# Relative Principal Components Analysis: Application to Analyzing Biomolecular Conformational Changes

*Mazen Ahmad*[*a], *Volkhard Helms*[b], *Olga V. Kalinina* [a] *and Thomas Lengauer* [a]

[a]Computational Biology Research Group, Max Planck Institute for Informatics, Saarland Informatics Campus, Campus E1 4, 66123 Saarbrücken, Germany.

[b]Center for Bioinformatics, Saarland University, 66123 Saarbrücken, Germany.

[*]Author to whom correspondence should be addressed: mahmad@mpi-inf.mpg.de

**Abstract**

A new method termed *"Relative Principal Components analysis*" (RPCA) is introduced that extracts optimal relevant principal components to describe the change between two data samples representing two macroscopic states. The method is applicable in all areas of data-driven science. Mining of the components is based on a unified physical framework for studying the change between two macroscopic states, represented by two data sets. To demonstrate the applicability of RPCA, we analyzed the energetically relevant conformational changes of the protein HIV protease upon binding to different drug molecules. In this case, the objective function of the RPCA method, namely the Kullback-Leibler divergence, provides a sound thermodynamic foundation for the analysis of the binding process to characterize both the collective and the locally relevant conformational changes.

# 1 INTRODUCTION

Studying the transitions and differences between multiple states populated by a dynamic system is a central topic in different fields including chemistry, physics, biology, machine learning and all of data-driven science. A typical task is to uncover how macroscopic changes of the dynamic system are related to the features (variables) that describe its microscopic individuals (instances). Two examples of such microscopic features would be the genetic sequences of a virus taken from snapshots during the course of evolution or the spatial conformations of two biomolecules when they bind to each other. The relationship between the "microscopic" factors of a system and the change of its macroscopic states requires the definition of an appropriate objective function for quantifying the change of the "macroscopic" state of the system. Such a rigorous definition of changes of the macroscopic state of a system in terms of its microscopic features is available for physical systems whose thermodynamic quantities can be measured or computed. For example, the change of free energy (a scalar value) is a suitable quantity to characterize macroscopic changes in physical, chemical and biochemical systems. However, in other areas of data-driven science, such a rigorous definition and quantification of macroscopic changes generally does not exist. Instead, various heuristic objective functions are used in practice. Examples include divergence measures from information theory[1] and the wide variety of objective functions which are used for prediction and feature extraction in pattern recognition[2]. Mining the factors informative for the change between two samples is of high importance and of general interest in all areas of data-driven science and is generally performed in a high-dimensional feature space. In fact, mining informative features is the central theme in a large domain of machine learning and includes methods such as dimensionality reduction[3,4], feature extraction[2], and latent variable models[5]. However, one needs to select an objective function that is appropriate for quantifying the change before applying a multivariate method to extract the informative features.

In this work, we introduce a unified framework rooted in statistical information theory and statistical mechanics[6–11] for the purpose of studying the change between two data sets representing two states. A new method termed *"Relative Principal Components Analysis*" is introduced to extract the optimal relevant principal components describing the change between two data samples (two states). This method is applicable in all areas of data-driven science. As an example, we apply RPCA for analyzing the energetically relevant conformational changes of a biomolecule (protease of the human immunodeficiency virus, HIV-1) upon binding to various ligands.

# 2 A general formalism for analyzing changes in dynamic systems based on information theory

Before going into the technical details of finding the directions in feature space that are informative of the change between two states, we first introduce a physical framework for defining and quantifying the change of dynamic systems in all areas of data-driven sciences and justify the objective function used for quantifying the macroscopic change. By a *system* we mean a collection of individuals or *instances,* e.g.

viruses or molecules, where each instance is defined via a set of (microscopic) *features*. The probability distribution of all instances together exhibits a macroscopic property, called the *macroscopic state*.

Let $\boldsymbol{x}$ be a random vector encoding the microscopic features (variables) which are necessary to identify the difference between the instances of a system of interest. A macroscopic state ($a$) of the system is defined by the probability density distribution $P(\boldsymbol{x}|a) = P_a(\boldsymbol{x})$ of $\boldsymbol{x}$ when the system is at equilibrium in this macroscopic state $a$. The macroscopic state $a$ can be changed into a new macroscopic state $b$ by perturbing the probability distribution of the microscopic instances $\boldsymbol{x}$ to $P(\boldsymbol{x}|b) = P_b(\boldsymbol{x})$. Here, a Bayesian approach is used whereby the macroscopic state is viewed as a random variable, and the conditional probability $P(\boldsymbol{x}|i) = P_i(\boldsymbol{x})$ is naturally interpreted as the equilibrium distribution of state $i$ that can be taken from a finite or discrete set. A relationship between the two macroscopic states can be obtained by applying Bayes' theorem, yielding the following equation for the probability density ratio $P_b(\boldsymbol{x})/P_a(\boldsymbol{x})$:

$$\frac{P(\boldsymbol{x}|b)}{P(\boldsymbol{x}|a)} = \frac{P_b(\boldsymbol{x})}{P_a(\boldsymbol{x})} = \frac{P(a)}{P(b)}\,\frac{P(b|\boldsymbol{x})}{P(a|\boldsymbol{x})} \tag{2.1}$$

