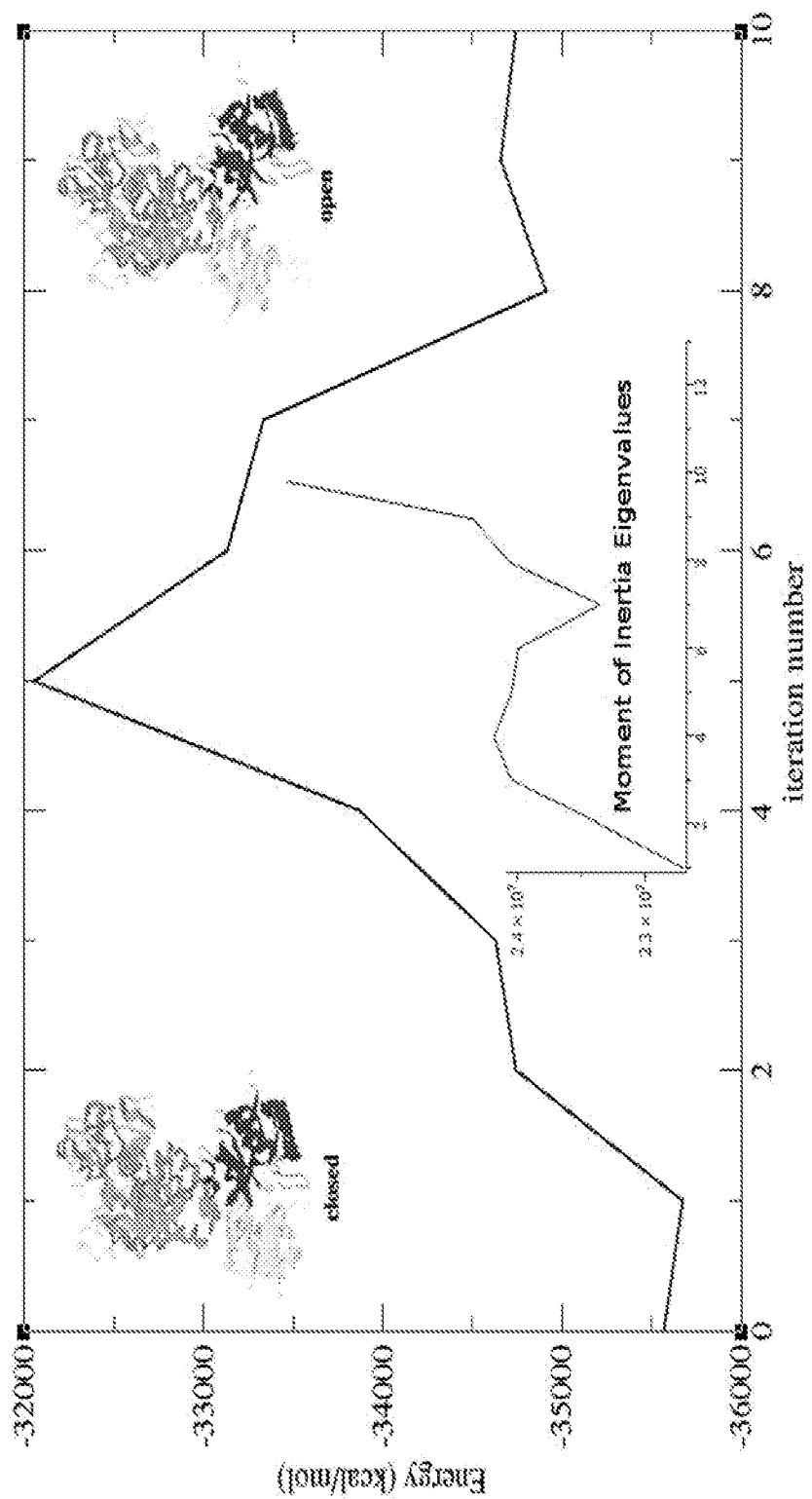


FIG. 5

6/14

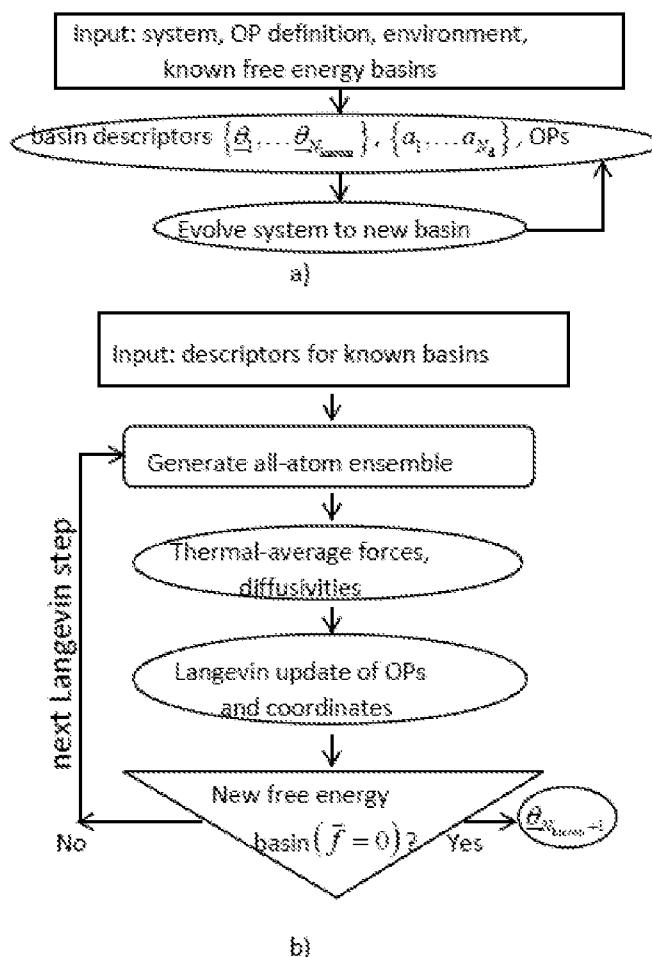


FIG. 6

