Learning dynamical information from static protein and sequencing data

Philip Pearce,[1] Francis G. Woodhouse,[2] Aden Forrow,[1] Ashley Kelly,[3] Halim Kusumaatmaja,[3] and Jörn Dunkel[1]

[1] *Department of Mathematics, Massachusetts Institute of Technology, Cambridge MA 02139-4307*
[2] *Department of Applied Mathematics and Theoretical Physics,*
Centre for Mathematical Sciences, University of Cambridge,
Wilberforce Road, Cambridge CB3 0WA, United Kingdom
[3] *Department of Physics, Durham University, Science Laboratories,*
South Road, Durham DH1 3LE, United Kingdom

(Dated: August 24, 2018)

Many complex processes, from protein folding and virus evolution to brain activity and neuronal network dynamics, can be described as stochastic exploration of a high-dimensional energy landscape. While efficient algorithms for cluster detection and data completion in high-dimensional spaces have been developed and applied over the last two decades, considerably less is known about the reliable inference of state transition dynamics in such settings. Here, we introduce a flexible and robust numerical framework to infer Markovian transition networks directly from time-independent data sampled from stationary equilibrium distributions. Our approach combines Gaussian mixture approximations and self-consistent dimensionality reduction with minimal-energy path estimation and multi-dimensional transition-state theory. We demonstrate the practical potential of the inference scheme by reconstructing the network dynamics for several protein folding transitions and HIV evolution pathways. The predicted network topologies and relative transition time scales agree well with direct estimates from time-dependent molecular dynamics data and phylogenetic trees. The underlying numerical protocol thus allows the recovery of relevant dynamical information from instantaneous ensemble measurements, effectively alleviating the need for time-dependent data in many situations. Owing to its generic structure, the framework introduced here will be applicable to modern cryo-electron-microscopy and high-throughput single-cell RNA sequencing data and can guide the design of new experimental approaches towards studying complex multiphase phenomena.

Energy landscapes encapsulate the effective dynamics of a wide variety of physical, biological and chemical systems[1,2]. Well-known examples include a myriad of biophysical processes[3–7], multiphase systems[2], thermally activated hopping in optical traps[8], chemical reactions[1], brain neuronal expression[9], and cellular development[10–14]. Energetic concepts have also been connected to machine learning[15] and to viral fitness landscapes, where pathways with the lowest energy barriers may explain typical mutational evolutionary trajectories of viruses between fitness peaks[16,17]. Recent advances in experimental techniques including cryo-electron microscopy (cryo-EM)[3] and single-cell RNA sequencing[18], as well as new online social interaction datasets[19], are producing an unprecedented wealth of high-dimensional instantaneous snapshots of biophysical and social systems. Although much progress has been made in dimensionality reduction[20–22] and the reconstruction of effective energy landscapes in these settings[3,11,14,23], the problem of inferring dynamical information such as protein folding or mutation pathways and rates from instantaneous ensemble data remains a major challenge.

To address this practically important question, we introduce here an integrated computational framework for identifying metastable states on reconstructed high-dimensional energy landscapes and for predicting the relative mean first passage times (MFPTs) between those states, without requiring explicitly time-dependent data. Our inference scheme employs an analytic representation of the data based on a Gaussian mixture model (GMM)[24] to enable efficient identification of minimum-energy transition pathways[25–27]. We show how the estimation of transition networks can be optimized by reducing the dimension of a high-dimensional landscape while preserving its topology. Our algorithm utilizes experimentally validated analytical results[8] for transition rates[1,28,29]. Thus, it is applicable whenever the time-evolution of the underlying system can be approximated by a Fokker-Planck-type Markovian dynamics, as is the case for a wide range of physical, chemical and biological processes[1].

Specifically, we illustrate the practical potential by inferring protein folding transitions and HIV evolution pathways. Current standard methods for coarse-graining the conformational dynamics of biophysical structures[30] typically estimate Markovian transition rates from time-dependent trajectory data in large-scale molecular dynamics simulations[31]. By contrast, we show here that protein folding pathways and rates can be recovered without explicit knowledge of the time-dependent trajectories, provided the system is sufficiently ergodic and equilibrium distributions are sampled accurately. The agreement with the trajectory-based estimates suggests that the inference of complex transition networks via reconstructed energy landscapes can provide a viable and often more efficient alternative to traditional time series estimates, particularly as new experimental techniques will offer unprecedented access to high-dimensional ensemble data.