Here, $P(b)$ and $P_a(\boldsymbol{x})$ are conditional probabilities under the (implicit) condition of the perturbing factors. Averaging the logarithm of the density ratio equation over the probability distribution of state $b$ ($P_b(\boldsymbol{x})$) yields:

$$\underbrace{\ln\left[\frac{P(a)}{P(b)}\right]}_{\Delta F} = \int P_b(\boldsymbol{x}) \underbrace{\ln\left[\frac{P(a|\boldsymbol{x})}{P(b|\boldsymbol{x})}\right]}_{U_p(x)} d\boldsymbol{x} + \underbrace{\int P_b(\boldsymbol{x}) \ln\left[\frac{P_b(\boldsymbol{x})}{P_a(\boldsymbol{x})}\right] d\boldsymbol{x}}_{D_{kl}} \tag{2.2}$$

Indeed, this is a general derivation of the Perturbation Divergence Formalism (PDF), which was previously derived for physical systems for the purpose of decomposing the change of free energy ($\Delta F$) between two macroscopic states $a$ and $b$ [6,7]:

$$\ln\left[\frac{P(a)}{P(b)}\right] = \Delta F = \langle U_p(\boldsymbol{x})\rangle_b + D_{kl}\big(P_b(\boldsymbol{x}) \parallel P_a(\boldsymbol{x})\big) \tag{2.3}$$

For physical systems, we use here the natural unit of the energy $(kT)^{-1} = 1$. $D_{kl}$ is the Kullback-Leibler (KL) divergence[12] (also termed the relative entropy[13] or the discriminant information[14]) between the probability distributions of states $a$ and $b$. Interestingly, equation (2.2) provides a purely probabilistic formalism of free energy change via a new probabilistic definition of the perturbation $U_p$ of the microscopic instances as the logarithm of the ratio of the posterior probabilities of two macroscopic states given a particular microscopic configuration $\boldsymbol{x}$:

$$U_p(\boldsymbol{x}) = \ln\left[\frac{P(a|\boldsymbol{x})}{P(b|\boldsymbol{x})}\right] \tag{2.4}$$

**Bridging the macroscopic change and the microscopic features.** The concept of the perturbation of a microscopic instance was originally introduced in statistical mechanics as an energetic quantity in order to formalize the relationship between the microscopic (atomistic) description of a physical system

and its macroscopic changes between states (free energy change)[15,16]. When considering the change between two macroscopic states, the perturbation $U_p(\boldsymbol{x})$ is a one-dimensional (microscopic) variable which has the same KL divergence (discriminant information) as the high dimensional feature $\boldsymbol{x}$[7]. In statistical inference theory[17,18], such a variable is termed a sufficient statistic and the sufficiency principles are fundamental for the field of data reduction[17,18]. A sound framework for the relationship between the change of macroscopic states of a dynamic system and its microscopic elements can be inherited from the formalism of exponential families[19,20] in statistical estimation theory.

Let us assume we have a dynamic system at equilibrium in a macroscopic reference state labeled by the so-called *natural parameters* $\boldsymbol{\lambda}$ with unknown probability density distribution $P_{\lambda_0}(\boldsymbol{x})$. The system can populate a new macroscopic state $P_\lambda(\boldsymbol{x})$ upon perturbing its original state. The principle of minimum discrimination information[21] (equivalent to the principle of maximum entropy[22]) is applied to find the new distribution $P_\lambda(\boldsymbol{x})$ via minimizing the KL divergence from the reference distribution $P_{\lambda_0}(\boldsymbol{x})$ to the new distribution $P_\lambda(\boldsymbol{x})$[21] under the constraint of a finite expected value of the sufficient statistic $\langle \boldsymbol{T}(\boldsymbol{x}) \rangle_\lambda$. The probability density distribution $P_\lambda(\boldsymbol{x})$ forms an exponential family of distributions in terms of the reference distribution:

$$P_\lambda(\boldsymbol{x}) = \exp[\boldsymbol{\lambda}^{\boldsymbol{T}} \boldsymbol{T}(\boldsymbol{x}) - \psi(\boldsymbol{\lambda})]\, P_{\lambda_0}(\boldsymbol{x})$$

$$\psi(\boldsymbol{\lambda}) = \ln\left[\int_x d\boldsymbol{x} P_{\lambda_0}(\boldsymbol{x}) \exp\left(\boldsymbol{\lambda}^{\boldsymbol{T}} \boldsymbol{T}(\boldsymbol{x})\right)\right] \tag{2.5}$$

$$D_{kl}\left(\langle \boldsymbol{T}(\boldsymbol{x}) \rangle_{\boldsymbol{\lambda}}\, ; P_\lambda(\boldsymbol{x}) \parallel P_{\lambda_0}(\boldsymbol{x})\right) = \boldsymbol{\lambda}^{\boldsymbol{T}} \langle \boldsymbol{T}(\boldsymbol{x}) \rangle_\lambda - \psi(\boldsymbol{\lambda}) \tag{2.6}$$

