Toward a direct and scalable identification of reduced models for categorical processes

Susanne Gerber[a,b] and Illia Horenko[c,1]

[a]Institute for Developmental Biology and Neurobiology, Faculty of Biology, Johannes-Gutenberg Universität Mainz, 55128 Mainz, Germany; [b]Center for Computational Sciences in Mainz (CSM), Johannes-Gutenberg Universität Mainz, 55128 Mainz, Germany; and [c]Institute of Computational Science, Faculty of Informatics, Università della Svizzera Italiana, 6900 Lugano, Switzerland

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved March 29, 2017 (received for review August 9, 2016)

The applicability of many computational approaches is dwelling on the identification of reduced models defined on a small set of collective variables (colvars). A methodology for scalable probability-preserving identification of reduced models and colvars directly from the data is derived—not relying on the availability of the full relation matrices at any stage of the resulting algorithm, allowing for a robust quantification of reduced model uncertainty and allowing us to impose a priori available physical information. We show two applications of the methodology: (*i*) to obtain a reduced dynamical model for a polypeptide dynamics in water and (*ii*) to identify diagnostic rules from a standard breast cancer dataset. For the first example, we show that the obtained reduced dynamical model can reproduce the full statistics of spatial molecular configurations—opening possibilities for a robust dimension and model reduction in molecular dynamics. For the breast cancer data, this methodology identifies a very simple diagnostics rule—free of any tuning parameters and exhibiting the same performance quality as the state of the art machine-learning applications with multiple tuning parameters reported for this problem.

dimension reduction | Markov state models | clustering | computer-aided diagnostics | Bayesian modeling

Model reduction and identification of a most appropriate (small) set of collective variables are essential prerequisites for many computational methods and modeling techniques in a number of applied disciplines ranging from biophysics and bioinformatics to computational medicine and image processing. A variety of methods for the identification of collective variables can be roughly subdivided into two major groups: (*i*) methods that are based on some user-defined agglomeration of the original degrees of freedom into collective variables (e.g., based on the physical intuition) (1) and (*ii*) methods that produce/derive these agglomerations of original system's variables based on a reduced approximation of some system-specific relation matrices. These matrices can be defined, for example, as covariance or kernel covariance matrices (2, 3), partial autocorrelation matrices of autoregressive processes (4), Gaussian distance kernel matrices (5, 6), Laplacian matrices [as in the case of spectral clustering methods for graphs (7, 8)], adjacency matrices [in community identification methods for networks (9)], or Markov transition matrices [as in spectral reduction methods for Markov processes (10, 11)]. In most of these reduction methods, the relation matrices are assumed a priori available—and this assumption is true, for example, in social sciences, network science, and many areas of biology. However, in many particular applications (e.g., in biophysics and many medical applications; examples 1 and 2 below), one first needs to estimate these matrices from available data. For systems with a large number of dimensions (for continuous data) or categories (for categorical data) and short available statistics, these matrix estimates will be subject to uncertainty and may lead to biasedness of the derived colvars. Some other reduction approaches that allow for computing the reduced representation from the data directly [e.g., the Probabilistic Latent Semantic Analysis (PLSA; used in mathematical linguistics and information retrieval for analysis and reduction of texts and documents)] (12–14) impose strong assumptions on the data and exhibit issues related to the computational cost scaling (Fig. 1 and *SI Appendix*, section S5 have detailed discussion), making them practically not applicable to nonsparse data in such areas as, for example, the model and data reduction in biophysics and bioinformatics. Another problem arises when trying to identify the colvars for dynamical systems while simultaneously trying to preserve some essential conservation properties (e.g., conservation of energy or probability) in the reduced representation. For example, deploying spectral methods based on Euclidean eigenvector projections [such as principal component analysis (PCA) and spectral clustering methods] to reduction of probability measures would not guarantee that the components of the projected/reduced representation will also add up to one and all be bigger than or equal to zero (i.e., the resulting reduced models may not be probability preserving).

In this paper, we present an algorithmic framework that is scalable for realistic dynamical systems and is designed for the inference and analytically computable uncertainty quantification of reduced probability-preserving Bayesian relation models directly from the data.

Methodology

Below, we will give a brief description of the methodology—detailed derivation can be found in *SI Appendix*, section S1. Our aim is to come out with a reduction method intending to preserve causality relations—measured in terms of the matrix of conditional probabilities between two categorical processes Y and X. Process Y will serve as a reference process, meaning that it will not change when process X is reduced. The terms "categorical process" and "categorical variables" mean that—in every particular case s (e.g., at any given time s or for

Significance

We derive a computational framework that allows highly scalable identification of reduced Bayesian and Markov relation models, their uncertainty quantification, and inclusion of a priori physical information. It does not rely on the prior knowledge or a necessity of estimation of the full matrix of system's relations in any step. Application to a molecular dynamics (MD) example showed that this methodology opens possibilities for a robust construction of reduced Markov state models directly from the MD data—providing ways of bridging the gap toward longer simulation times and larger systems in computational MD.

Author contributions: I.H. designed research; S.G. and I.H. performed research; S.G. and I.H. analyzed data; and S.G. and I.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: illia.horenko@usi.ch.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1612619114/-/DCSupplemental.

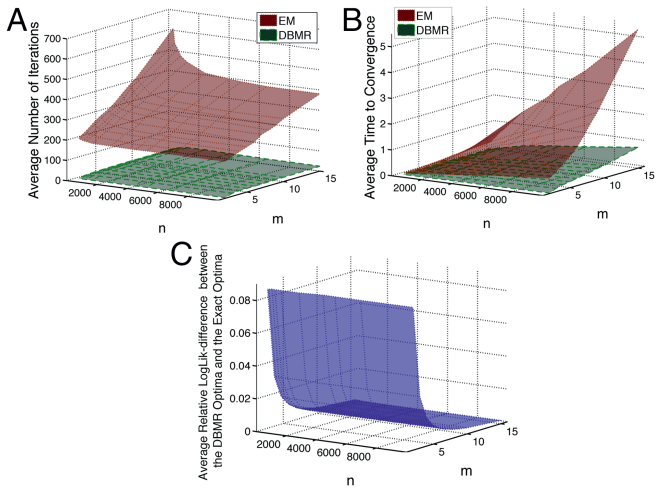

Fig. 1. Numerical comparison of the PLSA Expectation Maximization reduction (13, 14) and the DBRM reduction algorithms: (*A*) for the average number of iterations required until reaching the same convergence tolerance, (*B*) for the average central processing unit time until algorithms reach the same convergence tolerance, and (*C*) for the average relative log-likelihood difference between the optima achieved with the DBRM and the optima obtained with the exact iterative maximization of Eq. **3**. For every combination of problem dimensions *n* and *m*, averaging was performed over the ensemble of 1,000 randomly generated datasets that were subject to reduction with $K = 2$ for both of the algorithms. Convergence tolerance was measured in terms of the same normalized log-likelihood measure $1/mn\,\hat{L}$. Average relative log-likelihood difference was computed as $\mathbb{E}[|\hat{L}_{exact^*} - \hat{L}_{DBRM^*}|/|\hat{L}_{exact^*}|]$, where $\hat{L}$ is defined in Eq. **3**. MATLAB code generating this comparison is available at https://github.com/SusanneGerber. The code implementing PLSA Expectation Maximization methods (13, 14) is openly available at the MathWorks webpage (https://ch.mathworks.com/matlabcentral/fileexchange/56302-probabilistic-latent-sematic-analysis–tempered-em-and-em-). EM, Expectation Maximization.

any given instance s in the dataset)—$Y(s)$ is taking one and only one of the possible values from m categories $\{y(1), y(2), \ldots, y(m)\}$ and $X(s)$ is taking values from one and only one of the n categories $\{x(1), x(2), \ldots, x(n)\}$. For example, in biomolecular dynamics simulations of polypeptides with N amino acid residues, every peptide residue i at any time s can be assigned to one and only one of three Ramachandran states dependent on its current combination of torsion angle values $\phi_i(s)$ and $\psi_i(s)$ (*SI Appendix*, Fig. S1). Also, every global configuration/conformation X of the entire polypeptide molecule at any time can then be assigned to one of the $n \leq 3^N$ categories $\{x(1), x(2), \ldots, x(n)\}$—where every particular $x(k)$ is defined by a vector of Ramachandran state combinations [e.g., $x(k)$ is a category when junction 1 is being in state 1, junction 2 is being in state 2, and so on]. Efficient approaches based on the Markov state modeling (MSM) framework have been recently introduced, allowing for automated transformation of continuous-valued processes [e.g., molecular dynamics (MD) coordinates time series] into categorical time series (15, 16). Because the system cannot be in two different categories simultaneously, these categories are disjointed, and a relation between the probability for $Y(s)$ to attain a category $y(i)$ in its instance/realization s and the probabilities for $X(s)$ can be formulated exactly via the conditional probabilities and the law of the total probability (17). Defining the column vectors of probabilities $\Pi_Y(s) = \{\mathbb{P}[Y(s) = y(1)], \ldots, \mathbb{P}[Y(s) = y(m)]\}$, $\Pi_X(s) = \{\mathbb{P}[X(s) = x(1), \ldots,$ and $\mathbb{P}[X(s) = x(n)]\}$, we can write the exact relation between the variables X and Y in a matrix vector form:

$$\Pi_Y(s) = \Lambda \Pi_X(s), \qquad [1]$$

where matrix elements $\{\Lambda\}_{ij} = \mathbb{P}[Y(s) = y(i) | X(s) = x(j)]$ are conditional probabilities. If known, they can be used as indicators for existence of causality relations between the variables Y and X in the randomized studies: if $\{\Lambda\}_{ij} = \mathbb{P}[Y(s) = y(i)]$ for all j and s, the processes are then independent—meaning that information about the variable X provides no additional advantage in computing the probability of the outcomes of Y. If $\{\Lambda\}_{ij} \neq \mathbb{P}[Y(s) = y(i)]$ for some j, consequently, there exists some relation between X and Y (18). To be able to interpret these conditional probabilities as a measure of the true causality relations in practical studies when $\{\Lambda\}_{ij}$ are estimated from the available observations of X and Y, one needs to guarantee that the data are appropriately randomized.

In a particular case, where $m = n$, with s being the time index and $X(s) \approx Y(s - \tau)$ (where τ is a time step), the above formulation (Eq. **1**) is equivalent to a so-called master equation of a Markov process [and thereby, is a particular time-discrete case of the well-known time-continuous Fokker–Planck equation (17)]. The $n \times n$ matrix Λ in this case will be a transpose of the Markov transition operator (19). If matrix Λ is known, it provides full information about the relations between processes Y and X—and can be used to predict Y if X is available.

In many applications, the relation matrix Λ is not known and needs to be first estimated from the available observational data $\{X(1), X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2), \ldots, Y(S)\}$ [e.g., by means of the maximum log-likelihood approach that allows us to provide the analytical estimates of the most likely parameter values Λ^* and their uncertainties (lemma 1 in *SI Appendix*)]. However, in realistic applications (e.g., in the MD example below), the number of categories n can grow exponentially with the physical dimension of the problem ("curse of dimension")—leading to the exponential growth of overall uncertainty for the Λ^* estimates when the available statistics size S and a number m of Y-categories are fixed (lemma 2 in *SI Appendix*). This problem also means that the uncertainty of all additional physical observables obtained from Λ^* (e.g., the uncertainty of eigenvalues, eigenvectors, metastable sets, etc.) will be growing with the growing n, making practical deployment of Eq. **1** problematic for realistic systems with a "large" n and "small" S. Therefore, if we want to reduce the dimensionality n—for example, through identification of a small number K of collective categorical variables that agglomerate the original n categories of process X into K groups/boxes—then this methodology should not rely on a direct estimation of the full Bayesian causality matrix Λ in these situations.

To circumvent this problem, one can try to identify a latent reduced categorical process $\{\hat{X}(1), \hat{X}(2), \ldots, \hat{X}(S)\}$ (being a reduced representation of the full categorical process X) that is defined on a reduced statistically disjoint complete set of (the yet unknown) categories $\{\hat{x}(1), \hat{x}(2), \ldots, \hat{x}(K)\}$ with $K < n$. Deploying a law of a total probability, we can establish the Bayesian relations between $\hat{X}$ and X on one side (by means of the conditional probabilities $\hat{\Gamma}_{kj} = \mathbb{P}[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)]$) and between $\hat{X}$ and Y on the other side (by means of the conditional probabilities $\hat{\lambda}_{ik} = \mathbb{P}[Y(s) = y(i) | \hat{X}(s) = \hat{x}(k)]$). Then, it is straightforward to validate (a detailed derivation is in *SI Appendix*, section S2) that an optimal probability-preserving reduced approximation of the full relation model (Eq. **1**) for K colvars takes a form

$$\hat{\Pi}_Y(s) = \hat{\lambda} \hat{\Gamma} \Pi_X(s), \qquad [2]$$

where $\{\hat{\Pi}_Y(s)\}_i = \mathbb{P}[Y(s) = y(i)]$ and $i = 1, \ldots, m$. For every particular combination of k and j, $\hat{\Gamma}_{k,j}$ defines a probability for the colvar to be in a reduced collective categorical variable k when the observed original process X is in a category $x(j)$, and therefore, it can be understood as a discrete analog of the

continuous projection and reduction operators deployed in methods like PCA; $\hat{\lambda}$ is a reduced version of the matrix Λ from the full relation model (Eq. **1**). Please note that, being basically a reformulation of the exact law of the total probability, reduced model (Eq. **2**) is exact in the Bayesian sense, and no additional approximations have been involved.

A similar approach to latent variable dependency modeling is used in the PLSA (13, 14) (that is, used in mathematical linguistics and information retrieval for identification of latent dependency structures in texts and documents). Deploying the definition of a conditional probability, PLSA allows one to parameterize a joint probability distribution $\mathbb{P}[X \text{ and } Y]$ with the help of the latent process $\hat{X}$ as $\mathbb{P}[X(s)=x(j) \text{ and } Y(s)=y(i)] = \mathbb{P}[X(s)=x(j)] \sum_{k=1}^{K} \hat{\lambda}_{ik}\hat{\Gamma}_{kj}$. To estimate the parameters, one deploys an Expectation Maximization algorithm having the computational iteration complexity of $\mathcal{O}(K \cdot \min\{mn, S\})$ and requiring $\mathcal{O}((K+1) \cdot \min\{mn, S\})$ memory in a general nonsparse situation (i.e., when the underlying matrix Λ is not assumed to be sparse a priori). However, as shown in *SI Appendix*, section S5, this problem requires imposing additional strong independence and stationarity assumptions on the latent variable $\hat{X}$. Moreover, as shown in Fig. 1*A*, the total average number of Expectation Maximization iterations for this problem grows rapidly with problem dimensions m and n—resulting in the overall algorithm complexity that grows polynomially in n and m (Fig. 1*B*). Applying standard statistical methods of polynomial regression fitting and discrimination (20, 21), one obtains that the statistically optimal fit of the red surface (corresponding to the PLSA) from Fig. 1*B* is given by a polynomial of the third degree in n and m. Extrapolation to the typical physical problem sizes (e.g., $m=n=10^5$, $K=2$) that, for example, emerge in biophysical applications like the protein molecules indicates that such an inference procedure based on the Expectation Maximization algorithm and PLSA would require approximately 1,450 years of computations on a single laptop personal computer. Detailed methodological description of the PLSA methodology and its relation to the reduced Bayesian model reduction methods is provided in *SI Appendix*, section S5.