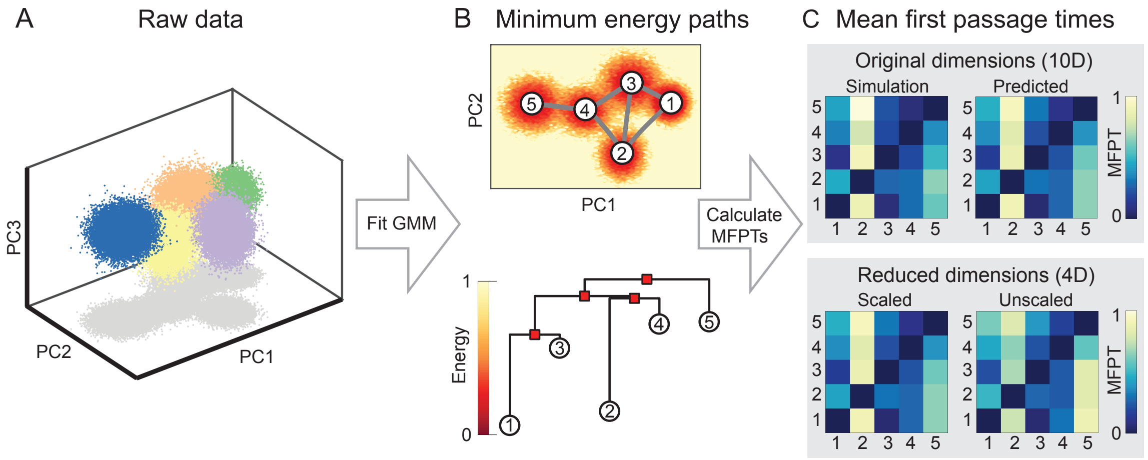

FIG. 1. Inference scheme for estimating transition networks and mean first passage times (MFPTs) from a stationary sample set, demonstrated on test data generated from a Gaussian Mixture Model (GMM; Supplementary Information). *(A)* Inputs are the instantaneously measured data, sampled here from a 10-dimensional GMM with 5 Gaussians, plotted in the first 3 principal components (PCs). *(B)* Top: A GMM is fit to the samples to construct the empirical distribution which is then converted to the energy landscape using Eq. (1). Background color indicates the projection of the empirical energy landscape onto the first two PCs. Minimum energy paths (MEPs, grey lines) between minima 1–5 on the landscape are calculated using the NEB algorithm (Supplementary Information). Bottom: Disconnectivity graph illustrating minima on the energy landscape (circles) and saddle points between them (squares). *(C)* A Markov state model (MSM) is constructed with transition rates given by Eq. (2) and solved to predict the MFPTs between discrete states (top right; Methods). MFPTs predicted by the MSM agree with direct estimates from Brownian dynamics simulations in the inferred energy landscape (top left; Supplementary Information). MFPTs calculated in a reduced 4-dimensional space using the scaling given in Eq. (3) recover the MFPTs accurately (bottom left). Without the appropriate scaling, the predicted MFPTs are inaccurate (bottom right).

RESULTS

Minimum-energy-path (MEP) network reconstruction

The equilibrium distribution $p\,(\mathbf{x})$ of a particle diffusing over a potential energy landscape $E(\mathbf{x})$ is the Boltzmann distribution $p(\mathbf{x}) = \exp\left[-E(\mathbf{x})/k_B T\right]/Z$, where k_B is the Boltzmann constant, T is the temperature and Z is a normalization constant. Given the probability density function (PDF) $p(\mathbf{x})$, the effective energy can be inferred from

$$E(\mathbf{x}) = -k_B T \ln[p(\mathbf{x})/p_{\max}], \qquad (1)$$

where $p_{\max}$ is the maximum value of the PDF, included to fix the minimum energy at zero. Our goal is to estimate the MFPTs between minima on the landscape using only sampled data. We divide this task into three steps, as illustrated in Fig. 1 for test data (Supplementary Information). In the first step, we approximate the empirical PDF by using the expectation maximization algorithm to fit a Gaussian mixture model (GMM) in a space of sufficiently large dimension d (Methods, Fig 1A). Mixtures with a bounded number of components can be recovered in time polynomial in both d and the required accuracy[32]. The resulting GMM yields an analytical expression for $E(\mathbf{x})$ via Eq. (1).

In the second step, the inferred energy landscape $E(\mathbf{x})$ is reduced to an MEP network whose nodes (states) are the minima of $E(\mathbf{x})$ (Fig. 1B top). Each edge represents an MEP that connects two adjacent minima and passes through an intermediate saddle point (Fig. 1B). The MEPs are found using the nudged elastic band (NEB) algorithm[25,26], which discretizes paths with a series of bead-spring segments (Supplementary Information).