Here, the sufficient statistic $\boldsymbol{T}(\boldsymbol{x})$ is a vector (or a scalar) function of the microscopic configuration $\boldsymbol{x}$ of the system. The cumulant generating function $\psi(\boldsymbol{\lambda})$ depends only on the natural parameters $\boldsymbol{\lambda}$. Besides being used to formulate a theoretical framework for studying the error bound of parameter estimation[21], the formalism of exponential families plays a central role in different fields of machine learning such as generalized linear models and variational inference[3]. In statistical thermodynamics, Kirkwood introduced his thermodynamic integration (TI) equation[15] using the exponential family to "*alchemically*" interpolate two macroscopic states[15]. Indeed, the one-dimensional sufficient statistic "the perturbation" is the appropriate tool for interpolating between two macroscopic states in free energy calculations. Generally, a higher-dimensional sufficient statistic is required when studying multiple macroscopic states[17]. The sufficient statistic and the corresponding cumulant generating function in equations (2.5) are not unique. Notably, comparing equations (2.2) and (2.6) shows that the sufficient statistic can be selected such that the corresponding cumulant generating function is a function of the change of free energy between the macroscopic states of dynamical systems. For example, when selecting the negative value of the perturbation $\left(-U_p(\boldsymbol{x})\right)$ as a sufficient statistic for studying the change between two macroscopic states, the corresponding cumulant generating function is the negative of the free energy change and equation (2.5) turns into the free energy perturbation equation that is well known in statistical thermodynamics[16]. Another example is the logarithm of the density ratio which is a well-known sufficient statistic in

statistical inference, and its corresponding cumulant generating function is the relative change of free energy (see Supplementary Material S5). An interesting known relationship of an exponential family is the functional dependence of the change of $\psi(\boldsymbol{\lambda})$ (the function of the macroscopic state) on the microscopic features $\boldsymbol{x}$ that is given by the average and the covariance of the sufficient statistic:

$$\nabla\psi(\boldsymbol{\lambda}) = \langle \boldsymbol{T}(\boldsymbol{x})\rangle_{\lambda}$$
$$\nabla^2\psi(\boldsymbol{\lambda}) = \mathrm{Cov}_{\boldsymbol{\lambda}}\big(\boldsymbol{T}(\boldsymbol{x})\big) \tag{2.7}$$

The free energy thermodynamic integration equation[15] is a special case of equation (2.7) in which the collected data of the known sufficient statistic (perturbation) is used to compute the differential changes of the free energy. However, quantification of the macroscopic change is not of general interest in the field of data driven sciences. The important task here is identifying the unknown sufficient statistic that explains the influence of the microscopic features of the macroscopic change. Unlike free energy change, KL divergence is a quantity familiar in machine learning and can be computed using parameteric or nonparametric models[23]. Indeed, equation (2.6) shows that KL divergence is a Legendre transformation[19,24] of the cumulant generating function and can be used to quantify the change of the macroscopic state. The PDF given by equation (2.2) is a special case of eq. (2.6) where we use the perturbation as a sufficient statistic for studying the change between two macroscopic states ($\lambda = 0,1$). Due to the Legendre duality[19,24] between the KL divergence and the change of the free energy, the relevance of the PDF goes beyond being a decomposition of the change of free energy. In fact, the terms of equation (2.2) are the perturbations (the features informative of the change) and their fingerprints (the configurational changes) which are quantified by the KL divergence. In this view, significant perturbations are reflected by significant changes of KL divergence.

# 3 Relative principal components analysis

This section presents a new method for analyzing the change between two states (data sets) using multivariate analysis methods[4]. Studying the change between multiple states is beyond the scope of this work and will be presented in a future publication. The newly introduced method termed "relative principal components analysis" (RPCA) computes collective canonical variables (linear combinations of the original features) termed the relative principle components (RPCAs) to which the KL divergence factorizes additively. Indeed, factorization of the KL divergence is equivalent to factorization of the logarithm of the density ratio which is a sufficient statistic of interest in machine learning[23] (see above). Factorization of KL divergence was introduced for multivariate normal distributions in the seminal work of Kullback[14]. However, the theoretical approach of factorizing KL divergence as introduced by Kullback was not accessible in practice. Specifically, a solution is needed around the singularity of the covariance matrices and the resulting features have to be optimal with respect to maximizing their KL divergences (see below).