In the following section, we will suggest several computational procedures for the scalable inference of reduced Bayesian relation model parameters (Eq. **2**) directly from the observed data $\{X(1), X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2), \ldots, Y(S)\}$. The optimal parameter estimates $\hat{\Gamma}^*$ and $\hat{\lambda}^*$ that maximize the observation probability (called likelihood) of the given data in Eq. **2** can be obtained by solving the following log-likelihood maximization problem subject to equality and inequality constraints:

$$\hat{L} = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij} \log\left(\left\{\hat{\lambda}\hat{\Gamma}\right\}_{ij}\right) \to \max_{\hat{\lambda},\hat{\Gamma}}, \qquad [3]$$

$$\hat{\lambda}_{ik} \geq 0, \quad \sum_{i=1}^{m} \hat{\lambda}_{ik} = 1, \text{ for all } i,k, \qquad [4]$$

$$\hat{\Gamma}_{kj} \geq 0, \quad \sum_{k=1}^{K} \hat{\Gamma}_{kj} = 1, \text{ for all } k,j, \qquad [5]$$

where $N_{ij} = \sum_{s=1}^{S} \chi(Y(s)=y_i)\chi(X(s)=x_j)$ (with χ being an indicator function). It is straightforward to observe that, for any fixed $\hat{\lambda}$, the original exact log-likelihood maximization problem (Eqs. **3–5**) can be decomposed into n optimization problems for the n columns of $\hat{\Gamma}$—and each of the column problems with $(K-1)$ optimization arguments is concave and can be solved independently from the other column problems. This observation can help in designing a convergent algorithm requiring much less memory than the Expectation Maximization $[\mathcal{O}(K(m+n)+\min\{mn, S\})$ instead of $\mathcal{O}((K+1)\cdot \min\{mn, S\})$ for Expectation Maximization] and with computational iteration complexity of $\mathcal{O}\left((m-1)^3K^3 + n(K-1)^3\right)$. It can be used for identification of the reduced Bayesian relation model parameters in the situations when m and K are relatively small and n is large (e.g., as in the medical example 2 below). Detailed derivation of this algorithm is given in *SI Appendix*, section S4. However, when m or K is large (as in a case of the MSM inference in MD, where $m=n\approx 10^3-10^9$), this scaling would not allow us to apply this method to large realistic systems.

It turns out that substituting the function $\hat{L}$ with its lower-bound approximation $\hat{L} \geq \hat{l} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} N_{ij}\hat{\Gamma}_{kj}\log(\hat{\lambda}_{ik})$ (which directly results from applying the Jensen's inequality to Eq. **3**) allows for providing a computational method that can solve this approximate model reduction problem with a much better scaling and allows analytically computable uncertainty estimates for the obtained reduced models.

Properties of this approximate model reduction procedure are summarized in the following theorem.

Theorem. *Given the two sets of categorical data* $\{X(1), X(2), \ldots, X(S)\}$ *and* $\{Y(1), Y(2), \ldots, Y(S)\}$ *(where for any* $s, X(s) \in \{x(1), x(2), \ldots, x(n)\}$ *and* $Y(s) \in \{y(1), y(2), \ldots, y(n)\}$*), the approximate maximum log-likelihood parameter estimates for* $\hat{\lambda}$ *and* $\hat{\Gamma}$ *in the reduced model* (Eq. **2**) *can be obtained via a maximization of the lower bound* $\hat{l}$ *of the above log-likelihood function* $\hat{L}$ *from* Eq. **3**:

$$\hat{L} \geq \hat{l} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} N_{ij}\hat{\Gamma}_{kj}\log\left(\hat{\lambda}_{ik}\right) \to \max_{\hat{\lambda},\hat{\Gamma}}, \qquad [6]$$

subject to the constraints (Eqs. **4** and **5**). *Solutions of this problem exist and are characterized by the discrete/deterministic optimal matrices* $\hat{\Gamma}$ *that have only elements zero and one. Solutions of* Eqs. **4–6** *can be found in a linear time by means of the monotonically convergent Direct Bayesian Model Reduction (DBMR) Algorithm shown below, with a computational complexity of a single-iteration scaling as* $\mathcal{O}(K\cdot\min\{mn, S\})$ *and requiring no more than* $\mathcal{O}(K(m-1)+n+\min\{mn, S\})$ *of memory. Asymptotic posterior uncertainty of the obtained parameters* $\hat{\lambda}^*$ *(characterized in terms of the posterior parameter variance) can be computed analytically as* $\mathrm{Var}\{\mathbb{P}[\hat{\lambda}_{ik}|\hat{\lambda}^*, \hat{\Gamma}^* X, Y]\} = \hat{\lambda}^*_{ik}(1-\hat{\lambda}^*_{ik})/\sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\hat{\Gamma}^*_{kj}$. *The least biased estimate of the ratio* ρ *for the expectations of posterior parameter variances from the resulting full and reduced models equals*

$$\rho = \frac{\mathbb{E}_{ij}\mathrm{Var}\{\mathbb{P}[\Lambda_{ij}|\Lambda^*, X, Y]\}}{\mathbb{E}_{ik}\mathrm{Var}\left\{\mathbb{P}\left[\hat{\lambda}_{ik}|\hat{\lambda}^*, \hat{\Gamma}^* X, Y\right]\right\}} = \frac{n}{K}. \qquad [7]$$

DBMR Algorithm.

Choose a random $\hat{\lambda}^{(0)}$ *(e.g., from the least biased uniform prior), set* $I = 0$.

Set $\Gamma^{(0)}_{kj}$ *to 1 if* $k = \underset{k'}{\mathrm{argmax}} \sum_{i=1}^{m} N_{ij}\log(\hat{\lambda}^{(0)}_{ik'})$ *and else to 0 for all* j *and* k.

Do until $\|\hat{\mathbf{l}}(\Gamma^{(I)}, \hat{\lambda}^{(I)}) - \hat{\mathbf{l}}(\Gamma^{(I-1)}, \hat{\lambda}^{(I-1)})\|$ *becomes less than a tolerance threshold.*

Step 1: set $\hat{\lambda}^{(I+1)}_{ik} = \frac{\sum_{j=1}^{n} N_{ij}\Gamma^{(I)}_{kj}}{\sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\Gamma^{(I)}_{kj}}$ for all i,k.

Step 2: set $\Gamma^{(I+1)}_{kj} = 1$ *if* $k = \underset{k'}{\mathrm{argmax}} \sum_{i=1}^{m} N_{ij}\log(\hat{\lambda}^{(I+1)}_{ik'})$ *and else* $\Gamma^{(I+1)}_{kj} = 0$ *for all* j,k.

$I = I+1$.

A proof is provided in *SI Appendix*, section S2.

As can be seen from Fig. 1*A*, the average number of DBMR iterations (green surface in Fig. 1*A*) (computed from a large ensemble of randomly generated Bayesian model reduction problems for $K=2$) does not change with the dimensions m and n. It implies that also the overall computational complexity of the DBMR is scaling as $\mathcal{O}(K\cdot\min\{mn,S\})$. DBMR estimation of the reduced MSM for a medium-sized protein MD with $m=n=10^5$ and $K=2$ takes 33 min (as mentioned above, the extrapolated estimate of the Expectation Maximization computational time was 1,450 years under the same optimization and hardware/software settings).

Fig. 1*C* represents the average relative log-likelihood differences between the results of exact iterative log-likelihood optimization of Eqs. **3**–**5** and the DBMR results (obtained under the same conditions). It reveals that the empirical average relative log-likelihood differences between the exact and the DBMR-approximated results converge to zero exponentially in m. This property implies that, for realistic high-dimensional applications, the log-likelihood difference between the reduced models obtained with the DBRM algorithm and those obtained with the optimization of the exact log-likelihood can be expected to become negligible—meaning that the reduced models obtained with the DBRM algorithm will have essentially the same posterior probability for explaining the observed full data as the exact reduced models.

The main feature of the two algorithms presented above is that they allow for obtaining the reduced model (Eq. **2**) directly from the available observational data $\{X(1),X(2),\dots,X(S)\}$ and $\{Y(1),Y(2),\dots,Y(S)\}$—completely omitting a need for computation/estimation of the full relation matrix Λ in Eq. **1**. The only tunable parameter in both of the algorithmic procedures introduced above (in the direct sequential optimization of the exact log-likelihood (Eqs. **4**–**6**) and the DBMR algorithm) is the reduced process dimension K. The optimal integer value of K can be obtained by performing the algorithms with different numbers of K (i.e., $K=1,2,3,\dots$) and then selecting the best reduced model (Eq. **2**) according to one of the standard model selection criteria [e.g., cross-validation criterion, information criteria like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), or approaches like L curve] (22, 23). To select an optimal K for the examples below, we have used the standard L-curve method (23) that identifies the optimal K as the edge point of the curve that describes a dependence of the optimal value of the maximized function (Eq. **3** or **6** in our case) from K (a practical example is in *SI Appendix*, Fig. S2).

When dealing with real-life applications, it is also important to have an option for adjusting a set of collective variables according to a physical intuition or some prior knowledge (1). For example, one could have some prior physical information that certain dimensions of the original problem have a higher relevance for the dynamics than some other physically less relevant dimensions. In *SI Appendix*, section S3, we present a computationally scalable way [with computational iteration complexity of $\mathcal{O}(mK+n(K-1)\log[n(k-1)])$] to impose such a priori information—cast into a form of the weighted graph—on the DBMR algorithm. The resulting DBMR graph algorithm is presented in *SI Appendix*, section S4, and a practical application of this information-imposing clustering method to reduced Bayesian model inference is given in the breast cancer diagnostics example 2 below.

A MATLAB library of algorithms implementing the methods introduced in this manuscript—as well as different variants of the constrained Nonnegative Matrix Factorization (12, 24) and PLSA methods (13, 14)—can be found in *SI Appendix* and is available as open access via a general public license from https://github.com/SusanneGerber.

Results

Example 1: Reduced Model of the 10-Alanine Dynamics in Water. First, we consider a colvar identification for a polypeptide molecule [deca-alanine (10-ALA)] from results of the MD simulation. This dataset represents an output of the 0.5-μs simulation (with a 2-fs time step) of a 10-ALA polypeptide in explicit water at the room temperature performed with the Amber99sb-ildn force field (25). These MD data were produced and provided by Frank Noe and Antonia Mey, Free University (FU) Berlin, Berlin. For additional analysis, the values of torsion angles ϕ_i and ψ_i $(i=1,\dots,8)$ inside of the molecular backbone (i.e., ignoring the two end groups and the ω_i angles) are grouped into the Ramachandran states 1–3 for every i (*SI Appendix*, Fig. S1, *Left*) with a time step resolution of 100 ps, resulting in eight categorical Ramachandran time series with 5,000 time points each. Based on these eight local junction time series, we create a series of global molecular states $X(s)$ $(s=1,\dots,5{,}000)$, where every particular combination of eight Ramachandran states is assigned to a particular category; in our case, it is a categorical series with $n=531$ of such eight-component combinations with $S=5{,}000$ time instances. As a 531D $X(s)$ variable to be reduced, we use this set of global states; as reference variables Y_i, we choose the individual Ramachandran series of junctions (i.e., with $m=3$ each) at time $s+1$. Thereby, we are casting the reduction problem to a setting of discrete Markov processes in time.

We start with setting $K=2$ and comparing the practical performance of algorithms introduced in this paper with the PLSA method (13, 14). Results of this comparison are summarized in *SI Appendix*, Fig. S5. As can be seen from *SI Appendix*, Fig. S5, methods based on optimization of Eqs. **3** and **6** provide colvars that are better in terms of the log-likelihood measure as well as in terms of the information theoretical measures, like the robust AIC and BIC (22). AIC and BIC take into account the model quality and penalize model uncertainty—for the same quality (log-likelihood), these measures would provide smaller values for the models that are less uncertain (22).

Second, we do the identification of reduced models (Eq. **2**) for each of the peptide junctions $(i=1,\dots,8)$. Values of the resulting optimal solutions for reduced log-likelihoods $\hat{\mathbf{l}}_i$ $(i=1,\dots,8)$ as functions of K are shown in *SI Appendix*, Fig. S2. These results reveal that the reduced log-likelihood does not exhibit any non-negligible increase for all i when the number of colvars K is becoming larger than three to seven, meaning that the maximal number of the nonredundant colvars is not greater than seven for this system. Next, we inspect the identified colvars for all of the Y_i. As can be seen from *SI Appendix*, Fig. S3 (as an example, representing a case of Y_i being the Ramachandran time series of the junction 4 for $K=3$), the three identified colvars almost perfectly—to 97%—coincide with the discretization that is solely based on this junction and disregard all other junctions in the peptide chain. In only 3% of the cases, the nonlocal information about the Ramachandran states of the peptide residues from other junctions is important. Therefore, relations in terms of temporal causality between the peptides MD dynamics can be almost (in 97% of the cases) described by a sequence of spatially independent Markov processes in each of the peptide junctions—for example, collected together in a form of the Ising model (26). To verify the validity of the obtained colvars as well as test the performance of the resulting reduced model (Eq. **2**), we use these colvars to produce a long Monte Carlo time series of the reduced molecular simulation (Eq. **2**) and compute statistics of the geometrical configurations for the entire molecule. As shown in Fig. 2, reduced dynamics based on just a few colvars can reliably represent the overall spatial statistics of molecular configurations in 3D—obtained from the full MD trajectory. These 3% of nonlocal dependence cases identified in *SI Appendix*, Fig. S3 seem to be crucial: without them, the corresponding box plot

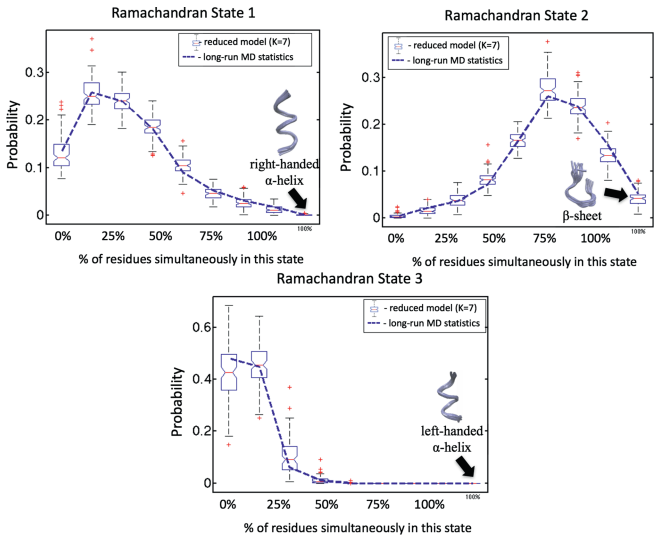

Fig. 2. Probabilities for different proportions of the chain in the same local Ramachandran state; 100% means that all of the residues in the chain are in this Ramachandran state, and 0% means that there is not a single residuum in this state. The blue lines indicate the values of this distribution obtained from the full MD simulation data, and the box plots show the probability distribution and its 95% confidence intervals obtained from the optimal reduced model run with seven colvars ($K = 7$) and nonlocal causality boxes. Red points denote the statistical outliers of the reduced model (meaning that they are outside of the 99% confidence interval). Respective distributions for a completely independent model (i.e., for a model where 100% of causality boxes are local) are shown in *SI Appendix*.

of the reduced Monte Carlo run is not capable of reproducing the true statistics of geometrical configurations from MD (*SI Appendix*, Fig. S4).