7/14

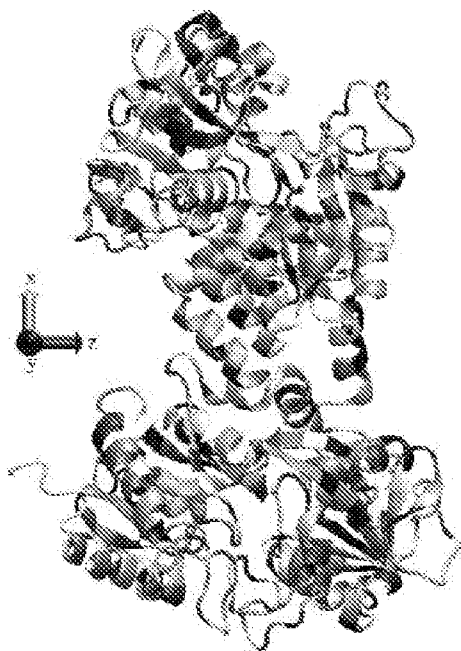


FIG. 7a

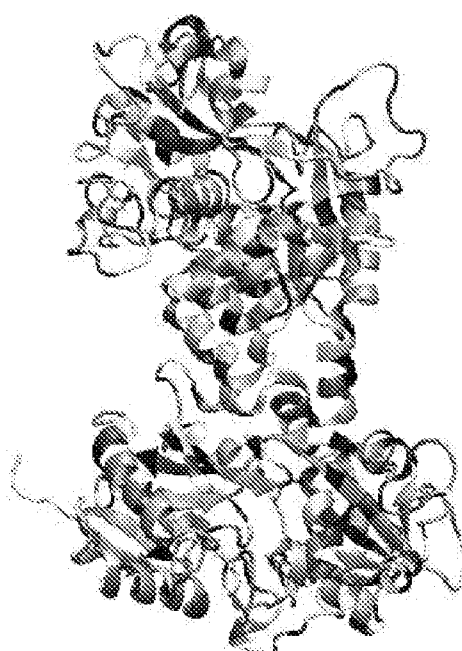


FIG. 7a'