Markov state model (MSM)

Given the MEP network, the final step is to infer the rates for transitioning from a minimum α to an adjacent minimum β. Assuming overdamped Brownian dynamics, the directed transition $\alpha \to \beta$ can be characterized by the generalized transition Kramers rate[1]

$$k_{\alpha\beta} = \frac{\omega_b}{2\pi\gamma} \frac{\prod_i \omega_i^\alpha}{\prod_i' \omega_i^S} \exp\left(-E_b\,/k_B T\,\right), \qquad (2)$$

where γ is the effective friction, E_b is the energy difference between the saddle point S on the MEP and the minimum α, ω_i^α are the stable angular frequencies at the minimum α, while ω_i^S and ω_b are the stable and unstable angular frequencies at the saddle. Eq. (2) assumes isotropic friction but can be generalized to a tensorial form[1] if anisotropies are relevant. In most practical applications, the error from assuming γ to be isotropic is likely negligible compared to other experimental noise sources. In principle, Eq. (2) can be refined further by including quartic (or higher) corrections to the prefactor ω_b/γ to account for details of the saddle shape[1]. Such corrections can be significant for GMMs (Supplementary Information).

Each edge $(\alpha\beta)$ has two weights, $k_{\alpha\beta}$ and $k_{\beta\alpha}$, assigned to it. The rate matrix $(k_{\alpha\beta})$ completely specifies the MSM on the network. Solving the MSM yields the matrix of pairwise mean first passage times (MFPTs) between states (Fig. 1C, Methods). In a simple two-state system, the MFPTs are determined up to a time scale by detailed balance, but for three or more states the influence of landscape topography and the associated state network topology (Methods) can lead to interesting hierarchical ordering of passage times. Identifying these hierarchies, and ways to manipulate them, is key to controlling protein folding or viral evolution pathways.

Topology-preserving dimensionality reduction

To ensure that the inference protocol can be efficiently applied to larger systems with a high-dimensional energy landscape, we derive a general method for reducing the dimension D of an energy landscape while preserving its topology. A probability density function with C well-separated Gaussians in D dimensions can be projected onto the $d = C - 1$ dimensional hyperplane spanning the Gaussian means using principal component analysis (PCA). In practice, it suffices to choose C to be larger than the number of energy minima if their number is not known in advance. Reduction to fewer than $d = C - 1$ dimensions does in general not allow a correct recovery of the MFPTs.

To preserve the topology under such a transformation – which is essential for the correct preservation of energy barriers and MEPs in the reduced-dimensional space – one needs to rescale GMM components in the low-dimensional space depending on the covariances of the Gaussians in the $D - d$ neglected dimensions (Fig. 1C). Explicitly, one finds that within the subspace spanned by the retained principal components (Supplementary Information)

$$p(\mathbf{x}_D) = \sum_{i=1}^{C} \phi_i \, p_i^d(\mathbf{x}_d) \frac{\sqrt{\det\left(2\pi \boldsymbol{U}_d^T \boldsymbol{\Sigma}_i \boldsymbol{U}_d\right)}}{\sqrt{\det\left(2\pi \boldsymbol{\Sigma}_i\right)}} \qquad (3)$$

as long as p satisfies certain minimally-restrictive conditions. Here, $\boldsymbol{U}_d$ denotes the first $d = C - 1$ columns of the matrix of sorted eigenvectors $\boldsymbol{U}$ of the covariance matrix of the Gaussian means, and ϕ_i, p_i^d and $\boldsymbol{\Sigma}_i$ are the mixing components, reduced-dimensional PDF and the covariance matrix of each individual Gaussian in the mixture, respectively (Supplementary Information). Neglecting the determinant scale-factors in Eq. (3), as is often done when GMM models are fitted to PCA-projected data, generally leads to inaccurate MFPT estimates (Fig. 1C, bottom). Note that Eq. (3) does not represent inversion of the transformation performed on the data by PCA, unless all D dimensions are retained; if some dimensions are neglected, Eq. (3) represents a rescaling of the marginal distribution in the retained dimensions to reconstruct the probability density function in the original dimension. In other words, the transition rates are best recovered from the conditional – not marginal – distributions, which are given by Eq. (3) up to a constant factor that does not affect energy differences.

Dimensionality reduction can substantially improve the efficiency of the NEB algorithm step: when the MEPs in the reduced d-dimensional space have been computed, the identified minima and saddles can be transformed back into the original data dimension D to calculate the Hessian matrices at these points, allowing Kramers' rates to be calculated as usual (Fig. 1C, Supplementary Information). Alternatively, in specific situations where the MEPs lie outside the hyperplane spanning the means (Supplementary Information), the MEP in the reduced d-dimensional space can be transformed back to the D-dimensional space and used as an initial condition in that space, significantly reducing computational cost. These results present a step towards a general protocol for identifying reaction coordinates or collective variables for projection of a high-dimensional landscape onto a reduced space while quantitatively preserving the topology of the landscape.