Example 2: Reduced Model for the Breast Cancer Diagnostics Based on the Standard BI-RADS Data. For the second example, we consider analysis and reduction of the standard Breast Imaging Reporting and Data System (BI-RADS) dataset for breast cancer diagnostics—available as an open access data file at the University of California Irvine (UCI) Machine-Learning Repository: mlr.cs.umass.edu/ml/datasets/Mammographic+Mass. This dataset contains information about 403 healthy (benign) and 427 malignant breast cancer patients. For each of the patient entries, the age and three categorical variables obtained from the mammography images are provided together with the basic result ("cancer"/"no cancer") obtained from the invasive analysis of the tissue—as well as assessments based on the standard noninvasive mammography diagnostics procedure called BI-RADS. The three categorical variables provide qualitative characteristics of the mammographic image features used in BI-RADS—such as the shape of the intrusion (with four categories), characteristics of the intrusion margins (with five categories), and intrusion density (with four categories). This standard categorical dataset is widely used to access the quality of various computer-aided diagnostic (CAD) tools, with the general aim of identifying such a CAD tool that would use the noninvasive information of age and mammographic image features for the precise diagnostics of breast cancer and providing lower rates of false positive and false negative diagnoses than the standard BI-RADS procedure currently used by medical doctors (27). The widely used measure of CAD performance adopted in the medical literature is called area under curve (AUC) (28). The closer the AUC value is to 1.0, the better the performance of the respective CAD tool and the lower the probability of false positive and false negative diagnoses. To compute the AUC values of different CAD tools together with the 99% confidence intervals of AUC, we use a methodology described in ref. 28 and available in the open source software library that can be downloaded at https://github.com/brian-lau/MatlabAUC. CAD tools based on the pattern recognition artificial neuronal networks (ANNs) have been reported to have the highest AUC for the breast cancer diagnostics issue (27). Training such ANNs on these data from 830 patients results in AUC of 0.85 with the 99% confidence interval of $[0.82, 0.91]$, whereas we obtain AUC of 0.82 with the 99% confidence interval of $[0.78, 0.87]$ for a standard BI-RADS diagnostics (on the same data and computed deploying the same methodology from ref. 28). Therefore, despite the fact that the AUC value of ANNs is somewhat larger than the AUC of BI-RADS, their confidence intervals are largely overlapping—meaning that, from the view point of statistics, this standard dataset does not reveal an advantage of the ANNs compared with BI-RADS. In addition, ANN has many free adjustable parameters (e.g., weights and biases of neurons, transfer functions, etc.), which increase the danger of overfitting for such relatively small data. The application of the categorical reduction procedure described in this paper results in an optimal set of just two collective variables that are both completely defined by information from a single categorical variable "margin." Very unexpectedly, obtained optimal decomposition into two colvars turns out to be completely independent from all other variables and can be summarized in a very simple diagnostic rule: if the intrusion margin on the mammography image is circumscribed, then the risk of breast cancer is low (12%), and if not, the risk of breast cancer is high (72%). Applying the same open source methodology for AUC confidence intervals on the same standard data as above, we find that this very simple rule (with no free tunable parameters at all) has the AUC value of 0.835 with the 99% confidence interval of $[0.79, 0.88]$ (i.e., in terms of the AUC performance, it is not worse than the ANN with approximately 20 free adjustable parameters).

Discussion

The most important features distinguishing the methodology presented in this paper from other approaches described in the literature are that it allows highly scalable (Fig. 1) identification of reduced Bayesian relation models, their uncertainty quantification, and inclusion of a priori physical information and does not rely on the prior knowledge or a necessity of estimation of the full matrix Λ (Eq. **1**) of system's relations in any step. It allows an identification of the colvars and the reduced relation models (Eq. **2**)—as well as the MSMs—directly from the observational data.

According to the above theorem, the least biased estimate of the ratio ρ for the expectations of posterior parameter variances from the resulting full and reduced models equals n/K—where n is the number of the original dimensions, and K is the reduced dimensionality. In application examples 1 and 2 shown above, ρ is of the order of 100—meaning that the reduced models (Eq. **2**) can be estimated from much shorter data series than those required for the full model without reduction. In the context of MD and other multiscale applications, this feature can be a used to bridge the gap toward longer time scales in simulations. In particular, in both examples, we have shown how the interpretation of the obtained colvars can provide clues about the locality or nonlocality of relations in the system. Application of this methodology in both examples revealed essentially local models (i.e., models where colvars mostly coincide with only one of the original systems dimensions). In example 1, we have shown that the relation between the local geometry changes of single peptide units in time is local to 97% and that, only in 3% of the original system's states assembled to colvar, they are nonlocal (and distinctively defined by the peptide junction

configuration farther away in the chain). In example 2, the two identified colvars were completely defined through only one of the original data dimensions and are entirely independent from all other information on the system. This seemingly oversimplification of the obtained reduced models could, however, be undermined by the comparison of results and predictions obtained for these very simple reduced models (Fig. 2 or the results of AUC comparison in example 2). The proposed methodology is very simple to implement and to use—we also provide a MATLAB toolbox with all of the methods from this manuscript as open access via the https://github.com/SusanneGerber. As was shown for two application examples, obtained results are straightforward to interpret and provide insights in the underlying systems as well as situations when the system's dimension n is large (e.g., $n = 531$ for the example 1) and standard approaches may be subject to the overfitting issues.

ACKNOWLEDGMENTS. S.G. was supported by Forschungsinitiative Rheinland-Pfalz through the Center for Computational Sciences in Mainz. I.H. is partly funded by the Swiss Platform for Advanced Scientific Computing, Swiss National Research Foundation Grant 156398 MS-GWaves, and the German Research Foundation (Mercator Fellowship in the Collaborative Research Center 1114 Scaling Cascades in Complex Systems).

1. Fiorin G, Klein M, Henin J (2013) Using collective variables to drive molecular dynamics simulations. *Mol Phys* 111:3345–3362.
2. Schölkopf B, Smola A, Müller KR (1997) *Kernel Principal Component Analysis*, eds Gerstner W, Germond A, Hasler M, Nicoud J (Springer, Berlin), pp 583–588.
3. Jolliffe I (2002) *Principal Component Analysis* (Springer, Berlin).
4. Schmid P (2010) Dynamic mode decomposition of numerical and experimental data. *J Fluid Mech* 656:5–28.
5. Donoho DL, Grimes C (2003) Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proc Natl Acad Sci USA* 100:5591–5596.
6. Coifman R, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc Natl Acad Sci USA* 102:7426–7431.
7. von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17:395–416.
8. Liou C, Cheng W, Liou J, Liou D (2014) Autoencoder for words. *Neurocomputing* 139:84–96.
9. Zhao Y, Levina E, Zhu J (2011) Community extraction for social networks. *Proc Natl Acad Sci USA* 108:7321–7326.
10. Perez-Hernández G, Paul F, Giorgino T, De Fabritiis G, Noe F (2013) Identification of slow molecular order parameters for markov model construction. *J Chem Phys* 139:015102.
11. Roblitz S, Weber M (2013) Fuzzy spectral clustering by pcca+: Application to markov state models and data classification. *Adv Data Anal Classif* 7:147–179.
12. Ding C, Li T, Peng W (2006) Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. *Proceedings of the 21st National Conference on Artificial Intelligence* (AAAI Press, Berkeley, CA), Vol 1, pp 342–347.
13. Hofmann T (1999) Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, New York), pp 50–57.
14. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42:177–196.
15. Prinz J, et al. (2011) Markov models of molecular kinetics: Generation and validation. *J Chem Phys* 134:174105.
16. Bowman G, Pande V, Noé F (2013) *An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation*, Advances in Experimental Medicine and Biology (Springer, Dordrecht, The Netherlands).
17. Gardiner H (2004) *Handbook of Stochastical Methods* (Springer, Berlin).
18. Holland P (1986) Statistics and causal inference. *J Am Stat Assoc* 81:945–960.
19. Schütte C, Sarich M (2013) *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*, Courant Lecture Notes (American Mathematical Society, New York).
20. Wahba G (1990) *Spline Models for Observational Data* (SIAM, Philadelphia).
21. Nocedal J, Wright SJ (2006) *Numerical Optimization* (Springer, New York), 2nd Ed.
22. Burnham K, Anderson D (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer, Berlin).
23. Hansen P, OLeary D (1993) The use of the l-curve in the regularization of discrete ill-posed problems. *SIAM J Sci Comput* 14:1487–1503.
24. Berry MW, Browne M, Langville AN, Pauca VP, Plemmons RJ (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput Stat Data Anal* 52:155–173.
25. Lindorff-Larsen K, et al. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78:1950–1958.
26. Gerber S, Horenko I (2014) On inference of causality for discrete state models in a multiscale context. *Proc Natl Acad Sci USA* 111:14651–14656.
27. Elter M, Schulz-Wendtland R, Wittenberg T (2007) The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Med Phys* 34:4164–4172.
28. Qin G, Hotilovac L (2008) Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Stat Methods Med Res* 17:207–221.

Supplement for the paper "Towards a direct and scalable identification of reduced models for categorical processes"

Susanne Gerber,[1] Illia Horenko,[2*]

[1] Johannes-Gutenberg University of Mainz, Staudinger Weg 9, 55128 Mainz, Germany
[2]Università della Svizzera Italiana Via G. Buffi 13, 6900 Lugano Switzerland
[*]To whom correspondence should be addressed; E-mail: horenkoi@usi.ch

This supplement is subdivided into the six following chapters:

1. Derivation of the full Bayesian relation model and its properties (maximum log-likelihood estimates, their uncertainty).

2. Derivation of the reduced Bayesian model formulation and its properties, proof of the theorem from the main manuscript.

3. Imposing a priori information on DBMR.

4. Description of the Bayesian Model Reduction algorithms introduced in the main manuscript.

5. Methodological comparison of the Bayesian Model Reduction methods introduced in the manuscript to the Probabilistic Latent Semantic Analysis (PLSA).

6. Additional figures.

1 Derivation of the full Bayesian relation model and its properties (maximum log-likelihood estimates, their uncertainty)

Here, we will present a derivation of a full model of Bayesian relations between the categorical processes and investigate some of it's properties. Process Y below will serve as a reference process - meaning that it will not change when process X is reduced. The terms 'categorical process' and 'categorical variables' mean that - in every particular case - s (e.g., at any given time s or for any given instance s in the data set) $Y(s)$ is taking only one of the possible values from m categories $\{y(1), y(2), \ldots, y(m)\}$ and $X(s)$ from only one of the n categories $\{x(1), x(2), \ldots, x(n)\}$. Since the system can not be in two different global categories simultaneously, these categories are disjoint and relation between the probability for $Y(s)$ to attain a category $y(i)$ in its instance/realisation s and the probabilities for $X(s)$ can be formulated exactly (i.e., without the model error [8]) via the conditional probabilities and the *law of the total probability* [6]:

$$\mathbb{P}\left[Y(s) = y(i)\right] \;\; = \;\; \sum_{j=1}^{n} \Lambda_{ij} \mathbb{P}\left[X(s) = x(j)\right], \tag{1}$$

where matrix elements $\{\Lambda\}_{ij} = \mathbb{P}[Y(s) = y(i)|X(s) = x(j)]$ are conditional probabilities. They can be used as indicators for existence of causality relations between the processes Y and X in the randomised studies: if $\{\Lambda\}_{ij} = \mathbb{P}[Y(s) = y(i)]$ for all j and s the processes are then independent - meaning that information about process X provides no additional advantage in computing the probability of the outcomes of Y. If $\{\Lambda\}_{ij} \neq \mathbb{P}[Y(s) = y(i)]$ for some j, consequently, there exists some relation between X and Y [12]. To be able to interpret these conditional probabilities as some measure of the true causality relations in practical studies when $\{\Lambda\}_{ij}$ are estimated from the available observations of X and Y, one needs to guarantee that the data is appropriately *randomised*, meaning that the resulting Λ-estimates are unbiased by the presence of hidden/latent variables [12, 9].

Defining the column vectors of probabilities $\Pi_Y(s) = \{\mathbb{P}[Y(s) = y(1), \ldots, \;\; \mathbb{P}[Y(s) = y(m)]\}$, $\Pi_X(s) = \{\mathbb{P}[X(s) = x(1), \ldots, \;\; \mathbb{P}[X(s) = x(n)]\}$, we can re-write the above equation in a matrix-

vector form:

$$\Pi_Y(s) \;\; = \;\; \Lambda \Pi_X(s). \tag{2}$$

In a particular case, when $m = n$, when s is the time index and $X(s) \equiv Y(s - \tau)$ (where τ is a time-step), the above formulation (2) is equivalent to a so-called master equation of a Markov process (and - thereby - is a particular time-discrete case of the well-known time-continuous Fokker-Planck equation [6]). The $n \times n$-matrix Λ in this case will be a transpose of the Markov transition operator [14]. If matrix Λ is known it provides full information about the relations between processes Y and X - and can be used to simulate Y, if X is available. However, in many application areas (e.g., in molecular dynamics, computer-aided disease diagnostics, in genome-wide association studies (GWAS) in genomics) this matrix is not directly available and can only be estimated from data.