FIG. 7b

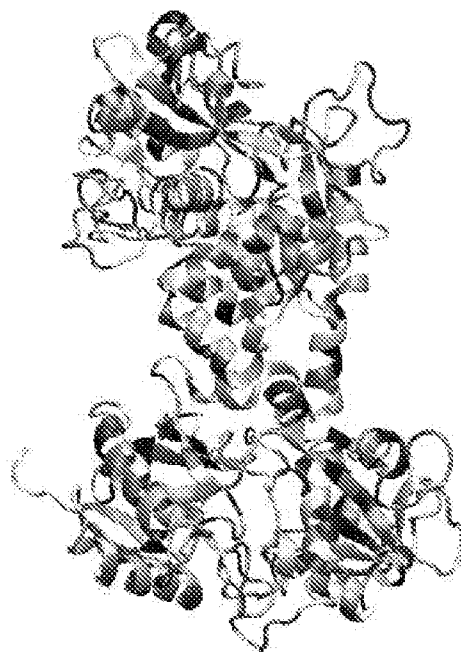


FIG. 7c

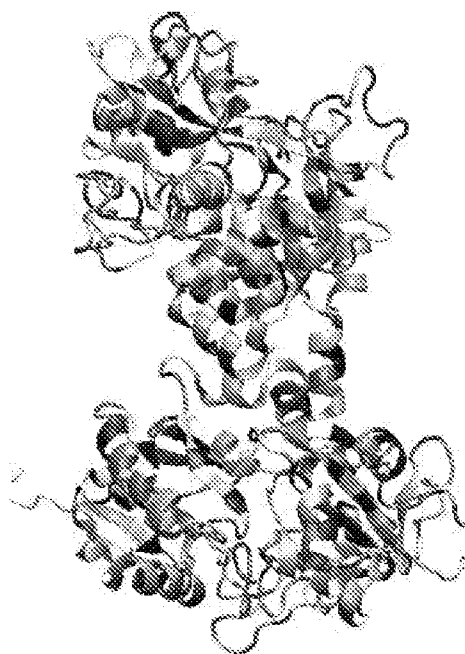


FIG. 7d

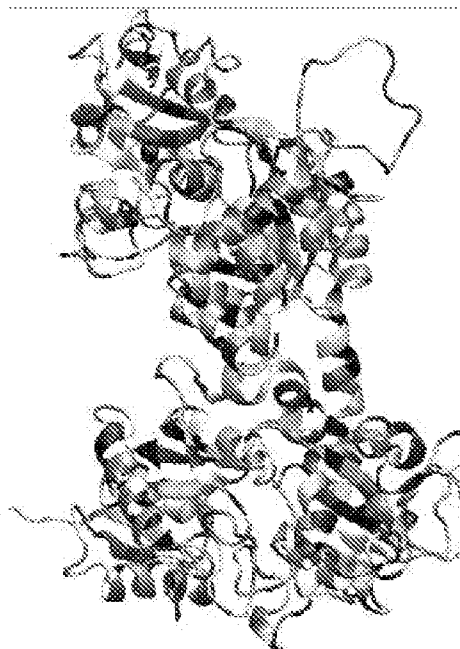


FIG. 7e