Protein folding

To illustrate the vast practical potential of the above scheme, we demonstrate the successful recovery of several protein folding pathways, using data from previous large-scale molecular dynamics (MD) simulations[31]. The protein trajectories, consisting of the time-dependent coordinates of the alpha carbon backbone, were pre-processed, treated as a set of static equilibrium measurements, and reduced in dimension before fitting a GMM (Methods). As is typical for high-dimensional parameter estimation with few structural assumptions, the fitting error due to a finite sample size n in d dimensions scales approximately as $\sqrt{d/n}$ (Supplementary Information); see Refs.[33,34] for advanced techniques tackling sample size limitations. Here, $d < 10$ so the sample size $n \sim 10^5$ suffices for effective recovery (Methods, Supplementary Information).

For each of the four analyzed proteins Villin, BBA, NTL9 and WW, the reconstructed energy landscapes reveal multiple states including a clear global minimum corresponding to the folded state (Fig. 2A,B). To estimate MFPTs, we determined the effective friction γ in Eq. (2) for each protein from the condition that the line of best fit through the predicted vs. measured MFPTs has unit gradient. Although not usually known, γ could in principle be calculated by comparing MD simulations with experimental data. Our MFPT predictions agree well with direct estimates (Supplementary Information) from

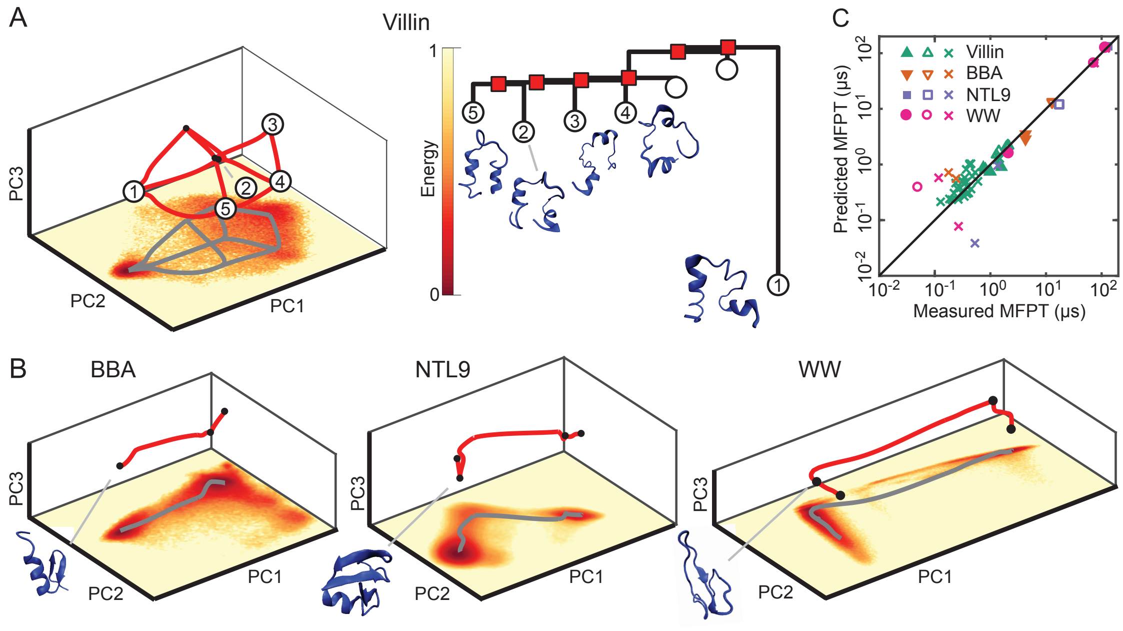

FIG. 2. Reconstructed MEP networks for protein folding transitions, and comparison of predicted MFPTs with direct estimates from molecular dynamics (MD) simulations (Supplementary Information). *(A)* Left: Low energy states and transition network in the first three principal components (PCs) for Villin including predicted transition paths between states (red lines); bottom coloring shows two-dimensional projection of the empirical energy landscape onto the first two PCs. Right: Associated disconnectivity graph and illustrations of the five lowest energy states, with state 1 corresponding to the folded state. *(B)* Low energy states, transition paths and empirical energy landscape for BBA, WW and NTL9 proteins, and sketches of their folded states. *(C)* Predicted MFPTs agree well with estimates from MD simulations when energy minima are well separated and become less accurate for fast transitions with small MFPTs. Filled shapes correspond to unfolding transitions and unfilled shapes correspond to folding transitions for Villin, BBA, NTL9 and WW. Crosses correspond to transitions between intermediate states.

the time-dependent MD trajectories (Fig 2C). Detailed analysis confirms that the MFPT estimates are robust under variations of the number of Gaussians used in the mixture (Fig. S1). Also, the estimated MEPs are in good agreement with the typical transition paths observed in the MD trajectories (Fig. S2).