If observational data $\{X(1), X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2), \;\; \ldots, Y(S)\}$ are available, an estimate Λ^* of the rectangular $m \times n$ matrix Λ can be obtained by applying the well-known maximum likelihood principle. Hereby one seeks for a set of such parameters $\hat{\Lambda}^*$ that would maximise the total probability (or *likelihood*) of simultaneously observing these two particular sequences of observations. In this procedure the observational data is assumed to be fixed, resulting in the following expression for the log-likelihood function (i.e., a logarithm of the observational probability):

$$\begin{aligned} \mathbf{LogL}\left(\Lambda\right) &= \log\left(\mathbb{P}\left[\{X(1), X(2), \ldots, X(S)\} \text{ and } \{Y(1), Y(2), \ldots, Y(S)\} \,|\Lambda\right]\right) = \\ &= \sum_{s=1}^{S} \sum_{i=1}^{m} \sum_{j=1}^{n} \chi(Y(s) = y_i) \chi(X(s) = x_j) \log P\left[Y(s) = y_i | X(s) = x_j\right] = \\ &= \sum_{s=1}^{S} \sum_{i=1}^{m} \sum_{j=1}^{n} \chi(Y(s) = y_i) \chi(X(s) = x_j) \log \Lambda_{ij} = \sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij} \log \Lambda_{ij}, \end{aligned} \tag{3}$$

with

$$\Lambda^* \;\; = \;\; \underset{\Lambda}{\operatorname{argmax}} \left\{\mathbf{LogL}\left(\Lambda\right)\right\} = \underset{\Lambda}{\operatorname{argmax}} \left\{ \sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij} \log \Lambda_{ij} \right\}, \tag{4}$$

where $N_{ij} = \sum_{s=1}^{S} \chi(Y(s) = y_i)\chi(X(s) = x_j)$ (with χ being an indicator function) is the total number of instances in the data when $X(s)$ was in category $x(j)$ and - simultaneously - $Y(s)$ was in category $y(i)$. To guarantee that the elements of the obtained Λ^* preserve the basic properties of conditional probability, maximisation of the log-likelihood function (3) must be subject to the following linear equality and inequality constraints

$$\Lambda_{ij} \geq 0, \quad \sum_{i=1}^{m} \Lambda_{ij} = 1, \text{ for all } i, j. \tag{5}$$

It turns out that analogously to how it is done for the standard estimation of Markov transition matrices [6], the optimal solution of this constrained minimisation problem (4-5) and further properties of the resulting estimates can be retrieved analytically. These results and properties are summarised in the following two Lemma.

Lemma 1: *for the given observational data* $\{X(1), X(2), \ldots, X(S)\}$ *and* $\{Y(1), Y(2), \ldots, Y(S)\}$ *the maximum log-likelihood problem (4-5) has a unique solution* $\Lambda^*_{ij} = N_{ij} / \sum_{i=1}^{m} N_{ij}$ *(where* $i = 1, \ldots, m$ *and* $j = 1, \ldots, n$*) if* $\sum_{i=1}^{m} N_{ij} \neq 0$*. Asymptotic posterior uncertainty* $\sigma^2_{ij} = \mathrm{Var}\left\{P\left[\Lambda_{ij} | \Lambda^*_{ij}, X, Y\right]\right\}$ *of this maximum-likelihood parameter estimate* $\Lambda^*_{i,}$ *(characterised in terms of the posterior parameter variance) can be computed analytically as* $\sigma^2_{ij} = \frac{\Lambda^*_{ij}(1-\Lambda^*_{ij})}{\sum_{i=1}^{m} N_{ij}}$.

Proof: First of all, it is straightforward to see that obtaining the solution for the above problem (4-5) is equivalent to obtaining the solutions for the following n independent maximisation problems (with $j = 1, \ldots, n$):

$$\Lambda^*_{:j} = \underset{\Lambda}{\mathrm{argmax}} \left\{ \sum_{i=1}^{m} \frac{N_{ij}}{\sum_{i=1}^{m} N_{ij}} \log \Lambda_{ij} \right\}, \text{s.t.} \tag{6}$$

$$\Lambda_{:j} \geq 0, \quad \sum_{i=1}^{m} \Lambda_{ij} = 1, \text{ for all } i. \tag{7}$$

Since $\sum_{i=1}^{m} \frac{N_{ij}}{\sum_{i=1}^{m} N_{ij}} \equiv 1$, for any $j = 1, \ldots, n$ these problems are convex combinations of the concave optimisation problems defined on the convex and bounded domains - therefore solutions to

all of these problems will exist and be unique if the respective $\sum_{i=1}^{m} N_{ij} \neq 0$. Therefore, also a solution of the original problem (4-5) will exist and be unique. Deploying a method of Lagrange multipliers and taking into account only the equality constraints from (5) first, for $\sum_{i=1}^{m} N_{ij} \neq 0$ we obtain that

$$\Lambda^{*}_{ij} \quad = \quad \frac{N_{ij}}{\sum_{i=1}^{m} N_{ij}}. \tag{8}$$

Alternatively, one can obtain the same result without the Lagrange multipliers methods, by observing that the above expression (6) represents a negative of the cross entropy distance between the unknown distribution $\Lambda_{:j}$ and a distribution $\frac{N_{:j}}{\sum_{i=1}^{m} N_{ij}}$ that is given by the observational data. Thereby, maximisation of (6-7) is equivalent to the minimisation of the cross entropy between the two distributions - attaining the minimum if the two distributions are equal, i.e., when (8) is fulfilled.

Next, we consider a problem of estimating a variance of the maximum log-likelihood estimator for Λ^{*} (8). For this purpose we can first deploy the Bayes theorem that will allows us to express the posterior distribution of the parameter Λ^{*} (conditioned on fixed observed sequences $\{X(1), X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2), \ \ldots, Y(S)\}$) in terms of the log-likelihood function (3) and the prior probabilities $\mathbb{P}_{\text{prior}}\left[\{X(1), X(2), \ldots, X(S)\}, \{Y(1), Y(2), \ldots, Y(S)\}\right]$ and $\mathbb{P}_{\text{prior}}\left[\Lambda\right]$. Applying the logarithm to both sides of the obtained expression we gain:

$$\log\left(\mathbb{P}_{post}\left[\Lambda | \left\{X(1), X(2), \ldots, X(S)\right\}, \left\{Y(1), Y(2), \ldots, Y(S)\right\}\right]\right) =$$

$$= \mathbf{LogL}\left(\Lambda\right) + \log\left(\mathbb{P}_{\text{prior}}\left[\Lambda\right]\right) - \log\left(\mathbb{P}_{\text{prior}}\left[\left\{X(1), X(2), \ldots, X(S)\right\}, \left\{Y(1), Y(2), \ldots, Y(S)\right\}\right]\right) =$$

$$= \mathbf{LogL}\left(\Lambda\right) + const1, \tag{9}$$

where in the second equality we used the fact that the least-biased prior for the bounded categorical variables and discrete stochastic matrices (bounded on the intervals $[0, 1]$) are the uniform priors[1]. In

[1]Please note that the essential issue here is the boundedness of the domain of realisations for categorical processes: if the X and Y where the unbounded random processes in R^{n}, then according to the maximum-entropy principle the least-biased prior with the fixed variance would be a Gaussian prior [6].

the following we will estimate the marginal variance of this posterior distribution $\mathbb{P}_{post}\left[\Lambda|\left\{X(1),X(2),\ldots,X(S)\right\},\left\{Y(1),Y(2),\ldots,Y(S)\right\}\right]$ with respect to a single matrix element Λ_{ij} - when all of the other matrix elements $\Lambda^*_{i'j'}$ (for all $i{\neq}i'$ and $j{\neq}j'$) are kept fixed for a given series of the observed data X and Y. In such a case we can formulate the maximum log-likelihood problem (4-5) as an unconstrained optimisation problem:

$$\mathbf{LogL}\left(\Lambda_{ij}\right) \;=\; N_{ij}\log(\Lambda_{ij}) + (N_j - N_{ij})\log(1-\Lambda_{ij}) + const, \tag{10}$$

where $N_j = \sum_{i=1}^{m} N_{ij}$ and $const$ is a factor that is independent of Λ_{ij}.

Again, multiplying (10) with $\frac{1}{N_j}$ (by assuming that $N_j{\neq}0$), we get an equivalent optimisation problem that is concave (being a convex combination of concave problems). Being defined on a bounded concave domain, this problem has a unique solution that can be obtained by setting the first derivative of (10) to zero and solving the resulting equation with respect to Λ_{ij}. It is instructive to see that it results in the same maximum log-likelihood estimate (8).

Considering the $\log(\Lambda_{ij})$ as random realisations of the i.i.d. random variable we can observe that the right-hand side of (4) is the expectation of this variable and deploying the central limit theorem[2], we get that asymptotically (for $S \to \infty$) the posterior distribution of the parameter Λ^*_{ij} is a Gaussian distribution of the form

$$\mathbb{P}\left[\Lambda_{ij}|\Lambda^*\right] \;=\; \exp\left[\mathbf{LogL}\left(\Lambda^*_{ij}\right) + 0.5(\Lambda_{ij}-\Lambda^*_{ij})^2\partial^2\mathbf{LogL}\left(\Lambda^*_{ij}\right)\right] \tag{11}$$

where we essentially have deployed the quadratic Taylor-approximation of the function $\log\left(\mathbb{P}_{post}\left[\Lambda|\left\{X(1),X(2),\ldots,X(S)\right\},\left\{Y(1),Y(2),\ldots,Y(S)\right\}\right]\right)$ around the maximum log-likelihood

[2]The line of the argument that is deployed here is exactly the same as the one used in the proofs of the asymptotic Akaike Information Criterium, we refer to [1] for further details.

estimate Λ^*_{ij}. Then, the variance of this Gaussian distribution can be expressed analytically as

$$\begin{aligned}\sigma^2_{ij} &= \mathrm{Var}\left\{P\left[\Lambda_{ij}|\Lambda^*, X, Y\right]\right\} = -\frac{1}{\partial^2 \mathbf{LogL}\left(\Lambda^*_{ij}\right)} = \\ &= -\frac{1}{-\frac{N_j^2}{N_{ij}} - \frac{N_j^2}{N_j - N_{ij}}} = \frac{N_{ij}(N_j - N_{ij})}{N_j^3} = \frac{\Lambda^*_{ij}(1-\Lambda^*_{ij})}{N_j}. \quad \square \end{aligned} \tag{12}$$

Lemma 2: *if the categorical process X is obtained from the box-discretisation of a d-dimensional process with N discretisation boxes per dimension, then the supremum of a least-biased estimate σ^2_{ij} for the uncertainty of the maximum log-likelihood parameter estimates Λ^*_{ij} is given by the expression* $\sup \sigma^2_{ij} = \frac{N^d(m-1)}{Sm^2}$

Proof: since the categories $x(1), \ldots, x(n)$ of the full process X should be statistically-disjoint (to allow for applying the law of the total probability (1)), every category $x(j)$ should correspond to a one of the N^d original dimension box-combinations - i.e., $n \leq N^d$. According to the maximum-entropy principle [6], the least-biased prior distribution for the data matrix elements N_{ij} is provided by the uniform prior, i.e., by $N_{ij} = S/(nm)$. It is easy to verify that this results in the maximum log-likelihood estimate

$$\Lambda^*_{ij} = 1/m, \tag{13}$$

and substituting (14) and $n \leq N^d$ into (11) we get

$$\sigma^2_{ij} = \frac{n(m-1)}{Sm^2} \leq \sup_{ij} \sigma^2_{ij} = \frac{N^d(m-1)}{Sm^2} \quad \square \tag{14}$$

Interpretation: In realistic applications the number of categories n can grow exponentially with the physical dimension of the problem ("curse of dimension") - leading to the exponential growth of uncertainty for the Λ^* estimation problem, when the available statistics size S and number m of Y-categories are fixed. Obtained formula (11) also means that the uncertainty of all further physical observables obtained from Λ^* will be also growing with the growing n, making practical deployment

of eq. (2) problematic for realistic systems with "large" n and "small" S. This also means that if we want to reduce the dimensionality n - for example, through identification of a small number K of collective categorical variables that agglomerate the original n categories of process X into K groups/boxes - then this methodology should rather not rely on a direct estimation of the full Bayesian causality matrix Λ^* in these situations when "n" is large and "S" is small. In the following we shall present a simple idea circumventing the need of direct estimation/computation of Λ^* and allowing for a direct computation of the reduced/agglomerated collective variables.

2 Derivation of the reduced Bayesian model formulation and its properties, proof of the theorem from the main manuscript.

To achieve this aim, we shall be looking for a categorical process $\{\hat{X}(1), \hat{X}(2), \ldots, \hat{X}(S)\}$ (being a reduced representation of the full categorical process X) that is defined on a reduced set of categories $\{\hat{x}(1), \hat{x}(2), \ldots, \hat{x}(K)\}$ with $K < n$. Deploying again the law of the total probability we can establish a relation between the probability density $\hat{\pi}_{\hat{X}}(s)$ of this - still unknown - process and the full probability density of the observed process $\Pi_X(s)$:

$$\hat{\pi}_{\hat{X}}(s) \;\; = \;\; \hat{\Gamma}\Pi_X(s), \tag{15}$$

where $\hat{\Gamma}_{kj} = \mathbb{P}\left[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)\right]$ is the matrix of Bayesian conditional probabilities relating the two processes X and $\hat{X}$. This matrix can also be understood as a discrete probabilistic projection operator, playing in the following a similar role as the linear projection operators built of the dominant eigenvectors of the relation matrices in standard reduction methods (e.g., projection matrices built from the dominant eigenvectors of the data covariance matrix deployed in the Principal Component Analysis method or the Frobenius-eigenvectors of propagator and generator matrices used in the spectral reduction theory of Markov chains). The main conceptual difference of the $\hat{\Gamma}$ matrix from the projection operators obtained in the standard reduction methods is that it is preserv-

ing the l_1-norm - thereby guaranteeing that the reduced process density $\pi_{\hat{X}}$ will always preserve the probability (i.e., $\sum_{i=1}^{K} \{\hat{\pi}_{\hat{X}}(s)\}_i \equiv 1$ and $\{\hat{\pi}_{\hat{X}}(s)\}_i \geq 0$ for all i). In contrast, the projection matrices in standard approaches like Markov spectral reduction theory and PCA are l_2 objects and do not automatically preserve a probability of the reduced density in this sense.

Next, we deploy a law of the total probability to establish a Bayesian relation between the reduced (and unobserved) process $\hat{X}$ and the observed process Y:

$$\hat{\Pi}_Y(s) \;\; = \;\; \hat{\lambda}\pi_{\hat{X}}(s), \tag{16}$$

where $\hat{\lambda}_{ik} = \mathbb{P}\left[Y(s) = y(i)|\hat{X}(s) = \hat{x}(k)\right]$ is the matrix of conditional probabilities connecting the two processes. Substituting (15) into (16) we obtain a reduced representation of the original model (2):

$$\hat{\Pi}_Y(s) \;\; = \;\; \hat{\lambda}\hat{\Gamma}\hat{\pi}_X(s), \tag{17}$$

that now connects the observed processes not directly (as was the case for the full model (2)) but rather indirectly - through a latent reduced process $\hat{X}$ that is defined on a categorical space of a smaller dimension K.

If the two sets of categorical data $\{X(1),\;\; X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2), \ldots, Y(S)\}$ (where for any s, $X(s) \in \{x(1), x(2), \ldots, x(n)\}$ and $Y(s) \in \{y(1), y(2), \ldots, y(n)\}$) are given, the maximum log-likelihood parameter estimates for $\hat{\lambda}$ and $\hat{\Gamma}$ in the reduced model (17) can be obtained with one of the two following computational approaches.