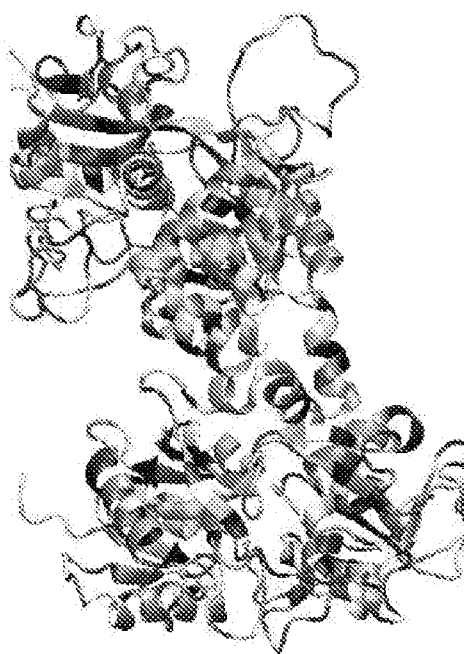


FIG. 7f

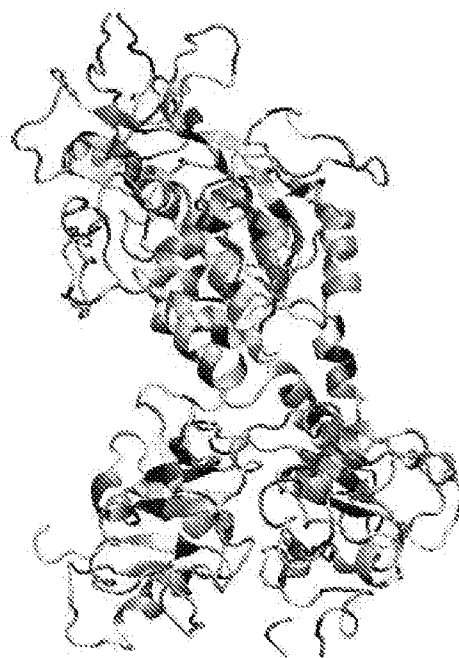


FIG. 7g

9/14

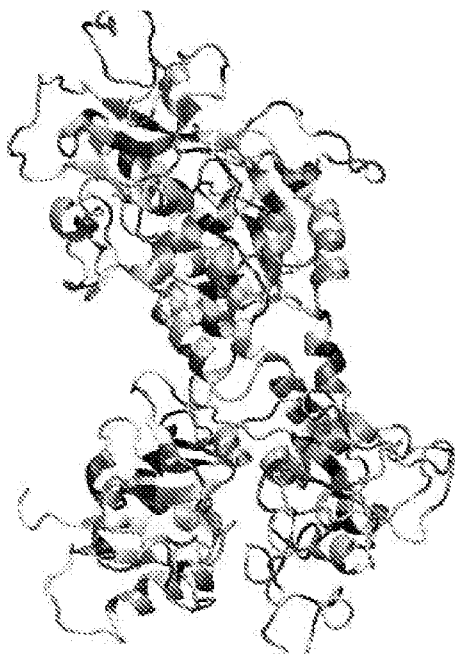


FIG. 7h