Let $\boldsymbol{x} = (x_1 \dots x_d)^T$ be a $d$-dimensional continuous random variable with two samples from two macroscopic states that will be labeled ($a$) and ($b$). RPCA aims at finding $k$ latent canonical

variables $\boldsymbol{y} = (y_1, y_2 \dots y_k)^T = f(\boldsymbol{x})$ which satisfy the following two conditions: (i) their marginal distributions are independent in the states *a* and *b* meaning that their KL divergences are additive, and (ii) optimal in terms of maximizing the KL divergence, such that we can use $m$ ($m \ll d$) latent variables to represent the significant directions informative of the change. The KL divergence of a new variable $\boldsymbol{y} = f(\boldsymbol{x})$ is always non-negative[13,14] and is bounded from above by the KL divergence of the original variable $\boldsymbol{x}$[6,14]:

$$D_{kl}\big(P_b(\boldsymbol{x}) \parallel P_a(\boldsymbol{x})\big) \geq D_{kl}\left(P_b\big(f(\boldsymbol{x})\big) \parallel P_a\big(f(\boldsymbol{x})\big)\right) \geq 0 \tag{3.1}$$

The new variable $\boldsymbol{y} = f(\boldsymbol{x})$ is "sufficient" if equality holds in (3.1). For example, a one-to-one mapping $\boldsymbol{x} \leftrightarrow f(\boldsymbol{x})$ returns a sufficient variable $f(\boldsymbol{x})$. When studying the change between two macroscopic states, a sufficient one-dimensional variable always exists regardless of the dimensionality of $\boldsymbol{x}$. Examples of one-dimensional sufficient variables include the perturbation[6] $U_p(\boldsymbol{x})$ and the logarithm of the density ratio $\ln[P_b(\boldsymbol{x})/P_a(\boldsymbol{x})]$ :

$$D_{kl}\big(P_b(\boldsymbol{x}) \parallel P_a(\boldsymbol{x})\big) = D_{kl}\left(P_b\big(U_p(\boldsymbol{x})\big) \parallel P_a\big(U_p(\boldsymbol{x})\big)\right) \tag{3.2}$$

Although the existence of a one-dimensional sufficient feature appears promising, it is not practically useful for two reasons: (i) the analytical nature of the sufficient one-dimensional variable is generally unknown and its nonparametric estimation (e.g. the density ratio) is subject to the curse of dimensionality[25]. (ii) The complexity and nonlinearity of a sufficient one-dimensional variable –if it is known– will hinder its interpretability in terms of the original features $\boldsymbol{x}$. In fact, its simple interpretability and analytical traceability is one reason for the wide-spread use of the Gaussian linear parametric model in latent variable models[5] (e.g. PCA). We will adopt this model here as well. Then the latent variables are linear combinations of the original variable and are normally distributed:

$$\begin{gathered} \boldsymbol{y} = f(\boldsymbol{x}) = \boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{x} \\ \boldsymbol{y_l} \sim \mathcal{N}\big(\boldsymbol{\mu}_l, \boldsymbol{\Lambda}_l\big) \ ; \ l = a, b \end{gathered} \tag{3.3}$$

Here, $\boldsymbol{G}$ is a $d \times k$ transformation matrix with columns $\boldsymbol{g_i}$. Thus $y_i = \boldsymbol{g_i}^{\boldsymbol{T}}\boldsymbol{x}$ and $\boldsymbol{\mu}_l$ ; $l = a, b$ are the averages of the two distributions. The covariance matrices $\boldsymbol{S}_l$ of the original variables are related to the covariance matrices of the latent variables by $\boldsymbol{\Lambda}_l = \boldsymbol{G}^{\boldsymbol{T}}\boldsymbol{S}_l\boldsymbol{G}$; $l = a, b$. Under the model assumption of normality, it is sufficient for the normally distributed vectors $\boldsymbol{y_i}$ to be uncorrelated (independent) in both states *a* and *b* to make their contributions to the KL divergences additive. This requires the covariance matrices of the latent variables to be diagonal. Here, the diagonal matrix of state *b:* $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1 \dots \lambda_k)$ contains the variances $\lambda_1 \dots \lambda_k$ of the latent variables at state *b* while the covariance matrix of the latent variables at state *a* is arbitrarily selected to be an identity matrix ($\boldsymbol{I}$). Briefly, the task of RPCA amounts to finding a transformation matrix $\boldsymbol{G}$ which simultaneously diagonalizes both covariance matrices:

$$\boldsymbol{G}^T\boldsymbol{S}_a\boldsymbol{G} = \boldsymbol{I}$$
$$\boldsymbol{G}^T\boldsymbol{S}_b\boldsymbol{G} = \boldsymbol{\Lambda} \tag{3.4}$$