Viral evolution

As a second proof-of-concept application, we demonstrate that our inference scheme recovers the expected evolution pathways between HIV sequences as well as the key features of a distance-based phylogenetic tree (Fig. 3). To this end, we reconstructed an effective energy landscape from publicly available HIV sequences sampled longitudinally at several points in time from multiple patients[35], assuming that the frequency of an observed genotype is proportional to its probability of fixation and that the high-dimensional discrete sequence space can be projected onto a continuous reduced-dimensional phenotype space (Fig. 3A; Supplementary Information). First, a Gaussian was fit to each patient and then combined in a GMM with equal weights, to avoid bias in the fitness landscape towards sequences infecting any specific patient (Supplementary Information). Thereafter, we applied our inference protocol to reconstruct the effective energy landscape, transition network (Fig. 3B) and disconnectivity graph (Fig. 3C), where each state is associated to a separate patient. As expected, states corresponding to patients infected with different HIV subtypes are not connected by MEPs (Fig 3A,B). The disconnectivity graph reproduces the key features of a coarse-grained patient-level representation of the phylogenetic tree (Fig. 3C). Using our inference scheme, vertical evolution in the tree can be tracked along the minimum energy paths in a reduced-dimensional sequence space (Fig. 3B). The energy barriers, represented by the lengths of the vertical lines in the disconnectivity graph (Fig. 3C), provide an estimate for the relative likelihood of evolution to fixation via point mutations between fitness peaks (energy minima). If mutation rates are known, the MEPs can also be used to estimate the time for evolution to fixation from one fitness peak to another[36].

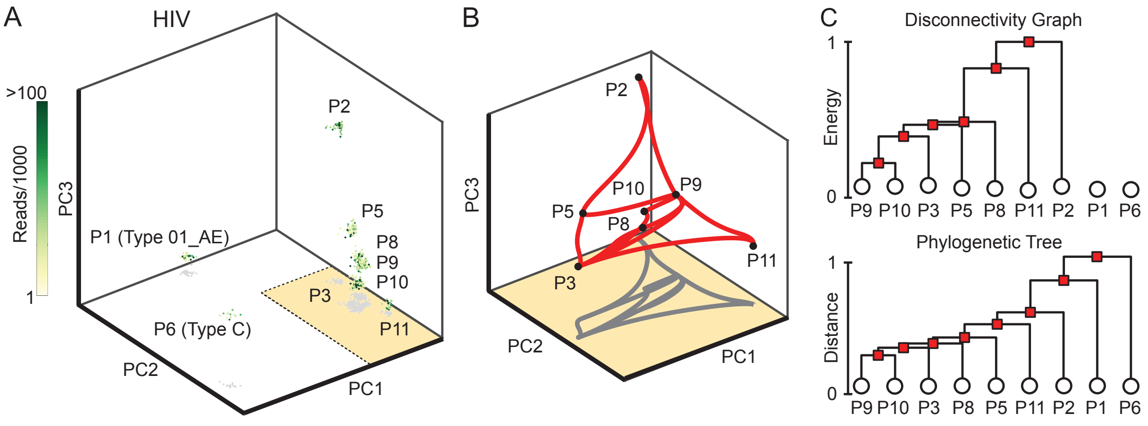

FIG. 3. Minimum energy paths (MEPs) on viral fitness landscapes reconstructed from publicly available HIV sequencing data[35]. *(A)* Longitudinal samples of the HIV virus are binarized after multiple sequence alignment (Supplementary Information) and plotted in the first 3 PCs. Samples of the same HIV subtype are closer in PC-space. Patient labels correspond to those used in[35]. *(B)* MEPs between minima corresponding to patients infected with Type B HIV, plotted in the first 3 PCs. Paths between minima indicate likely evolutionary pathways. Minima corresponding to patients with Type 01_AE and Type C HIV were unconnected to the other minima. *(C)* Disconnectivity graph for connected minima, where vertical evolution frequency is assumed to be proportional to the normalized energy barriers (top). The disconnectivity graph reproduces the majority of the structure of a distance-based phylogenetic tree (bottom), where the lengths of vertical lines are proportional to the Jukes-Cantor sequence distance (scaled to $[0, 1]$).