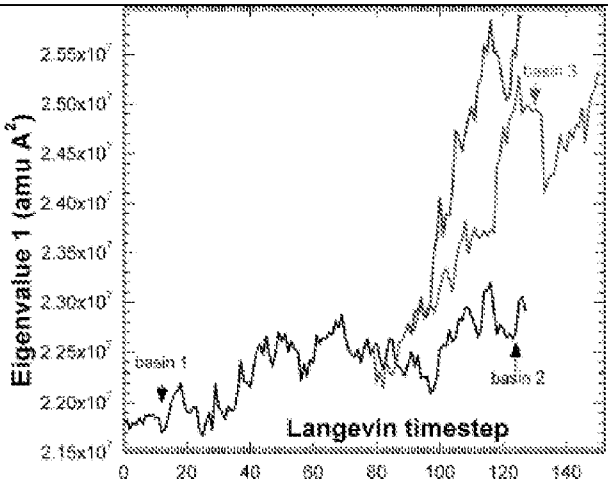


FIG. 8a

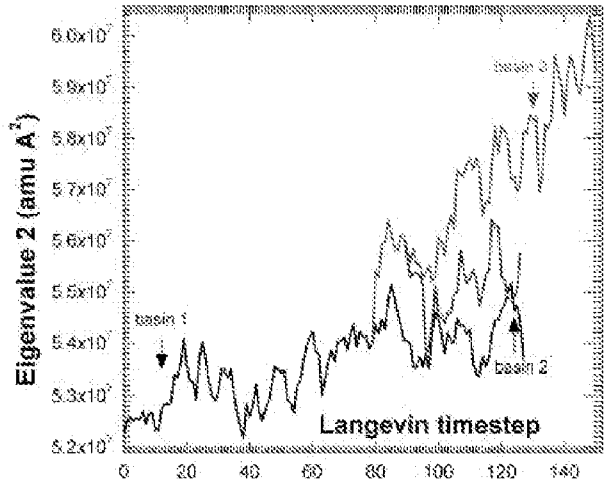


FIG. 8b

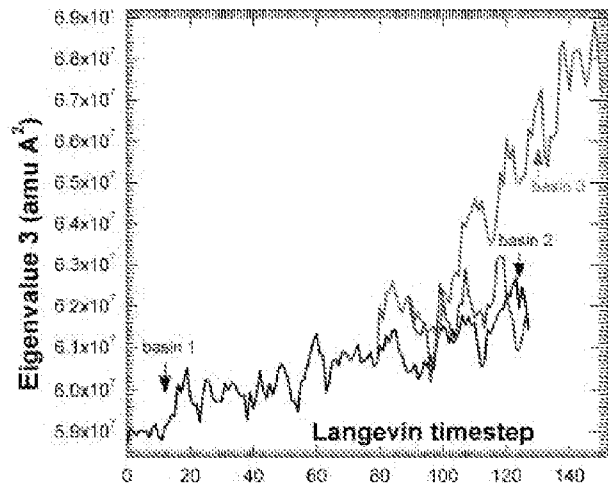


FIG. 8c

11/14