Analogously to the procedure deployed in the **Lemma 1** above, one can try to estimate the matrices $\hat{\Gamma}$ and $\hat{\lambda}$ directly from the observed data $\{X(1), X(2), \ldots, X(S)\}$ and $\{Y(1), Y(2),\;\; \ldots, Y(S)\}$. The optimal parameter estimates can be obtained by solving the exact log-likelihood maximisation

problem subject to equality and inequality constraints:

$$\hat{L} = \sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\log\left\{\hat{\lambda}\hat{\Gamma}\right\}_{ij} \rightarrow \max_{\hat{\lambda},\hat{\Gamma}}, \quad \text{s.t.} \tag{18}$$

$$\hat{\lambda}_{ik} \geq 0, \quad \sum_{i=1}^{m}\hat{\lambda}_{ik} = 1, \text{ for all i,k}, \tag{19}$$

$$\hat{\Gamma}_{kj} \geq 0, \quad \sum_{k=1}^{K}\hat{\Gamma}_{kj} = 1, \text{ for all } k,j, \tag{20}$$

Structure of this problem motivates deployment of the iterative methods (e.g., of the sequential quadratic programming procedures [13]) - since the parameters $\hat{\lambda}$ and $\hat{\Gamma}$ naturally separate the problem into two concave maximisation problems with linear equality and inequality constraints. However, following the standard procedure for this particular problem (i.e., substitution of the linear equality constraints into (18) - followed by taking the partial derivatives of the resulting function with respect to the arguments $\hat{\lambda}$ and $\hat{\Gamma}$ and setting the obtained derivatives to zero) - results in the nonlinear system of equations that can not be solved analytically. Moreover, resulting system of equations does not include the inequality constraints, providing no guarantee that the obtained solutions will be non-negative. And - as it is straightforward to verify - the full numerical solution of the problem (18,19,20) by means of gradient-based optimisation methods would require $\left(\mathcal{O}\left(\left(2K-1\right)^{3}\left(n+m\right)^{3}\right)+\mathcal{O}\left(S\right)\right)$ of operations. It means that the numerical cost of this reduced model identification procedure will scale polynomially with the dimension n - prohibiting an application of this method to realistic problems with large n. Computer code implementing this seqential quadratic optimisation algorithm is provided for open access as a part of the BMR-toolbox (in Matlab) that can be found in the supplemental information to this manuscript and over the GitHub-portal `www.github.com`.

It turns out that substituting the function $\hat{L}$ with its lower-bound approximation $\hat{L} \geq \hat{l} = \\ = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} N_{ij}\hat{\Gamma}_{kj}\log\left(\hat{\lambda}_{ik}\right)$ (which directly results from applying the Jensen's inequality to (18)) allows providing a computational method that can solve this approximate model reduction problem with a linear scaling of the computational cost in n. Moreover, as will be demonstrated

below, it provides those solutions that are more simple and "regular" - meaning that the obtained model matrices (e.g., the projector matrix $\hat{\Gamma}$) require much less parameters to encode them[3]. As shown on the application example, this results in two further practical advantages: (i) the reduced approximate models allow a better physical interpretation (since the original systems dimensions are thereby sharply and "deterministically" assigned to the reduced systems dimensions); and (ii) obtained models are less subject to the overfitting issues and are more advantageous in terms of the model quality measures (like information criteria) - measures that take into account both the model quality and the model complexity [3].

Properties of this approximate model reduction procedure are summarised in the following theorem.

Theorem: *Given two sets of categorical data* $\{X(1),\ X(2),\ldots,X(S)\}$ *and* $\{Y(1),Y(2),\ldots,Y(S)\}$ *(where for any* s, $X(s) \in \{x(1),x(2),\ldots,x(n)\}$ *and* $Y(s) \in \{y(1),y(2),\ldots,y(n)\}$*), the approximate maximum log-likelihood parameter estimates for* $\hat{\lambda}$ *and* $\hat{\Gamma}$ *in the reduced model (17) can be obtained via a maximisation of the lower bound* $\hat{l}$ *of the above log-likelihood function* $\hat{L}$ *from (18):*

$$\hat{L} \ \geq \ \hat{l} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} N_{ij}\hat{\Gamma}_{kj}\log\left(\hat{\lambda}_{ik}\right) \rightarrow \max_{\hat{\lambda},\hat{\Gamma}}, \tag{21}$$

subject to the constraints (19,20). Solutions of this problem exist and are characterised by the discrete/deterministic optimal matrices $\hat{\Gamma}$ *that have only elements zero and one. Solutions of (21,19,20) can be found in a linear time, by means of the monotonically-convergent DBMR-Algorithm shown below, with a computational complexity of a single iteration scaling as* $\mathcal{O}\left(K\cdot\min\{mn,S\}\right)$ *and requiring no more then* $\mathcal{O}\left(K(m-1)+n+\min\{mn,S\}\right)$ *of memory. Asymptotic posterior uncertainty of the obtained parameters* $\hat{\lambda}^*$ *(characterised in terms of the posterior parameter variance) can*

[3]The optimal assignment of projector elements $\hat{\Gamma}_{ij}$ resulting from the DBMR-algorithm introduced below appears to be either zero or one - which requires much less basis functions (e.g., wavelet basis functions) to represent the rows of the obtained matrices as compared to $\hat{\Gamma}$ with elements being everywhere between zero and one.

be computed analytically as $\mathrm{Var}\left\{\mathbb{P}\left[\hat{\lambda}_{ik}|\hat{\lambda}^*,\hat{\Gamma}^*X,Y\right]\right\} = \frac{\hat{\lambda}^*_{ik}(1-\hat{\lambda}^*_{ik})}{\sum_{i=1}^m\sum_{j=1}^n N_{ij}\hat{\Gamma}^*_{kj}}$. *The least-biased estimate of the ratio* ρ *for the expectations of posterior parameter variances from the resulting full and reduced models equals to*

$$\rho = \frac{\mathbb{E}_{ij}\,\mathrm{Var}\left\{\mathbb{P}\left[\Lambda_{ij}|\Lambda^*,X,Y\right]\right\}}{\mathbb{E}_{ik}\,\mathrm{Var}\left\{\mathbb{P}\left[\hat{\lambda}_{ik}|\hat{\lambda}^*,\hat{\Gamma}^*X,Y\right]\right\}} = \frac{n}{K}. \tag{22}$$

Direct Bayesian Model Reduction Algorithm (DBMR-Algorithm)

Choose random $\hat{\lambda}^{(0)}$, *set* $I = 0$.

Set $\Gamma^{(0)}_{kj}$ to 1 if $k = \underset{k'}{\mathrm{argmax}}\sum_{i=1}^m N_{ij}\log(\hat{\lambda}^{(0)}_{ik'})$ and else to 0 - for all j and k.

Do until $\|\hat{\mathrm{I}}(\Gamma^{(I)},\hat{\lambda}^{(I)}) - \hat{\mathrm{I}}(\Gamma^{(I-1)},\hat{\lambda}^{(I-1)})\|$ **becomes less then a tolerance threshold**:

Step 1: set $\hat{\lambda}^{(I+1)}_{ik} = \frac{\sum_{j=1}^n N_{ij}\Gamma^{(I)}_{kj}}{\sum_{i=1}^m\sum_{j=1}^n N_{ij}\Gamma^{(I)}_{kj}}$ *for all* i,k

Step 2: set $\Gamma^{(I+1)}_{kj} = 1$ if $k = \underset{k'}{\mathrm{argmax}}\sum_{i=1}^m N_{ij}\log(\hat{\lambda}^{(I+1)}_{ik'})$

and else $\Gamma^{(I+1)}_{kj} = 0$ - for all j,k .

$I = I + 1$.

Proof: *Step 1 (existence of solution)* Since $0 \geq \hat{L} \geq \hat{l}$, function $\hat{l}$ is bounded with zero from above. Existence of a solution for the respective optimisation problem then follows straightforwardly from the boundedness of the function (21) and boundedness of a convex $[0,1]$-simplex domain defined by the linear constraints (19,20) [13]. Please note that this solution might not be unique.

Step 2 (uniqueness of the analytical solution wrt. $\hat{\lambda}$ *for a fixed parameter* $\hat{\Gamma}$*)* For any fixed $\hat{\Gamma}$ that satisfies (20), the problem (21,19) becomes a concave maximisation problem wrt. $\hat{\lambda}$ that is subject to linear equality and inequality constraints. Deploying a standard method of Lagrange multipliers for equality constraints only, if $\sum_{i=1}^m\sum_{j=1}^n\left\{\hat{\Gamma}\right\}_{kj} N_{ij}\neq 0$ (for all $k = 1,\ldots,K$) one obtains a unique

optimal solution:

$$\left\{\hat{\lambda}\right\}^{*}_{ik} \;=\; \frac{\sum_{j=1}^{n}\left\{\hat{\Gamma}\right\}_{kj} N_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}\left\{\hat{\Gamma}\right\}_{kj} N_{ij}}, \tag{23}$$

that apparently also satisfies the inequality constraints in (19). Therefore, it will be also a unique solution of the full problem (21,19,20) when $\hat{\Gamma}$ is fixed.

Step 3 (discrete analytical solution wrt. $\hat{\Gamma}$ for a fixed parameter $\hat{\lambda}$) For any fixed $\hat{\lambda}$ that satisfies (19), the problem (21,20) is a linear maximisation problem (LP) with block-diagonal matrices of linear equality and inequality constraints. Due to this block-diagonal structure of constraints, solution of this LP-problem is equivalent to an independent solution of the n following LP-problems - separately for every j:

$$\left\{\begin{array}{ll} \sum_{k=1}^{K}\alpha_{kj}\hat{\Gamma}_{kj} & \rightarrow \max\limits_{\hat{\Gamma}_{1j},\ldots,\hat{\Gamma}_{Kj}} \\ \text{s.t.} \quad \sum_{k=1}^{K}\hat{\Gamma}_{kj}=1, & \hat{\Gamma}_{kj}\geq 0, \end{array}\right. , \tag{24}$$

where $\alpha_{kj} = \sum_{i=1}^{m} N_{ij}\log\hat{\lambda}_{ik}$ are fixed non-positive constants when $\hat{\lambda}$ is fixed. Then, as it is very easy to check, substituting the following expression for Γ^{*}

$$\hat{\Gamma}^{*}_{kj} \;=\; \left\{\begin{array}{ll} 1, & \text{if } k=\operatorname{argmax}\limits_{k'}\{\alpha_{k'j}\} \\ 0, & \text{else} \end{array}\right. , \tag{25}$$

into (24) (when $\operatorname{argmax}\limits_{k'}\{\alpha_{k'j}\}$ is unique for all j) provides a maximum value to the LP-functions that also satisfies the constraints. When the $\operatorname{argmax}\limits_{k'}\{\alpha_{k'j}\}$ is not unique (i.e., when there are some j for which there exists some set $k=\{k_1,k_2,\ldots,k_p\}$ such that $\alpha_{k_1 j}=\cdots=\alpha_{k_p j}=\max\limits_{k'}\{\alpha_{kj}\}$) then the solution of (24) is not unique and every combination of $\hat{\Gamma}_{kj}$ that satisfies $\hat{\Gamma}_{k_1 j}+\cdots+\hat{\Gamma}_{k_p j}=1$, $\hat{\Gamma}_{:j}\geq 0$ - including a deterministic one (where one arbitrarily-selected $\Gamma_{k'j}$ ($k'\in k$) is set to one and all other $\Gamma_{k''j}$ ($k''\neq k'$) are set to zero) - would provide an optimum of the problem (21,20) for a fixed $\hat{\lambda}$.

Step 4 (monotonic convergence of the DBMR-algorithm, computational iteration cost) According to the above Step 2 and Step 3 of this Proof, in every step (I) the problem can be solved via the iterative optimisation procedure switching between the optimisations for fixed iterated parameter values $\hat{\Gamma}^{(I)}$ (in Step 3) and $\hat{\lambda}^{(I)}$ (in Step 2). Iterative repetition of these two steps - starting at some arbitrarily chosen value $\hat{\Gamma}^{(1)}$ or $\hat{\lambda}^{(1)}$ in the first algorithm iteration - will result in a monotonic increase of the respective function value $\hat{l}^{(I)}$ when I increases (i.e., $\hat{l}^{(I)} < \hat{l}^{(I+k)}$, where $k \geq 1$). Since the overall problem (21,19,20) is bounded from above with zero and is defined on a bounded domain, this iterations will monotonically converge to some local maximum of the function (21) - dependent on the initial choice of the iteration parameters $\hat{\Gamma}^{(1)}$ or $\hat{\lambda}^{(1)}$.

Computational cost of this algorithm consists of the cost $\mathcal{O}(S)$ (for creating a matrix N with $N_{ij} = \sum_{s=1}^{S} \chi(Y(s) = y_i)\chi(X(s) = x_j)$ from the observational data) and the computational iteration complexity of $\mathcal{O}(K \cdot \min\{mn, S\})$ for analytical computation of the optima (25) and (23) in every iteration of the DBMR-algorithm, as derived in the Step 2 and the Step 3 above.

Step 5 (asymptotic uncertainty estimator for the reduced model) Following the same line of argument as introduced in the proof of the **Lemma 1** (i.e., deploying the Bayes theorem to express the logarithm of the posterior reduced parameter log-likelihood through the reduced log-likelihood $\hat{l}$ and expanding the obtained logarithm with Taylor series up to the second order), we obtain the expression for the asymptotic marginal variance of the $\hat{\lambda}^*$:

$$\mathrm{Var}\left\{\mathbb{P}\left[\hat{\lambda}_{ik} | \hat{\lambda}^*, \hat{\Gamma}^* X, Y\right]\right\} \quad = \quad \frac{\hat{\lambda}^*_{ik}(1 - \hat{\lambda}^*_{ik})}{\sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\hat{\Gamma}^*_{kj}}, \tag{26}$$

when all other parameters and the observation data are fixed.