DISCUSSION

Preserving landscape topology under dimensionality reduction

Finding the appropriate number of collective macrovariables to describe an energy landscape is a generic problem relevant to many fields. For example, although some proteins can be described through effective one-dimensional reaction coordinates[5], the accurate description of their diffusive dynamics over the full microscopic energy landscape requires many degrees of freedom[37]. Whenever dynamics are inherently high-dimensional, topology-preserving dimensionality reduction can enable a much faster search of the energy landscape for minima and MEPs. In practice, data dimension is often reduced with PCA or similar methods before constructing an energy landscape[37,38]. The extent to which commonly used dimensionality reduction techniques alter MEP network topology or quantitatively preserve energy barriers is not well understood. Eq. (3) suggests that reducing dimensions using PCA should not introduce significant errors if the variance of the landscape around each state (energy minimum) in the neglected dimensions is similar. For instance, we found that the protein folding data could be reduced to five dimensions while maintaining accuracy (Fig. S1), although additional higher energy states may become evident in higher dimensions. Overall, our theoretical results demonstrate the benefits of combining an analytical PDF with a linear dimensionality reduction technique so that the neglected dimensions can be accounted for explicitly.

Biological and biophysical applications

Rapidly advancing imaging techniques, such as cryogenic electron microscopy (cryo-EM), will allow many snapshots of biophysical structures to be taken at the atomic level in the near future[3]. A biologically and biophysically important task will be to infer dynamical information from such instantaneous static ensemble measurements. The protein folding example in Fig. 2 suggests that the framework introduced here can help overcome this major challenge. Another promising area of future application is the analysis of single-cell RNA-sequencing data quantifying the expression within individual cells[18]. In related recent work, an effective energy landscape of single-cell expression snapshots was inferred using the Laplacian of a k-nearest neighbor graph on the data, allowing lineage information to be derived via a Markov chain[13]. The GMM-based framework here provides a complementary approach for reconstructing faithful low-dimensional transition state dynamics from such high-dimensional data.

Furthermore, the proof-of-concept results in Fig. 3 suggests that our inference scheme for Markovian network dynamics can be useful for studying viral and bacterial evolution, which are often modeled as movements through a series of DNA or protein sequences[39]. The fitness landscape of an organism in sequence space is analogous to the negative of an effective energy landscape. The process of fixation by a succession of mutants in a population, whereby each mutant replaces the previous lineage as the population's most recent common ancestor, has been modeled as a Markov process[40]. Successive sweeps to fixation have been observed in long-term evolution experiments, promising groundbreaking data for future analysis as whole-genome sequencing technologies improve[41].

Outlook and extensions

The inference protocol opens the possibility to analyze previously intractable multi-phase systems: many high-dimensional physical, chemical and other stochastic processes can be described by a Fokker-Planck dynamics[1], with phase equilibria corresponding to maxima of the stationary distribution. By taking near-simultaneous measurements of many subsystems within a large multistable Fokker-Planck system, the above scheme allows the inference of coexisting equilibria and transition rates between them. Other possible applications may include neuronal expression[9] and social networks[19], which have been described in terms of effective energy landscapes.

While we focused here on normal white-noise diffusive behavior, as is typical of protein folding dynamics, the above ideas can in principle be generalized to other classes of stochastic exploration processes. Such extensions will require replacing Eq. (2) through suitable generalized rate formulas, as have been derived for correlated noise[1,42]. Conversely, the present framework provides a means to test for diffusive dynamics: if the MFPTs of an observed system differ markedly from those inferred by the above protocol, then either important degrees of freedom have not been measured; the system is out of equilibrium on measurement time scales; or the system does not have Brownian transition statistics, necessitating further careful investigation of its time dependence.

To conclude, the conformational dynamics of biophysical structures such as viruses and proteins are characterized by their metastable states and associated transition networks, and can often be captured through Markovian models. Current experimental techniques, such as cryo-EM or RNA-sequencing, provide limited dynamical information. In these cases, transition networks must be inferred from structural snapshots. Here, we have introduced a numerical framework for inferring Markovian state-transition networks via reconstructed energy landscapes from high-dimensional static data. The successful application to protein folding and viral evolution pathways illustrates that high-dimensional energy landscapes can be reduced in dimension without losing relevant topological information. Generally, the inference scheme presented here is applicable whenever the dynamics of a high-dimensional physical, biological or social system can be approximated by diffusion in an effective energy landscape.

METHODS

Population landscapes. A Gaussian mixture model (GMM) was used to represent the probability density function (PDF), or population landscape, of samples. The PDF at position $\mathbf{x}$ of a GMM with C mixture components in d dimensions is

$$p(\mathbf{x}) = \sum_{i=1}^{C} \phi_i p_i(\mathbf{x})$$

$$p_i(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)^T \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_i\right)\right)}{\sqrt{\det\left(2\pi\boldsymbol{\Sigma}_i\right)}},$$

where ϕ_i are the weights of each component, $\boldsymbol{\mu}_i$ are the means and $\boldsymbol{\Sigma}_i$ are the covariance matrices. More details on GMMs and how they were fit to data is given in the Supplementary Information.