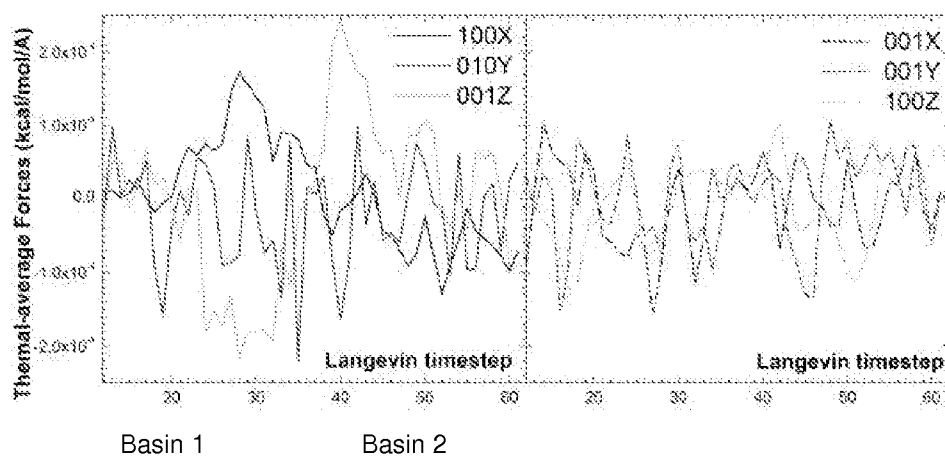


FIG. 9a

FIG. 9b

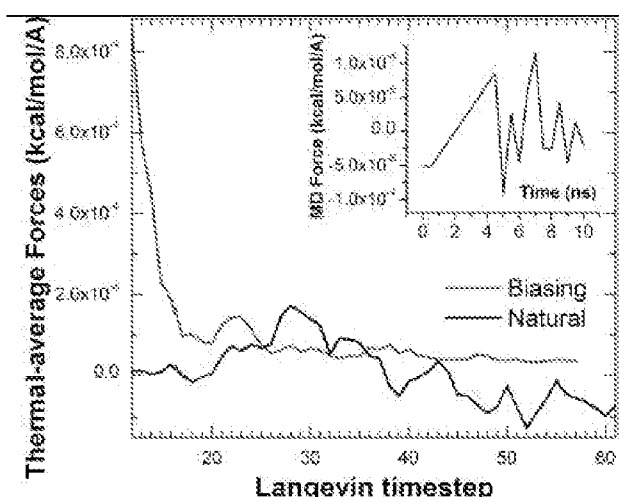


FIG. 9c

12/14

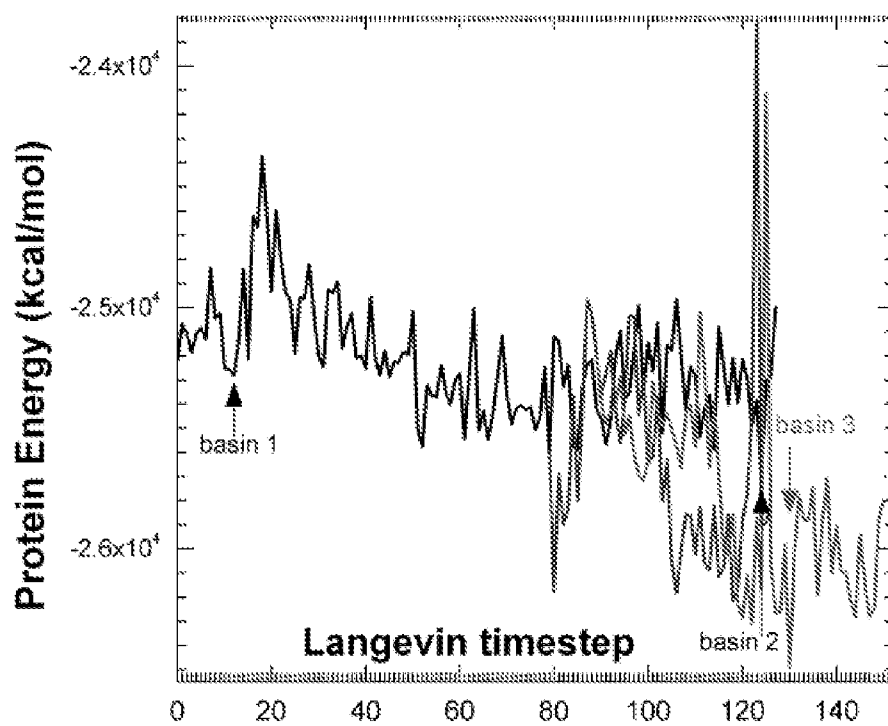


FIG. 10