Step 6 (optimal reduced model in the least-biased situation)

As shown in the proof of the **Lemma 2** above, the least biased prior for the data matrix N is the uniform prior of the form

$$N_{ij} \quad = \quad \frac{S}{nm}, \quad \text{for all } i, j, \tag{27}$$

that leads to a maximum log-likelihood estimator $\Lambda^*_{ij} = 1/m$. Then, substituting this estimator into (21) one can see that maximisation of the reduced log-likelihood function (21) becomes equivalent to a minimisation of the cross entropy between the two unknown discrete distributions $\hat{\Gamma}$ and $\hat{\lambda}$. Since the absolute minimum of the cross entropy is attained for the uniformly-distributed $\hat{\lambda}$ (for any fixed distribution $\hat{\Gamma}$), it means that the optimal solution of the reduced problem (21,19,20) for the least-biased prior is given by

$$\left\{\hat{\lambda}^*\right\}_{ik} \;=\; \frac{1}{m}, \quad \text{for all } i,k. \tag{28}$$

Step 7 (expected ratio of parameter uncertainties in the least-biased situation) Substituting (27) and (28) into (26) we obtain

$$\mathbb{E}_{ik}\,\mathrm{Var}\left\{\mathbb{P}\left[\hat{\lambda}_{ik}|\hat{\lambda}^*,\hat{\Gamma}^*X,Y\right]\right\} \;=\; \frac{K(m-1)}{Tm^2}, \tag{29}$$

Dividing the expectation of the left side of (14) with (29) we obtain (22). □

3 Imposing a priori information on DBMR

When dealing with real-life applications, it is also important to have an option for adjusting a set of collective variables according to a physical intuition or some prior knowledge [5]. For example, one could have some prior physical information that certain dimensions of the original problem have a higher relevance for the dynamics then some other physically less-relevant dimensions. This is especially relevant for very short data series with small S (as in the medical Example 2 below) - due to an imminent danger of "overfitting" the data through a model with a large number of free adjustable parameters. One straightforward way of resolving this problem would be in deploying some ad hoc smoothing or sparsifying strategies (e.g., Ridge- or LASSO-regularizations) that would add additional convex constraints to (19,20) or some concave penalty terms to (21). However, because of the particular analytical structure of the obtained non-concave optimisation problems it is more

appropriate to use the non-concave regularization strategies that were systematically derived for this type of clustering problems, e.g., the regularization methodologies that allow to impose a priori available expert information cast into a form of a network/graph on clustering algorithms in optimisational formulation [7].

Defining the original categories $x = \{x(1), x(2), \ldots, x(n)$ as the edges E of a graph G, in many situations we will be able to formulate the a priori available physical information as an $n \times n$ matrix of weights W_G for the vertices V connecting the edges of this graph (please see [7] for particular examples). Then, since the DBMR optimisation problem (21) has a form of the clustering problem, we can deploy a graph-based regularisation methodology introduced in [7] in order to impose this a priori information on the problem of finding an optimal reduced Bayesian model. This will result in the following maximisation problem:

$$\hat{L} \;\geq\; \hat{l} \geq \hat{l}^{D_G,\epsilon} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{K} N_{ij}\hat{\Gamma}_{kj}\log\left(\hat{\lambda}_{ik}\right) - \epsilon^2 \sum_{j_1,j_2=1}^{n}\sum_{k=1}^{K} D_G(j_1,j_2)\hat{\Gamma}_{kj_1}\hat{\Gamma}_{kj_2} \rightarrow \max_{\hat{\lambda},\hat{\Gamma}}, \quad (30)$$

where $D_G = P - 2W + Q$ (with diagonal matrices $P_{uu} \equiv \sum_{v|(v,u)\in E} W_{v,u}$ and $Q_{uu} \equiv \sum_{v|(u,v)\in E} W_{u,v}$) is kernel distance matrix of the graph. Computationally, this results in adding an additional penalty term to the right-hand side of (3), which practically means that the analytically-computable explicit formulas from Steps 1 and 2 in the original DBRM algorithm need to be substituted by solutions of sparse quadratic programming problems and resulting in the overall iteration cost of $\mathcal{O}\left(mK + n(K-1)\log[n(k-1)]\right)$ (pseudocode description of the *DBMR-graph* algorithm can be found in the Section 5 below). More details on imposing available information on clustering methods can be found in [7], a practical application of this information-imposing clustering method to the analysis of relatively short data series with a small n is given in the breast cancer diagnostics Example 2 from the main manuscript.

4 Description of the Bayesian Model Reduction algorithms introduced in the main manuscript.

Exact iterative log-likelihood maximisation First of all, we observe that for any fixed $\hat{\lambda}$, the original exact log-likelihood maximisation problem (18,19,20) can be decomposed into n independent optimisation problems of the form

$$\hat{L}_j = \sum_{i=1}^{m} N_{ij} \log \left(\sum_{k=1}^{K} \hat{\lambda}_{ik} \hat{\Gamma}_{kj} \right) \rightarrow \max_{\hat{\lambda}, \hat{\Gamma}_{1j}, \ldots, \hat{\Gamma}_{Kj}}, \tag{31}$$

$$\hat{\Gamma}_{kj} \geq 0, \quad \sum_{k=1}^{K} \hat{\Gamma}_{kj} = 1, \text{ for all } k, j, \tag{32}$$

and $\hat{L} = \sum_{j=1}^{n} \hat{L}_j$. For every $j = 1, \ldots, n$ the problem (31, 32) is a concave maximisation problem on a simplex domain - and can be approached with standard methods of constrained convex optimisation [13] with the iteration cost of $\mathcal{O}\left((K-1)^3\right)$.

For a fixed $\hat{\Gamma}$ the original exact log-likelihood maximisation problem (18,19) is also a concave maximisation problem on a simplex domain - and can be solved with the iteration cost of $\mathcal{O}\left((m-1)^3 K^3\right)$. Then, it is straightforward to validate that in order to achieve a monotonic maximisation of the exact log-likelihood $\hat{L}$, it would be enough to iterate the procedure where only one iteration step is performed for each of these concave maximisations problems. Then, the overall iteration cost for a full iterative optimisation of the original exact log-likelihood maximisation problem will be $\mathcal{O}\left((m-1)^3 K^3 + n(K-1)^3\right)$. This algorithmic procedure was deployed in order to obtain the exact log-likelihood optima that were used to create the Figure 1.c) from the main manuscript. Pseudocode description of this algorithmic procedure is provided below:

Algorithm 1 (direct iterative maximisation of the exact log-likelihood $\hat{L}$)

Choose a random $\hat{\Gamma}^{(0)}$,

compute $\hat{\lambda}^{(I)}$ *from one iteration of the problem (18,19) for a fixed* $\hat{\Gamma}^{(0)}$

set $I = 1$.

Do until $\|\hat{L}(\Gamma^{(I)}, \hat{\lambda}^{(I)}) - \hat{L}(\Gamma^{(I-1)}, \hat{\lambda}^{(I-1)})\|$ ***becomes less then a tolerance threshold****:*

Step 1:

for *j=1,. . . ,n*

for a given j, compute $\hat{\Gamma}_{1j}^{(I)}, \ldots, \hat{\Gamma}_{Kj}^{(I)}$ *from one iteration of the problem (31, 32) with a fixed* $\hat{\lambda}^{(I-1)}$

end

Step 2: compute $\hat{\lambda}^{(I)}$ *from one iteration of the problem (18,19) for a fixed* $\hat{\Gamma}^{(I)}$

$I = I + 1$.

Algorithm 2 (DBMR)

Direct Bayesian Model Reduction Algorithm (DBMR-Algorithm)

Choose random $\hat{\lambda}^{(0)}$, set $I = 0$.

Set $\Gamma_{kj}^{(0)}$ *to 1 if* $k = \underset{k'}{\operatorname{argmax}} \sum_{i=1}^{m} N_{ij} \log(\hat{\lambda}_{ik'}^{(0)})$ *and else to 0 - for all j and k.*

Do until $\|\hat{\mathrm{l}}(\Gamma^{(I)}, \hat{\lambda}^{(I)}) - \hat{\mathrm{l}}(\Gamma^{(I-1)}, \hat{\lambda}^{(I-1)})\|$ ***becomes less then a tolerance threshold****:*

Step 1: set $\hat{\lambda}_{ik}^{(I+1)} = \frac{\sum_{j=1}^{n} N_{ij} \Gamma_{kj}^{(I)}}{\sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij} \Gamma_{kj}^{(I)}}$ for all i, k

Step 2: set $\Gamma_{kj}^{(I+1)} = 1$ *if* $k = \underset{k'}{\operatorname{argmax}} \sum_{i=1}^{m} N_{ij} \log(\hat{\lambda}_{ik'}^{(I+1)})$

and else $\Gamma_{kj}^{(I+1)} = 0$ *- for all* j, k *.*

$I = I + 1$.

Algorithm 3 (DBMR-graph)

Algorithm 3 (DBMR-graph, with the imposed a priori informotation in the graph form)

Choose a random $\hat{\Gamma}^{(0)}$,

compute $\hat{\lambda}^{(I)}$ *from one iteration of the problem (18,19) for a fixed* $\hat{\Gamma}^{(0)}$

set $I = 1$.

Do until $\|\hat{l}^{D_G,\epsilon}(\Gamma^{(I)}, \hat{\lambda}^{(I)}) - \hat{l}^{D_G,\epsilon}(\Gamma^{(I-1)}, \hat{\lambda}^{(I-1)})\|$ ***becomes less then a tolerance threshold****:*

Step 1: compute $\hat{\Gamma}^{(I)}$ *by solving the sparse quadratic problem (30) subject to (20) for a fixed* $\hat{\lambda}^{(I-1)}$

Step 2: set $\hat{\lambda}_{ik}^{(I)} = \frac{\sum_{j=1}^{n} N_{ij}\Gamma_{kj}^{(I)}}{\sum_{i=1}^{m}\sum_{j=1}^{n} N_{ij}\Gamma_{kj}^{(I)}}$ for all i, k

$I = I + 1$.

---

5 Methodological comparison of the Bayesian Model Reduction methods introduced in the manuscript to the Probabilistic Latent Semantic Analysis (PLSA)

In the following we present a list of methodological issues that arise from comparison of the model reduction framework proposed in this manuscript to the Probabilistic Latent State Analysis (PLSA) approaches that were developed in the area of information retrieval for analysis and reduction of documents and texts [10, 11, 4].

PLSA: formulation, EM optimisation, limitations in a context of large-scale dynamical systems

PLSA is a methodology developed in the areas of information retrieval and mathematical linguistics in order to perform a semantic analysis and to identify the latent semantic spaces of texts and documents [10, 11]. Adopting the notation used so far, we can write the PLSA model (e.g., equations 1 and 2

from [10] and the equation 3 from [11]) as:

$$\mathbb{P}\left[X(s)=x(j) \text{ and } Y(s)=y(i)\right]= \\ =\mathbb{P}\left[X(s)=x(j)\right]\sum_{k=1}^{K}\mathbb{P}\left[Y(s)=y(i)|\hat{X}(s)=\hat{x}(k)\right]\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)|X(s)=x(j)\right]. \quad (33)$$

An equivalent symmetric form of this model (equation 4 from [10] and equation 5 from [11]) has the form:

$$\mathbb{P}\left[X(s)=x(j) \text{ and } Y(s)=y(i)\right]= \\ =\sum_{k=1}^{K}\mathbb{P}\left[X(s)=x(j)|\hat{X}(s)=\hat{x}(k)\right]\mathbb{P}\left[Y(s)=y(i)|\hat{X}(s)=\hat{x}(k)\right]\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)\right], \quad (34)$$

with $x=\{x(1),\ldots,x(n)\}$ being some fixed set of n documents and $y\,\{y(1),\ldots,y(m)\}$ being a predefined collection of m different words (a vocabulary).

Please note that by this construction used in [10, 11], both sets x and y are not assumed to build complete coverages of the respective realisation spaces - i.e., $\{x(1),\ldots,x(n)\}$ should not necessary build a full set of all documents and $\{y(1),\ldots,y(m)\}$ should not be a full vocabulary of all words. This fact implies that in general, for any s it is true that $0\leq\sum_{j=1}^{n}\mathbb{P}\left[X(s)=x(j)\right]\leq 1$ and $0\leq\sum_{i=1}^{m}\mathbb{P}\left[Y(s)=y(i)\right]\leq 1$. Another implication of this fact is that if we would cast this model into the particular context of discrete Markov processes in time - by setting $x\equiv y$, $Y(s)\equiv X(s+1)$ (with s being a discrete time index) - and shall be using the PLSA model as an iterative dynamic model to propagate the probability density from some initial value $X(0)$ for $s=1,2,...,S$ (as in the molecular dynamics application Example 1 from the main manuscript), then it is straightforward to validate that the obtained vectors will not remain the probability densities - i.e., the resulting propagation of the Markov chain in time will not be probability-preserving.

Representing the available data X and Y in the form of the *term frequency matrix* $N_{ij}=\sum_{s=1}^{S}\chi(Y(s)=y_i)\chi(X(s)=x_j)$ (with χ being an indicator function), in PLSA framework one would like to seek for the values of conditional probabilities $\mathbb{P}\left[Y(s)=y(i)|\hat{X}(s)=\hat{x}(k)\right]$ and $\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)|X(s)=x(j)\right]$

that would maximise the following log-likelihood function (equation 4 from [11]):

$$\mathcal{L} = \sum_{j=1}^{n} N_i \left[\pi_X(j) + \right.$$
$$\left. + \sum_{i=1}^{m} \frac{N_{ij}}{N_j} \log \left(\sum_{k=1}^{K} \mathbb{P}\left[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)\right] \mathbb{P}\left[Y(s) = y(i) | \hat{X}(s) = \hat{x}(k)\right] \right) \right], \quad (35)$$

where $N_j = \sum_{i=1}^{m} N_{ij}$ and

$$\mathbb{P}\left[X(s) = x(j)\right] \equiv \pi_X(j), \quad \text{for all } s = 1, \ldots, S; j = 1, \ldots, n, \quad (36)$$

i.e., in addition it is implicitly assumed that $X(s)$ is i.i.d. (i.e., independent/memoryless and identically-distributed process) in s. As demonstrated in [4], this maximum log-likelihod formulation of the PLSA inference problem is equivalent to the Nonnegative Matrix Factorisation method (NMF) in the Kulback-Leibler norm [2][4].

However, a direct maximisation of the log-likelihood function (35) is hampered by the fact that due to a specific form of the non-linearity in the right-hand side of the expression (being a logarithm of the sum over k of the nonlinear expressions $\mathbb{P}\left[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)\right] \mathbb{P}\left[Y(s) = y(i) | \hat{X}(s) = \hat{x}(k)\right]$) one can not find the analytical expressions for the arguments $\mathbb{P}\left[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)\right]$ and $\mathbb{P}\left[Y(s) = y(i) | \hat{X}(s) = \hat{x}(k)\right]$ that would simultaneously maximise (35) and remain probabilities (i.e., would satisfy $0 \leq \mathbb{P}\left[\hat{X}(s) = \hat{x}(k) | X(s) = x(j)\right] \leq 1$ and $0 \leq \mathbb{P}\left[Y(s) = y(i) | \hat{X}(s) = \hat{x}(k)\right] \leq 1$). Application of the standard numerical optimisation methods for this problem would result in the prohibitively-large computational complexity of the iterations in the optimisation procedure[5], scaling as $\mathcal{O}\left((m+n)^3 K^3\right)$.

[4]Equivalence of the two optimisational formulation does not imply the equivalence of the algorithmic procedures: e.g., in context of the large-scale dynamical systems discussed above, the NMF algorithm implementation would rely on the availability of the empirical frequency estimator of the full relation matrix Λ. As shown in the Lemma 1 and 2, these estimates will be subject to the uncertainty that may grow exponentially with the physical problem dimension.