Mean first passage times. We form a discrete-state continuous-time Markov chain on states given by the minima of the energy landscape. For a pair of states α and β directly connected by a minimum-energy pathway via a saddle, we approximate the transition rate $\alpha \to \beta$ by the Kramers rate $k_{\alpha\beta}$ in Eq. (2), while if α and β are not directly connected we set $k_{\alpha\beta} = 0$. Given these rates, the Markov chain has generator matrix $M_{\alpha\beta}$ where $M_{\alpha\beta} = k_{\alpha\beta}$ for $\alpha \neq \beta$ and $M_{\alpha\alpha} = -\sum_{\beta:\beta\neq\alpha} k_{\alpha\beta}$. Then the matrix $\tau_{\alpha\beta}$ of MFPTs (hitting times) for transitions $\alpha \to \beta$ satisfies

$$\sum_{\gamma} M_{\alpha\gamma}\tau_{\gamma\beta} = -1 \,\text{for}\, \alpha \neq \beta, \quad \tau_{\alpha\alpha} = 0.$$

Protein data pre-processing. Protein folding trajectories were obtained from all-atom molecular dynamics (MD) simulations performed by D.E. Shaw Research[31]. Data was subsampled by a factor of 5 to reduce the size. For some proteins, residues at the flexible tails of proteins were removed from the dataset to reduce noise. Pairwise distances between carbon alpha atoms on the protein backbone were taken, with a cut off of 6-8 Å, depending on the size of the protein. Samples were reduced in dimension using principal component analysis (PCA). The first five principle components of the protein data were found to be sufficient for inference of energy landscapes and transition networks (Fig. S1).

Code availability. The source code used in this study to learn a dynamical transition network and mean first passage times from a Gaussian mixture model is publicly available from Github (https://github.com/philip-pearce/learning-dynamical). Also included are all data processing codes required to convert the raw data used in this study into the appropriate format.

Data availability. Two publicly available datasets were used in this study. Protein folding trajectories[31] are available from D.E. Shaw Research (https://www.deshawresearch.com/). HIV sequences[35] are available from https://hiv.biozentrum.unibas.ch/.

ACKNOWLEDGMENTS

We thank D.E. Shaw Research for protein folding trajectories and Stefano Piana-Agostinetti of D.E. Shaw Research for helpful discussions. This work was supported by the Royal Society International Exchanges award IE160909 (H.K. and J.D.) and Complex Systems Scholar Award from the James S. McDonnell Foundation (J.D.).

[1] Hänggi, P., Talkner, P. & Borkovec, M. Reaction-rate theory: Fifty years after Kramers. *Rev Mod Phys* **62**, 251 (1990).

[2] Yukalov, V. Phase transitions and heterophase fluctuations. *Phys Rep* **208**, 395–489 (1991).

[3] Dashti, A. *et al.* Trajectories of the ribosome as a Brownian nanomachine. *Proc Natl Acad Sci USA* **111**, 17492–17497 (2014).

[4] Chung, H. S., Piana-Agostinetti, S., Shaw, D. E. & Eaton, W. A. Structural origin of slow diffusion in protein folding. *Science* **349**, 1504–1510 (2015).

[5] Neupane, K., Manuel, A. P. & Woodside, M. T. Protein folding trajectories can be described quantitatively by one-dimensional diffusion over measured energy landscapes. *Nat Phys* **12**, 700–703 (2016).

[6] Hosseinizadeh, A. *et al.* Conformational landscape of a virus by single-particle X-ray scattering. *Nat Methods* **14**, 877–881 (2017).

[7] Best, R. B. & Hummer, G. Diffusive model of protein folding dynamics with Kramers turnover in rate. *Phys Rev Lett* **96**, 228104 (2006).

[8] Rondin, L. *et al.* Direct Measurement of Kramers Turnover with a Levitated Nanoparticle. *Nat Nanotechnol* **12**, 1130–1133 (2017).

[9] Ezaki, T., Watanabe, T., Ohzeki, M. & Masuda, N. Energy landscape analysis of neuroimaging data. *Phil Trans R Soc A* **375**, 20160287 (2016).

[10] Corson, F. & Siggia, E. D. Geometry, epistasis, and developmental patterning. *Proc Natl Acad Sci USA* **109**, 5568–5575 (2012).