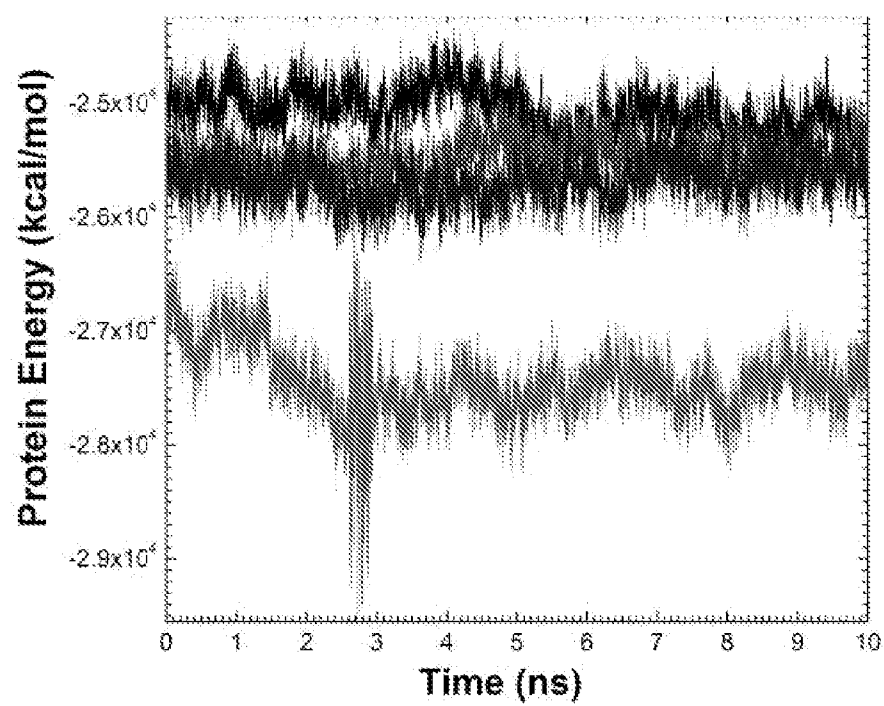


FIG. 11

13/14

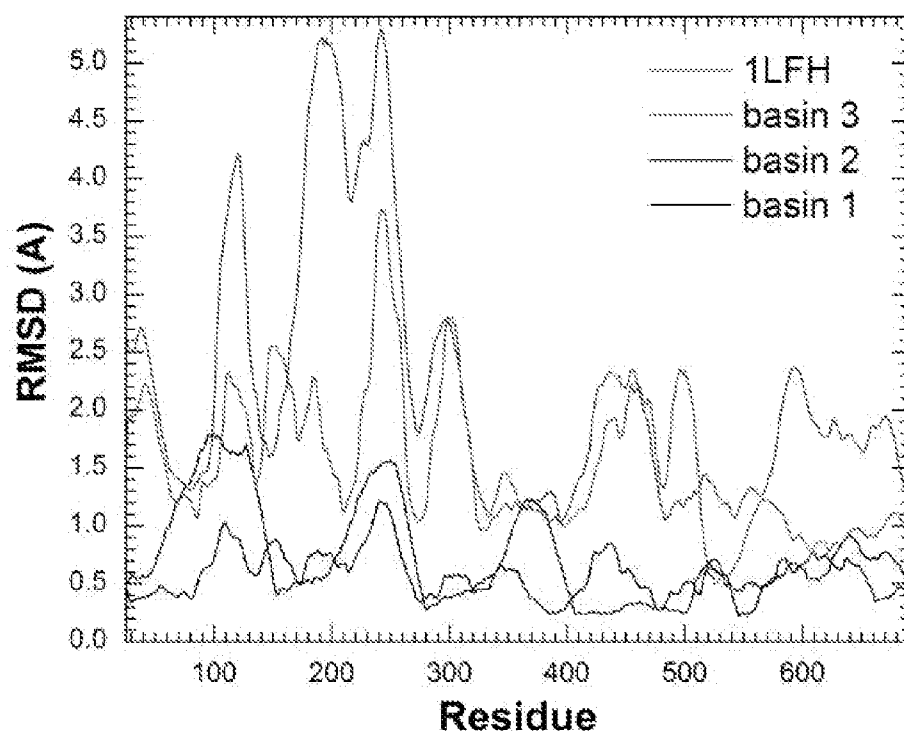


FIG. 12

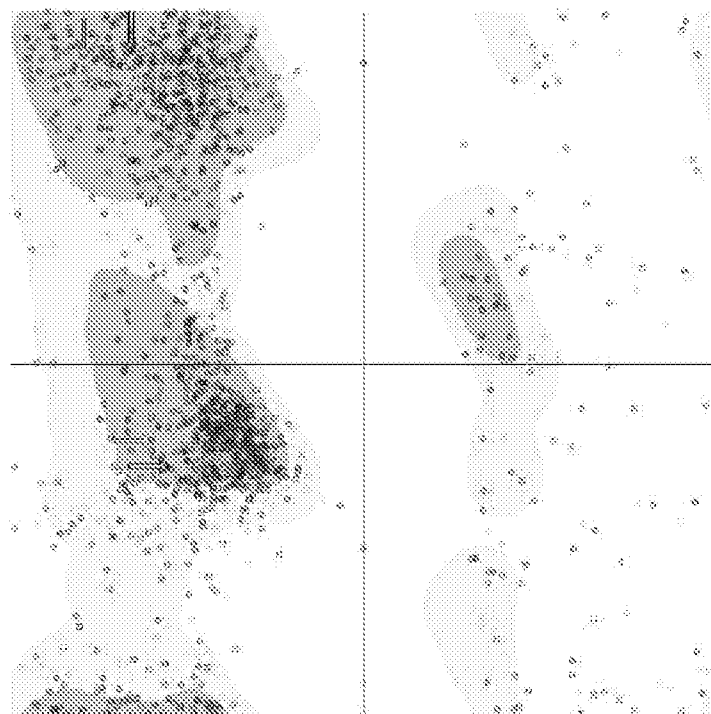


FIG. 13