The independence (orthogonality) of the latent variables in the two states implies that their KL divergences are additive and can be analytically computed based on the model assumption of normality [14]:

$$D_{\mathrm{KL}}(\boldsymbol{y}; a{:}\, b) = \sum_{i=1}^{k} D_{\mathrm{KL}}(y_i = \boldsymbol{g}_i^T\boldsymbol{x}; a{:}\, b) = \sum_{i=1}^{k}\frac{1}{2}\left[\underbrace{-\ln\lambda_i + \lambda_i - 1}_{\text{Variance change}} + \underbrace{\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i}_{\text{Average change}}\right] \tag{3.5}$$

Here $\Delta = \boldsymbol{\mu}_b - \boldsymbol{\mu}_a$ is the change of the average of the distributions of $\boldsymbol{x}$ in states $a$ and $b$. In case $\boldsymbol{S}_a$ is a well-conditioned nonsingular (positive definite) matrix (e.g. a multivariate normal distribution), Kullback[14] formulated the problem of solving (3.4) for $\boldsymbol{G}$ as solving the generalized eigenvalue problem[14] $|\boldsymbol{S}_b - \lambda\boldsymbol{S}_a|$ where the generalized eigenvectors are the columns of $\boldsymbol{G}$. The generalized eigenvalue problem can be solved using Wilkinson's algorithm[26] involving the Cholesky decomposition of $\boldsymbol{S}_a$[27]. Practically, the generalized eigenproblem is not solvable when the covariance matrix $\boldsymbol{S}_a$ is singular. Indeed, the covariance matrices of real data are mostly singular or ill-conditioned. Singularity arises due to the fact that the real dimensionality of the probability distributions is smaller than the apparent dimensionality of $\boldsymbol{x}$[28].

## RPCA via simultaneous diagonalization of two matrices.

Now we present a general algorithm for the simultaneous diagonalization of the matrices in equation (3.4) which can be used even if $\boldsymbol{S}_a$ is singular. Optimizing the solution to having a relative principal component with maximum KL divergence will be introduced in the next section.

A $d \times k$ transformation matrix $\boldsymbol{G}$ that simultaneously diagonalizes the symmetric $d \times d$ matrices $\boldsymbol{S}_a$ (of rank $k$) and $\boldsymbol{S}_b$ (equation (3.4)) can be found by a combination of two transformation matrices:

$$\boldsymbol{G} = \boldsymbol{W}\,\boldsymbol{\Psi} \tag{3.6}$$

1) The so-called *whitening transformation*[2] matrix $\boldsymbol{W}$ of the matrix $\boldsymbol{S}_a$:

$$\boldsymbol{W} = \boldsymbol{U}_k\boldsymbol{D}_k^{-1/2} \tag{3.7}$$

can be computed from an eigendecomposition of $\boldsymbol{S}_a$:

$$\boldsymbol{S_a} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T = [\boldsymbol{U}_k\ \ \boldsymbol{U_0}]\begin{bmatrix}\boldsymbol{D}_k & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_k^T\\ \boldsymbol{U}_0^T\end{bmatrix} = \boldsymbol{U}_k\boldsymbol{D}_k\boldsymbol{U}_k^T \tag{3.8}$$

Here, the $k$ eigenvectors in the columns of $\boldsymbol{U}_k$ correspond to the $k$ nonzero eigenvalues in the diagonal matrix $\boldsymbol{D}_k$ so that the $d-k$ eigenvectors $\boldsymbol{U_0}$, which are corresponding to the eigenvalues of zeros, are discarded. $\boldsymbol{0}$ denotes a matrix or a vector of zeros of a suitable dimensionality. The $d \times k$ whitening transformation $\boldsymbol{W}$ reduces (whitens) $\boldsymbol{S}_a$ to a $k \times k$ identity matrix:

$$\boldsymbol{W}^T\boldsymbol{S}_a\boldsymbol{W} = \boldsymbol{D}_k^{-1/2}\boldsymbol{U}_k^T\boldsymbol{S}_a\boldsymbol{U}_k\boldsymbol{D}_k^{-1/2} = \boldsymbol{D}_k^{-1/2}\boldsymbol{D}_k\boldsymbol{D}_k^{-1/2} = \boldsymbol{I} \tag{3.9}$$

In case $\boldsymbol{S}_a$ is nonsingular ($d = k$), all eigenvalues are positive and the algorithm above is well known[2].