[5]This is explained by the fact that this optimisation problem has $(m+n)K$ arguments and the limiting step in the computations based on gradient-ascent methods would be the inversion of the respective Jacobian (having a dimension of $(m+n)K$ times $(m+n)K$). The cost of this operation scales cubically for general unstructured matrices [13].

So, in order to obtain the optimisers, PLSA deploys the Expectation Maximization (EM) algorithm, that - instead of a direct maximisation of (35) - maximises (35) over its approximate lower bound that is obtained by deploying the Jensen's inequality to the expectation of the log-likelihood function (35) taken with respect to the latent process $\hat{X}(s)$.

$$\mathcal{E}_{\hat{X}}\left(\mathcal{L}\right)=\sum_{j=1}^{n}\sum_{i=1}^{m}N_{ij}\sum_{k=1}^{K}\pi_{\hat{X}}(k,i,j)\log\left(\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)|X(s)=x(j)\right]\mathbb{P}\left[Y(s)=y(i)|\hat{X}(s)=\hat{x}(k)\right]\right), \tag{37}$$

where

$$\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)|X(s)=x(j)\text{ and }Y(s)=y(i)\right]\equiv\pi_{\hat{X}}(k,i,j),\quad\text{for all }s,j,i, \tag{38}$$

i.e., it is also implicitly assumed that $\hat{X}(s)$ is i.i.d. (i.e., independent/memoryless and identically-distributed process) in s for any fixed combination of $X(s)$ and $Y(s)$. The fact that the numerical inference procedures in PLSA are not based on the direct maximisation - but are actually maximising the lower bound approximation (obtained with the help of the Jensen's inequality) is a standard part of the EM and is not mentioned explicitly in the PLSA literature. However, it becomes clear when for example comparing the formula 4 and 7 in [11].

In contrast to the original log-likelihood function (35), its lower bound approximation (38) allows a direct analytical computation of the extrema arguments (by taking the function derivatives, setting them to zero and solving the obtained equations analytically). This is performed in the E-step (for the hidden process probabilities $\pi_{\hat{X}}(k,i,j)$) and in the M-step (for the probability vector $\pi_X(j)$ and for the conditional probability matrices $\mathbb{P}\left[\hat{X}(s)=\hat{x}(k)|X(s)=x(j)\right]$ and $\mathbb{P}\left[Y(s)=y(i)|\hat{X}(s)=\hat{x}(k)\right]$). Again, the reason allowing this was given by the fact that deploying the Jensen's inequality allowed approximating the analytically-intractable nonlinearity (logarithm of the sum) through the analytically-tractable one (sum of the logarithms).

Compared with the direct numerical optimisation of (35) that has a computational iteration complexity of $\mathcal{O}\left(\left(m+n\right)^{3}K^{3}\right)$, the main advantage of this iterative EM procedure is its linear complex-

ity scaling - with iteration cost scaling as $\mathcal{O}\left(K \min\{mn, S\}\right)$. However, as discussed above, this advantage was coming at a price. The following list summarises the "cost items" that were needed to achieve this advantage - and that might hamper the applicability of the PLSA in context of the large-scale dynamical systems:

(i) introduction of the new optimisation arguments $\pi_X(j)$ and $\pi_{\hat{X}}(k,i,j)$ (that come in addition to the original quantities of interest $\mathbb{P}\left[\hat{X}(s) = \hat{x}(k)|X(s) = x(j)\right]$ and $\mathbb{P}\left[Y(s) = y(i)|\hat{X}(s) = \hat{x}(k)\right]$) means increasing the overall number of free parameters - and the required program memory - from $\mathcal{O}\left(K(m+n) + \min\{mn, S\}\right)$ to $\mathcal{O}\left(K(m+n) + Kmn + n + \min\{mn, S\}\right)$;

(ii) construction of the EM for PLSA requires imposing additional i.i.d. assumptions (36) and (38) - and introduces a bias if these assumptions are not fulfilled for the underlying dynamical system;

(iii) as was shown in the Lemma 1 and 2 above, in context of the large-scale dynamical systems the number of categorical problem dimensions may grow exponentially with the physical dimension of the underlying system. Following the same line of argument as deployed in the Lemma 1 and 2 on can demonstrate that this will be resulting in the exponential growth of uncertainty for the respective EM-estimators of the additional variables $\pi_X(j)$ and $\pi_{\hat{X}}(k,i,j)$. Since all of the variables are iteratively coupled in the estimation procedure, this will also result in the additional growth of estimation uncertainty for the original values of interest $\mathbb{P}\left[\hat{X}(s) = \hat{x}(k)|X(s) = x(j)\right]$ and $\mathbb{P}\left[Y(s) = y(i)|\hat{X}(s) = \hat{x}(k)\right]$;

(iv) as explained above, EM algorithm is optimising the lower bound approximation (37) of the original log-likelihood function (35), meaning that thereby obtained quantities of interest $\mathbb{P}\left[\hat{X}(s) = \hat{x}(k)|X(s) = x(j)\right]$ and $\mathbb{P}\left[Y(s) = y(i)|\hat{X}(s) = \hat{x}(k)\right]$ might not necessary also be the global optimisers of the original log-likelihood. And quantifying the deviations between the optima of (35) and the solutions obtained by the EM algorithm is not straightforward due to the nonlinearity and nonconvexity of both functions.

(v) In context of application to the large-scale dynamical systems, matrices obtained with EM may be difficult to interpret and to understand since they provide only 'fuzzy' and probabilistic - and not deterministic - relations between the original and the reduced dimensions of the underlying problem.

(vi) EM algorithm for PLSA does not allow a direct possibility to impose an a priori available "physical" information on the problem (e.g., an information that certain dimensions of the original problem have a higher relevance for the dynamics then some other physically less-relevant dimensions).

Comparison of the PLSA methodology with the three algorithms introduced in this manuscript

Below we provide a list of items that arise when comparing the PLSA with the Bayesian Model Reduction (BMR) algorithms introduced in the manuscript:

a) BMR algorithm 1 introduced above allows a direct explicit optimisation of the exact log-likelihood $\hat{L}$, whereas the *DBMR*, *DBMR+graph* and PLSA-EM all maximise different lower bound approximations of the exact log-likelihood (please see the Figure S6 for a graphic representation).

b) BMR algorithms do not require introduction and iterative estimation of the additional model parameters $\pi_X(j)$ and $\pi_{\hat{X}}(k,i,j)$ (that are crucial for the PLSA). It means that BMR algorithms do not rely on - and are not biased by - the additional strong i.i.d. assumptions (36) and (38) that are imposed in the PLSA. Another implication of the much smaller number of optimisation parameters in the case of the DBMR are much more favourable memory requirements (scaling as $\mathcal{O}\left(K(m+n)+\min\{mn,S\}\right)$) and a more favourable computational scaling for single iterations (see the Figure S6). It also allows to avoid an additional estimation uncertainty and reduces the risk of overfitting - that can, for example, be reflected in a more favourable

(i.e., smaller) values of information-theoretical measures like AIC and BIC measured for BMR results (please see the Figure S5).

c) Numerical tests performed on large ensembles of randomly generated data sets revealed that the average number of iterations required until reaching the same tolerance (measured in terms of the normalized log-likelihood $\frac{1}{mn}\hat{L}$) for different combinations of dimensions m and n for the PLSA model with the EM algorithm grows rapidly with problem dimensions (please see the Figure 1 from the main manuscript). In contrast, application of the DBMR to the same problems with the same tolerance of the normalized log-likelihood results in the average number of iterations that practically does not change with problem dimensions. This strong dimension-dependence of the number of EM iterations is then further reflected in the dimension-dependence of the overall computational time until convergence: deploying standard statistical tools of polynomial regression fitting and discrimination [15] one obtains that the statistically-optimal fit of the red surface (corresponding to the PLSA) from the Figure 1.b) in the main manuscript is given by a polynomial of the third degree - whereas the green surface from the same Figure 1.b) (corresponding to the DBMR results) is optimally fitted by a function that is linear in m and n. Extrapolation to the typical physical problem sizes (e.g., $m = n = 10^5$, $K = 2$) that, e.g., emerge in biophysical applications like the Markov State Model inference based on Molecular Dynamics simulations of medium-size protein molecules, indicates that it would require approximately 1450 years of computations with the PLSA method on a single PC. In contrast, application of the DBMR algorithm to the same data under the same conditions and settings requires 33 minutes on a single LapTop-PC.

d) As can be seen from the Figure 1.c) in the main manuscript, the average relative log-likelihood difference between the results of exact iterative log-likelihood optimisation (Algorithm 1) and the DBRM results (obtained under the same conditions) converges to zero exponentially in

m. This implies that for realistic high-dimensional applications the log-likelihood difference between the reduced models obtained with the DBRM algorithm and with the optimisation of the exact log-likelihood will become negligible - meaning that the reduced models obtained with the DBRM algorithm will have essentially the same posterior probability for explaining the observed full data as the exact reduced models.

e) Theorem 1 above provides very straightforward analytical formula for quantification of uncertainty of the reduced Bayesian models - and provides an understanding of how does the uncertainty change when changing the ratio of the reduced and the original problem's dimensions. Because of the additional parameters $\pi_X(j)$ and $\pi_{\hat{X}}(k,i,j)$, uncertainty quantification for the PLSA and related approaches is much less straightforward.

f) DBMR provides deterministic ($\{0\}/\{1\}$) relations between the original and the reduced dimensions of the analyzed problem - making understanding and interpretation of the obtained reduced relation models much easier.

g) DBMR-graph algorithm allows a robust and numerically-scalable (please see Figure S6) incorporation of a priori available physical information - if this information can be cast into a form of a weighted graph with the kernel weights matrix D_G.

6 Additional figures

- **Figure S1** *Left*: Ramachandran plot for the MD simulation of 10-alanine peptide in water and its decomposition into three categorical states. *Right*: representation with n categories for the whole polypeptide molecule with N residues.

- **Figure S2** Values of $\hat{\mathrm{l}}_i$ in the optimal solution of the reduced problem (8,6-7) obtained for different numbers of colvars (or *collective causality boxes*) K (on x-axis) when process Y_i are

the Ramachandran dynamics of residues number one (i=1) to eight (i=8).

- **Figure S3** Three optimal colvars (or *collective causality boxes*) obtained for $K = 3$ with process Y_4 being the Ramachandran dynamics of residuum number four.

- **Figure S4** Probabilities for different proportions of the chain in the same local Ramachandran state 1 to 3 from Fig.1 in the main manuscript. 100% means that all of the residues in the chain are in this Ramachandran state, 0% means that there is not a single residuum in this state. The blue line indicates the values of this distribution obtained from the full MD simulation data and the boxplot shows the probability distribution and its 95% confidence intervals obtained from the completely local reduced model (i.e. with 100% of local causality boxes). Red points denote the statistical outliers of the reduced model (meaning that they are outside of the 99% confidence interval). This plot should be compared to the Fig. 4 - implying that the 3% of non-local causality boxes used in the Fig. 4 are essential to match the statistics of reduced model with the full MD-data.

- **Figure S5** Practical comparison of PLSA methods and the approaches introduced in this manuscript - the direct $\hat{L}$-optimisation algorithm and the $\hat{l}$-optimisation with the DBMR-algorithm. Comparison of the optimal colvars obtained by different algorithms with $K = 2$ for the MD data in Example 1 from the main manuscript.

- **Figure S6** Graphic representation of the three algorithms introduced in the manuscript in comparison with the PLSA methodology from [10, 11].

References and Notes

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

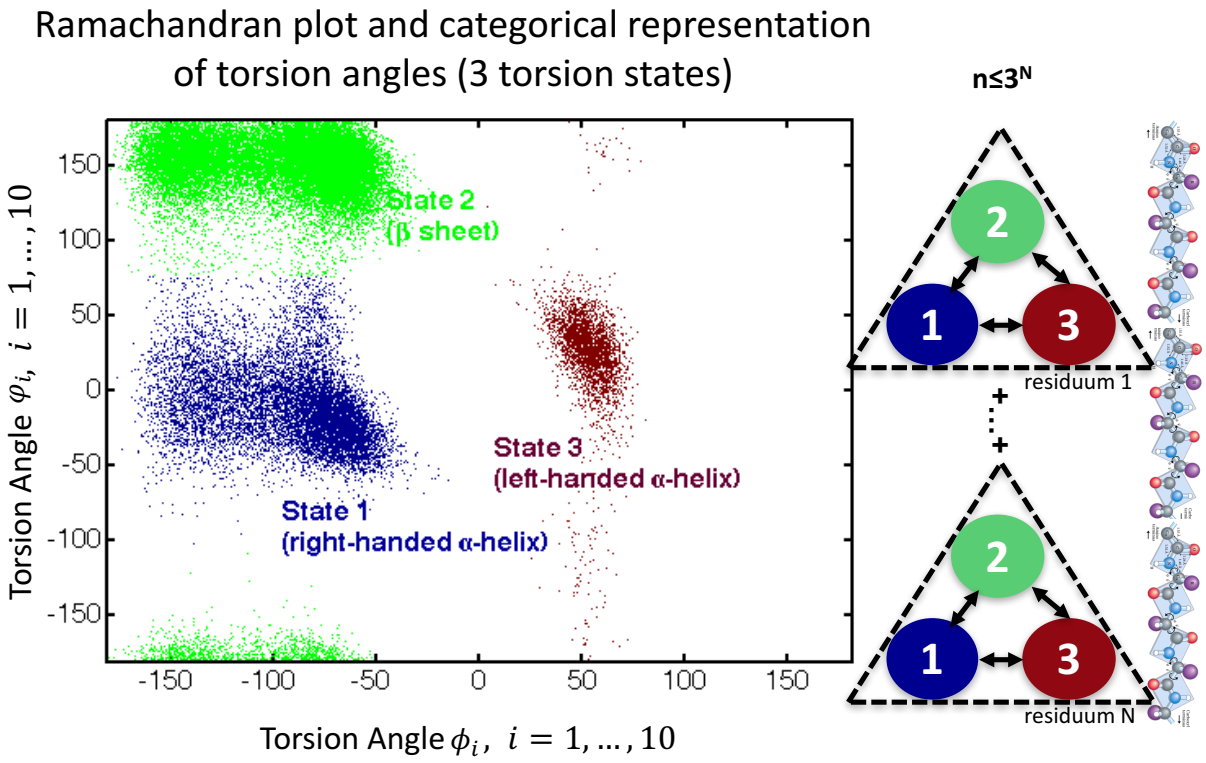

Figure 1: *Left*: Ramachandran plot for the MD simulation of 10-alanine peptide in water and its decomposition into three categorical states. *Right*: representation with n categories for the whole polypeptide molecule with N residues.