14/14

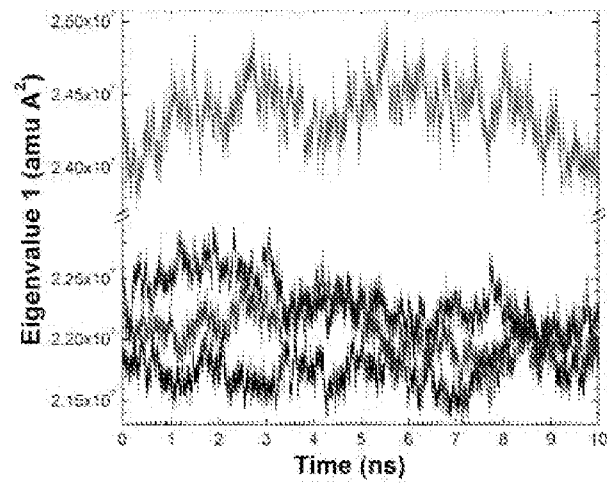


FIG. 14a

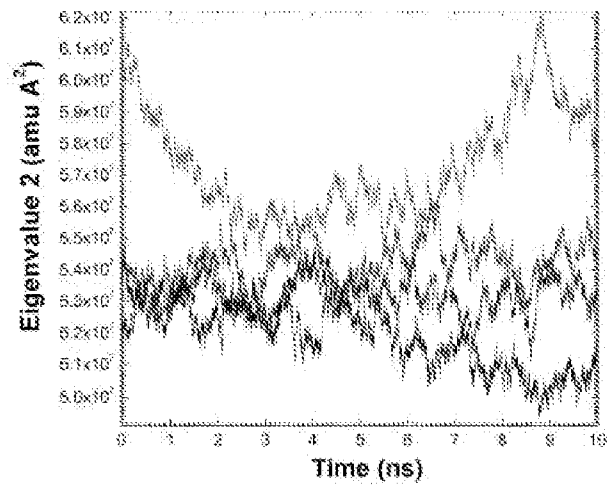


FIG. 14b

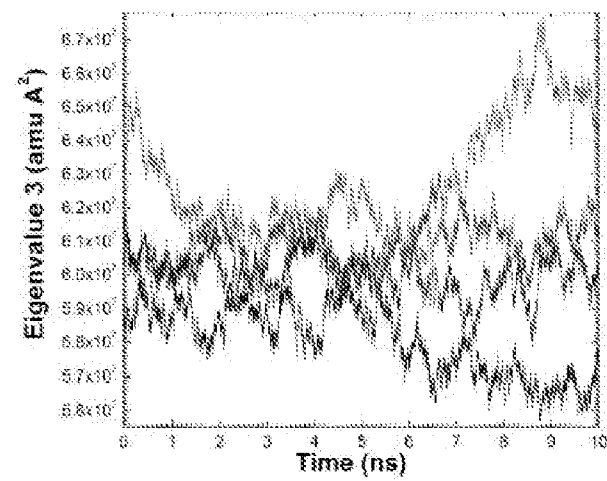


FIG. 14c