2) The second transformation matrix $\boldsymbol{\Psi}$ is formed using the eigenvectors from the eigendecomposition of the symmetric matrix $\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_b\boldsymbol{W}$:

$$\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_b\boldsymbol{W} = \boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^{\boldsymbol{T}} \tag{3.10}$$

Combining the transformations matrices from (3.7) and (3.10) yields a $d \times k$ transformation matrix $\boldsymbol{G} = \boldsymbol{W}\,\boldsymbol{\Psi}$ which simultaneously diagonalizes $\boldsymbol{S}_a$ and $\boldsymbol{S}_b$:

$$\begin{gathered}\boldsymbol{G}^T\boldsymbol{S}_a\boldsymbol{G} = \boldsymbol{\Psi}^T\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_a\boldsymbol{W}\boldsymbol{\Psi} = \boldsymbol{\Psi}^T\boldsymbol{I}\boldsymbol{\Psi} = \boldsymbol{I} \\ \boldsymbol{G}^T\boldsymbol{S}_b\boldsymbol{G} = \boldsymbol{\Psi}^T\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_b\boldsymbol{W}\right)\boldsymbol{\Psi} = \boldsymbol{\Lambda}\end{gathered} \tag{3.11}$$

The $k$ columns of $\boldsymbol{G}$ are the generalized eigenvectors with the corresponding generalized eigenvalues in the diagonal matrix $\boldsymbol{\Lambda}$. The relative principle components $y_i = \boldsymbol{g}_i^T\boldsymbol{x}$ can be reordered based on their KL divergences using equation (3.5). However, it is important to keep in mind that the value of $D_{\mathrm{KL}}$ of the components from equation (3.5) is computed based on the model assumption (normality). Fortunately, the one-dimensional distributions of the $y_i$ (projected points $y_i = \boldsymbol{g}_i^T\boldsymbol{x}$) of the two states can be used to assess the performance of the model in identifying the directions in which projections of the data have significant KL divergences. In other words, a significant KL divergence of a variable $y_i$ has to be reflected in a significant change between its distributions in states $a$ and $b$ (see the example below).

## Optimal RPCAs via average-covariance sub-spacing

The $m$-dimensional latent variable with maximum KL divergence can be found using the $m$ generalized eigenvectors corresponding to the largest KL divergence. The KL divergences of the generalized principle components in equation (3.5) can be decomposed into the terms holding the change of the variances $\frac{1}{2}[-\ln\lambda_i + \lambda_i - 1]$ and the terms holding the change of the averages $\frac{1}{2}\left(\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i\right)$. Indeed, the generalized eigenvectors and eigenvalues obtained by the algorithms described above are optimal with respect to maximizing the KL divergences of the components in cases when the averages are the same[14] ($\Delta = \boldsymbol{0}$). Unfortunately, this optimality of the components is violated when the averages differ due to their contributions to the KL divergences $\frac{1}{2}\left(\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i\right)$ which are not related to the generalized eigenvalues, but instead to the projections of the generalized eigenvectors onto the change of the averages ($\boldsymbol{g}_i^T\Delta$). Therefore, the following average-covariance sub-spacing algorithm is introduced to achieve the optimality of RPCAs with respect to maximizing their KL divergences. The idea is to find a transformation $\boldsymbol{G} = [\boldsymbol{g}_\mu \quad \boldsymbol{G}_{\boldsymbol{v}}]$ such that the KL divergence of the variable $y_\mu = \boldsymbol{g}_{\boldsymbol{\mu}}^T\boldsymbol{x}$ summarizes the total KL divergence due to changes of the averages while the KL divergences of the variables $\boldsymbol{y}_{\boldsymbol{v}} = \boldsymbol{G}_{\boldsymbol{v}}^T\boldsymbol{x}$ are purely due to the change of covariance. The vector $\boldsymbol{S}_{\boldsymbol{a}}^{-\boldsymbol{1}}\Delta$ is obtained by maximizing the total contributions to the divergence due to the change of the averages $\frac{1}{2}\left(\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i\right)$ with respect to $\boldsymbol{g}_i$. Thus, $\boldsymbol{g}_\mu$ is given by[14]:

$$\boldsymbol{g_m} = \frac{\boldsymbol{S_a^-}\Delta}{\sqrt{\boldsymbol{g_\mu^T S_a g_\mu}}} \; ; \; \boldsymbol{S_a^-} = \boldsymbol{U}_k \boldsymbol{D}_k^{-1} \boldsymbol{U}_k^T \tag{3.12}$$