[11] Lang, A. H., Li, H., Collins, J. J. & Mehta, P. Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. *PLOS Comput Biol* **10**, e1003734 (2014).

[12] Pusuluri, S. T., Lang, A. H., Mehta, P. & Castillo, H. E. Cellular reprogramming dynamics follow a simple 1D reaction coordinate. *Phys Biol* **15**, 016001 (2017).

[13] Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M. & Klein, A. M. Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci USA* (2018).

[14] Jin, S., MacLean, A. L., Peng, T. & Nie, Q. scEpath: Energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. *Bioinformatics* **34**, 2077–2086 (2018).

[15] Ballard, A. J. *et al.* Energy landscapes for machine learning. *Phys Chem Chem Phys* **19**, 12585–12603 (2017).

[16] Ferguson, A. L. *et al.* Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design. *Immunity* **38**, 606–617 (2013).

[17] Ebeling, W. & Feistel, R. Studies on Manfred Eigen's model for the self-organization of information processing. *Eur Biophys J* 395–401 (2018).

[18] Shalek, A. K. *et al.* Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* **498**, 236–240 (2013).

[19] Marvel, S. A., Strogatz, S. H. & Kleinberg, J. M. Energy landscape of social balance. *Phys Rev Lett* **103**, 198701 (2009).

[20] Stephens, G. J., Osborne, L. C. & Bialek, W. Searching for simplicity in the analysis of neurons and behavior. *Proc Natl Acad Sci USA* **108**, 15565–15571 (2011).

[21] Wasserman, L. Topological data analysis. *Annu Rev Stat Appl* **5**, 501–532 (2018).

[22] Mattingly, H. H., Transtrum, M. K., Abbott, M. C. & Machta, B. B. Maximizing the information learned from finite data selects a simple model. *Proc Natl Acad Sci USA* **115**, 1760–1765 (2018).

[23] Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods* **100**, 61–67 (2016).

[24] Westerlund, A. M., Harpole, T. J., Blau, C. & Delemotte, L. Inference of Calmodulin's Ca^{2+}-Dependent Free Energy Landscapes via Gaussian Mixture Model Validation. *J Chem Theory Comput* **14**, 63–71 (2018).

[25] Jónsson, H., Mills, G. & Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, 385–404 (World Scientific, 1998).

[26] Trygubenko, S. A. & Wales, D. J. A doubly nudged elastic band method for finding transition states. *J Chem Phys* **120**, 2082–2094 (2004).

[27] Kusumaatmaja, H. Surveying the free energy landscapes of continuum models: Application to soft matter systems. *J Chem Phys* **142**, 124112 (2015).

[28] Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304 (1940).

[29] Dunkel, J., Ebeling, W., Schimansky-Geier, L. & Hänggi, P. Kramers problem in evolutionary strategies. *Phys Rev E* **67**, 061118 (2003).

[30] Mardt, A., Pasquali, L., Wu, H. & Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat Commun* **9**, 5 (2018).

[31] Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).

[32] Kalai, A. T., Moitra, A. & Valiant, G. Disentangling Gaussians. *Commun ACM* **55**, 113–120 (2010).

[33] Bühlmann, P., Kalisch, M. & Meier, L. High-Dimensional Statistics with a View Toward Applications in Biology. *Annu Rev Stat Appl* **1**, 255–278 (2014).

[34] Brenner, M. P., Colwell, L. J. *et al.* Optimal design of experiments by combining coarse and fine measurements. *Phys Rev Lett* **119**, 208101 (2017).

[35] Zanini, F. *et al.* Population genomics of intrapatient HIV-1 evolution. *Elife* **4**, e11282 (2015).

[36] Gokhale, C. S., Iwasa, Y., Nowak, M. A. & Traulsen, A.

The pace of evolution across fitness valleys. *J Theor Biol* **259**, 613–620 (2009).

[37] Ferguson, A. L., Panagiotopoulos, A. Z., Kevrekidis, I. G. & Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem Phys Lett* **509**, 1–11 (2011).

[38] Ernst, M., Sittel, F. & Stock, G. Contact- and distance-based principal component analysis of protein dynamics. *J Chem Phys* **143**, 244114 (2015).

[39] Orr, H. A. Fitness and its role in evolutionary genetics. *Nat Rev Genet* **10**, 531–539 (2009).

[40] Sella, G. & Hirsh, A. E. The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* **102**, 9541–9546 (2005).

[41] Barrick, J. E. & Lenski, R. E. Genome dynamics during experimental evolution. *Nat Rev Genet* **14**, 827–839 (2013).

[42] Sharma, A., Wittmann, R. & Brader, J. M. Escape rate of active particles in the effective equilibrium approach. *Phys Rev E* **95**, 012115 (2017).