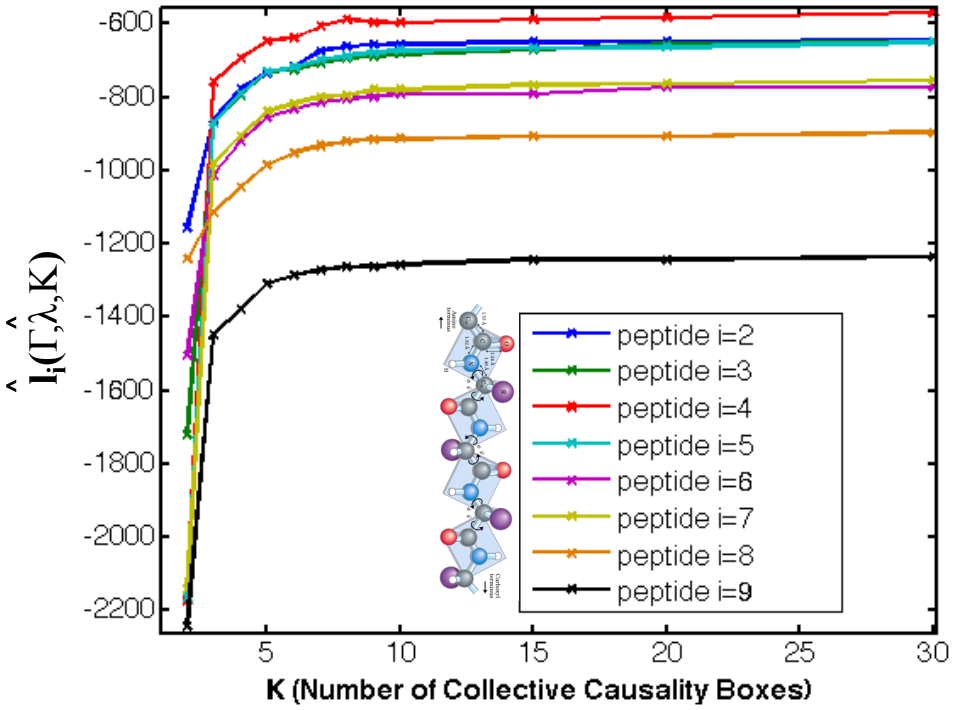

Figure 2: Values of $\hat{\mathrm{l}}_i$ in the optimal solution of the reduced problem (8,6-7) obtained for different numbers of colvars (or *collective causality boxes*) K (on x-axis) when process Y_i are the Ramachandran dynamics of residues number one (i=1) to eight (i=8).

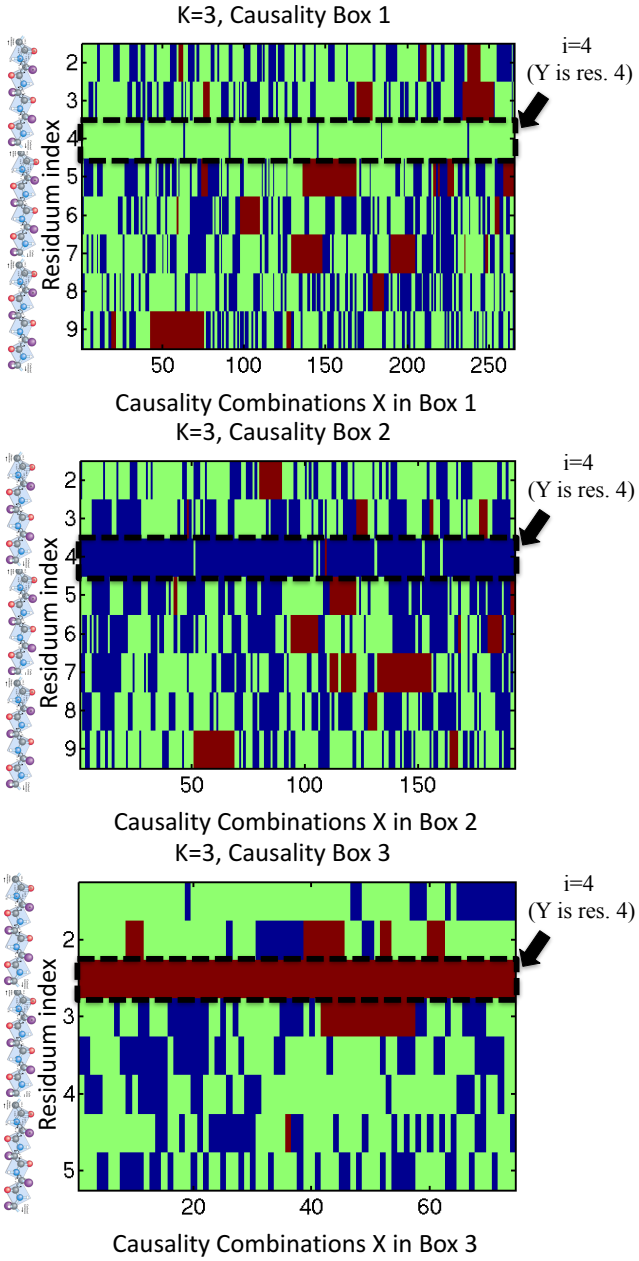

Figure 3: Three optimal colvars (or *collective causality boxes*) obtained for $K = 3$ with process Y_4 being the Ramachandran dynamics of residuum number four.

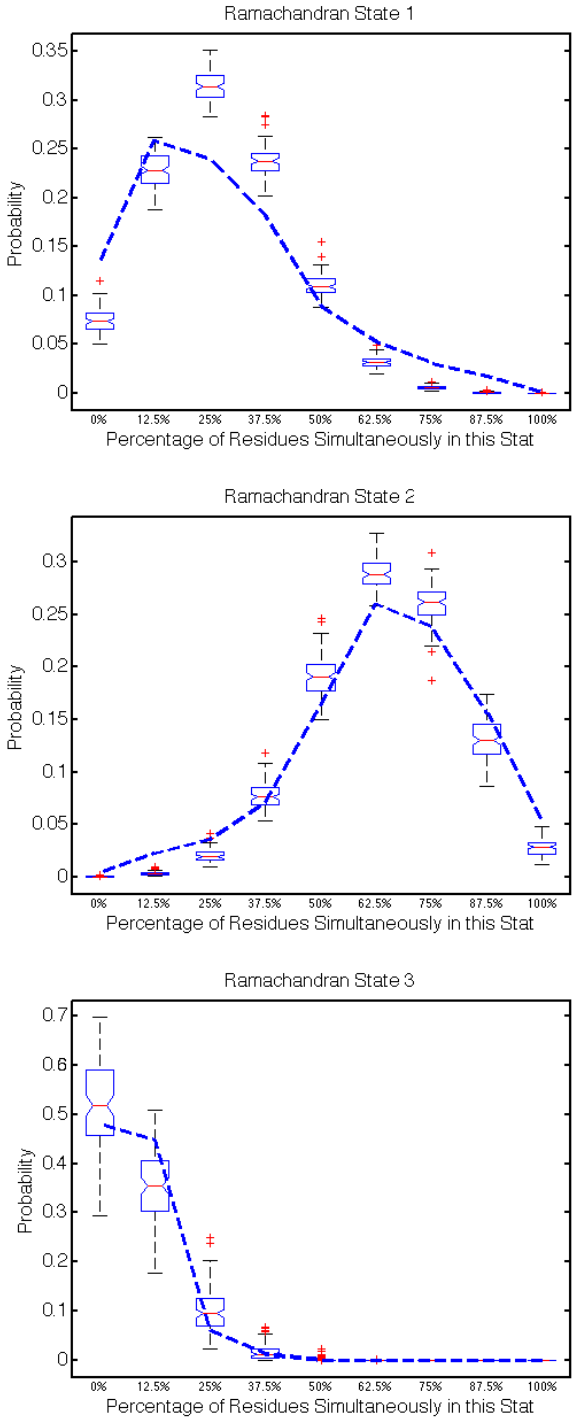

Figure 4: Probabilities for different proportions of the chain in the same local Ramachandran state 1 to 3 from Fig.1 in the main manuscript. 100% means that all of the residues in the chain are in this Ramachandran state, 0% means that there is not a single residuum in this state. The blue line indicates the values of this distribution obtained from the full MD simulation data and the boxplot shows the probability distribution and its 95% confidence intervals obtained from the completely local reduced model (i.e. with 100% of local causality boxes). Red points denote the statistical outliers of the reduced model (meaning that they are outside of the 99% confidence interval). This plot should be compared to the Fig. 4 - implying that the 3% of non-local causality boxes used in the Fig. 4 are essential to match the statistics of reduced model with the full MD-data.

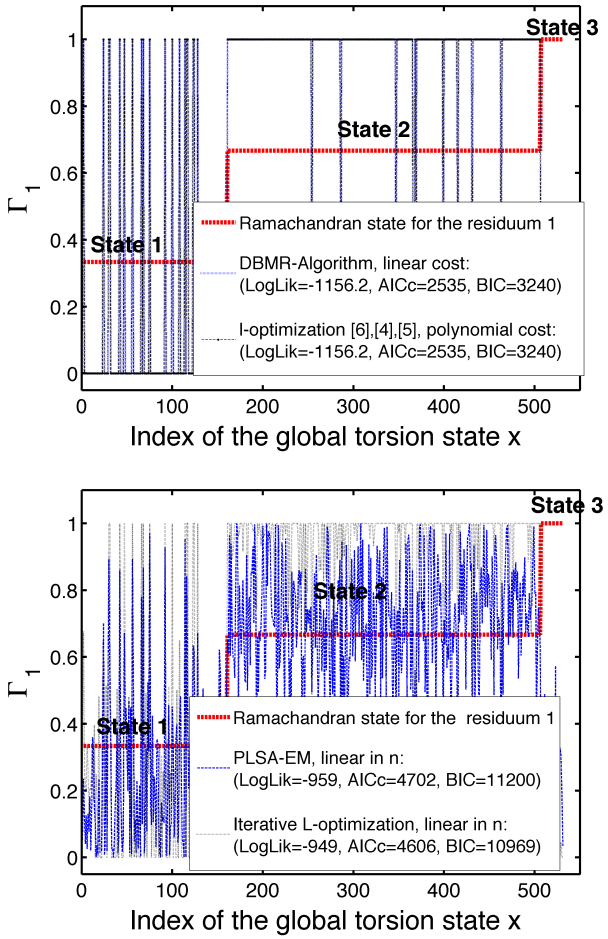

Figure 5: Practical comparison of PLSA methods and the approaches introduced in this manuscript - the direct $\hat{L}$-optimisation algorithm and the $\hat{l}$-optimisation with the DBMR-algorithm. Comparison of the optimal colvars obtained by different algorithms with $K = 2$ for the MD data in Example 1 from the main manuscript.

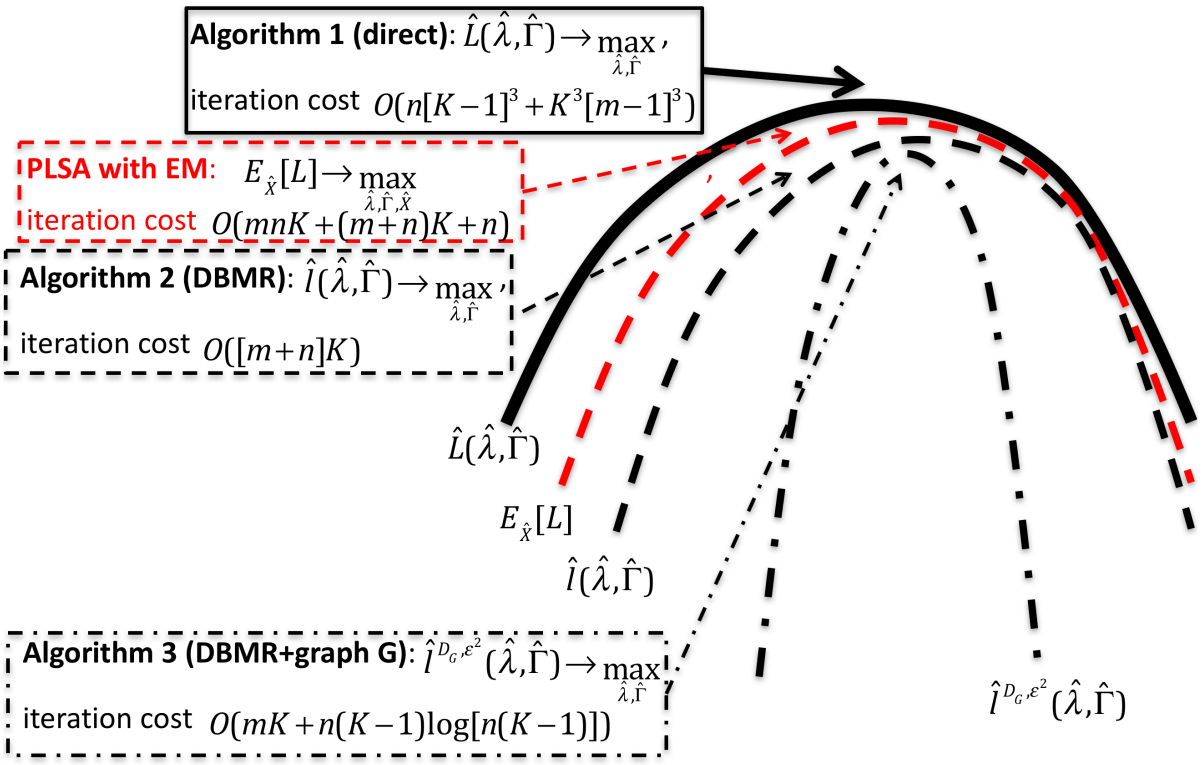

Figure 6: Graphic representation of the three algorithms introduced in the manuscript (black) in comparison with the PLSA methodology from [10, 11] (red).

[2] Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155 – 173, 2007.

[3] K.P. Burnham and D.R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer-Verlag, 2002.

[4] Chris Ding, Tao Li, and Wei Peng. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 342–347. AAAI Press, 2006.

[5] G. Fiorin, M. Klein, and J. Henin. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.*, 111(22-23):3345–3362, 2013.

[6] H. Gardiner. *Handbook of stochastical methods*. Springer, Berlin, 2004.

[7] S. Gerber and I. Horenko. Improving clustering by imposing network information. *Science Advances (AAAS)*, 1(7):e1500163, 2015.

[8] D. Giannakis, A. Majda, and I. Horenko. Information theory, model error, and predictive skill of stochastic models for complex nonlinear systems. *Physica D: Nonlinear Phenomena*, 241(20):1735–1752, 2012.

[9] D. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1-3):404–425, 1992.

[10] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[11] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[12] P. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[13] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006.

[14] Ch. Schütte and M. Sarich. *Metastability and Markov State Models in Molecular Dynamics: Modeling, Analysis, Algorithmic Approaches*. American Mathematical Society, Courant Lecture Notes, 2013.

[15] G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.