Here, the generalized pseudoinverse $\boldsymbol{S_a^-}$ is used for the general case (e.g. singular $\boldsymbol{S_a}$; see the proof in the Supplementary Material S1) and $\boldsymbol{g}_\mu$ is normalized with respect to the covariance matrix $\boldsymbol{S_a}$ such that $\boldsymbol{g_\mu^T S_a g}_\mu = 1$. The KL divergence of the new variable $y_\mu = \boldsymbol{g_\mu^T x}$ includes the total KL divergence due to the change of the averages which in turn equals half of the Mahalanobis distance between the averages $\frac{\boldsymbol{1}}{\boldsymbol{2}}\Delta^T \boldsymbol{S_a^-}\Delta$:

$$\begin{gathered} D_{\mathrm{KL}}\big(y_\mu = \boldsymbol{g_\mu^T x}; a{:}\, b\big) = \frac{1}{2}\Bigg[\underbrace{-\ln\lambda_\mu + \lambda_\mu - 1}_{\substack{\text{Variance}\\ \text{change}}} + \underbrace{\Delta^T \boldsymbol{S_a^-}\Delta}_{\substack{\text{Average}\\ \text{change}}}\Bigg] \\ \boldsymbol{g_\mu^T}\Delta\Delta^T \boldsymbol{g}_\mu = \Delta^T \boldsymbol{S_a^-}\Delta \\ \mathrm{Var}\big(y_\mu | b\big) = \lambda_\mu = \boldsymbol{g_\mu^T S_b g}_\mu \end{gathered} \tag{3.13}$$

Here, the KL divergence of the new variable $y_\mu = \boldsymbol{g_\mu^T x}$ also includes a contribution due to the change of the variance and the corresponding generalized eigenvalue (variance of $y_\mu$) is given by $\lambda_\mu$.

The remaining generalized eigenvectors $\boldsymbol{G_v}$, which do not contain contributions arising from the change of the averages (they are orthogonal to the change of the averages $\boldsymbol{G_v^T}\Delta = \boldsymbol{0}$), are computed after deflating the contributions to the KL divergence due to vector $\boldsymbol{g}_\mu$ from the matrices $\boldsymbol{S_a}$ and $\boldsymbol{S_b}$. Given the vector $\boldsymbol{g}_\mu$, the conditions for $\boldsymbol{G_v}$ to fulfill the restricted simultaneous diagonalization problem can be presented in the equations:

$$\begin{gathered} \boldsymbol{G^T S_a G} = \begin{bmatrix} \boldsymbol{g_\mu^T} \\ \boldsymbol{G_v^T} \end{bmatrix} \boldsymbol{S_a} [\boldsymbol{g}_\mu \quad \boldsymbol{G_v}] = \begin{bmatrix} \boldsymbol{g_\mu^T S_a g}_\mu & \boldsymbol{g_\mu^T S_a G_v} \\ \boldsymbol{G_v^T S_a g}_\mu & \boldsymbol{G_v^T S_a G_v} \end{bmatrix} = \boldsymbol{I} \\ \boldsymbol{G^T S_b G} = \begin{bmatrix} \boldsymbol{g_\mu^T} \\ \boldsymbol{G_v^T} \end{bmatrix} \boldsymbol{S_b} [\boldsymbol{g}_\mu \quad \boldsymbol{G_v}] = \begin{bmatrix} \boldsymbol{g_\mu^T S_b g}_\mu & \boldsymbol{g_\mu^T S_b G_v} \\ \boldsymbol{G_v^T S_b g}_\mu & \boldsymbol{G_v^T S_b G_v} \end{bmatrix} = \begin{bmatrix} \lambda_\mu & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda_v} \end{bmatrix} \end{gathered} \tag{3.14}$$

$$\text{subject to} \quad \boldsymbol{G_v^T}\Delta = \boldsymbol{0} \; (\boldsymbol{g}_i^T\Delta = 0 \text{ ; no KL divergence due the average change}) \tag{3.15}$$

Here, $\boldsymbol{\Lambda_v}$ is a diagonal matrix. To fulfill the conditions in (3.14) and (3.15), the generalized eigenvectors $\boldsymbol{G_v}$ can be constructed using a combination of two transformation matrices (see the details in the Supplementary Material S2):

$$\boldsymbol{G_v} = \boldsymbol{W_v \Psi_v} \tag{3.16}$$

(i) The whitening transformation $\boldsymbol{W_v}$, similarly to equation (3.7), is constructed here from the eigendecomposition of the projection of $\boldsymbol{S_a}$ on the subspace, which is orthogonal to the vector $(\boldsymbol{S_a g}_\mu)$.
(ii) The second transformation $\boldsymbol{\Psi_v}$ is obtained from the eigenvectors of the projection of $\boldsymbol{W_v^T S_b W_v}$ onto

the subspace which is orthogonal to the vector $\boldsymbol{W_v^T S_b g}_\mu$. Here, the number of generalized eigenvectors from the optimal RPCA sub-spacing is less by one than the number of generalized eigenvectors from the non-optimal RPCA algorithm.

### Data reconstruction from the latent variable

The influence of the $m$-dimensional ($m < d$) latent variable $\boldsymbol{y}_m = \boldsymbol{G}_m^{\boldsymbol{T}}\boldsymbol{x}$ on the original $d$-dimensional variable $\boldsymbol{x} \in \mathbb{R}^d$ can be presented via the reconstruction $\hat{\boldsymbol{x}} \in \mathbb{R}^d$ in the subspace which is spanned by the corresponding $m$ generalized eigenvectors $\boldsymbol{G}_m = [\boldsymbol{g}_1 \dots \boldsymbol{g}_m]$. The projection $\hat{\boldsymbol{x}}$ (reconstruction) of a data point $\boldsymbol{x}$ on the subspace, which is spanned by columns of $\boldsymbol{G_m}$, is given by the relationship[29]:

$$\hat{\boldsymbol{x}} = \underbrace{\boldsymbol{G}_m\left(\boldsymbol{G}_m^{\boldsymbol{T}}\boldsymbol{G}_m\right)^{-}\boldsymbol{G}_m^{\boldsymbol{T}}}_{\boldsymbol{P_m}}\boldsymbol{x} = \boldsymbol{P}_m\boldsymbol{x} \tag{3.17}$$

$$\hat{\boldsymbol{x}} = \underbrace{\boldsymbol{G}_m\left(\boldsymbol{G}_m^{\boldsymbol{T}}\boldsymbol{G}_m\right)^{-}}_{\boldsymbol{R}}\underbrace{\boldsymbol{G}_m^{\boldsymbol{T}}\boldsymbol{x}}_{\boldsymbol{y}_m} = \boldsymbol{R}\boldsymbol{y}_m \tag{3.18}$$

Here, $\boldsymbol{P}_m \in \mathbb{R}^{d\times d}$ is the projection matrix[29] on the $d$-dimensional subspace which is spanned by columns of $\boldsymbol{G}_m$. $\boldsymbol{R} \in \mathbb{R}^{d\times m}$ is the reconstruction matrix facilitating the transformation from the latent variable. It is straightforward (see the Supplementary Material S3) to show that the KL divergence between the distributions of the reconstructed variable $\hat{\boldsymbol{x}}$ equals the KL divergence between the distributions of its corresponding latent variable $\boldsymbol{y}_m$. In other words, the most informative approximation (the one maximizing KL divergence) of the original variable $\boldsymbol{x}$ using a restricted number ($m < d$) of variables is obtained using the $m$-dimensional latent variables $\boldsymbol{y}_m$, corresponding to the $m$ generalized eigenvectors, with the highest KL divergences.

### Change hotspots from RPCA

Besides being very useful for dimensionality reduction, the collective canonical nature of the RPCA variables $y_i = \boldsymbol{g}_i^T\boldsymbol{x}$ reflects the fact that the change is a collective process includes all the elements (features) of the original variable $\boldsymbol{x}$. Nevertheless, it is also interesting to have the possibility to map the featurewise (local) contributions to the change from the elements $x_i$ of $\boldsymbol{x}$ so that the elements of $\boldsymbol{x}$ with larger contribution to the change (divergence) can be interpreted as the hotspots of the change. The contributions to the divergence in equation (3.5) can be approximated as a sum of two quadratic terms:

$$D_{\mathrm{KL}}(y_l; a{:}\,b) \approx \frac{1}{2}\left[\lambda_l - 1 + \boldsymbol{g}_l^T\Delta\Delta^T\boldsymbol{g}_l\right]$$

$$\lambda_l = \boldsymbol{g}_l^T\boldsymbol{S}_b\boldsymbol{g}_l = \underbrace{\sum_{\boldsymbol{i}} g_{li}(\boldsymbol{S_b})_{ii}g_{li}}_{\text{local}} + \underbrace{\sum_{i\neq j} g_{li}(\boldsymbol{S_b})_{ij}g_{lj}}_{\text{cooperative}}$$

$$\boldsymbol{g}_l^T\Delta\Delta^T\boldsymbol{g}_l = \underbrace{\sum_{\boldsymbol{i}} g_{li}(\Delta\Delta^T)_{ii}g_{li}}_{\text{local}} + \underbrace{\sum_{i\neq j} g_{li}(\Delta\Delta^T)_{ij}g_{lj}}_{\text{cooperative}} \tag{3.19}$$

Here, $(\boldsymbol{A})_{\boldsymbol{ij}}$ denotes element (*i,j*) of matrix $\boldsymbol{A}$. The contributions to the quadratic terms are due to the local contributions from the elements and their cooperative (cross) terms taking into account that each element of a generalized eigenvector ($g_{li}$) corresponds to an element of $x_i$. However, it is clear that the contributions to the quadratic terms in equation (3.19) can be collected via arbitrary grouping of the elements into subgroups. The computation of the (local) group-wise contributions and their cooperative contributions is equivalent to the computation of the quadratic terms using the corresponding submatrices of $\boldsymbol{S_b}$ and $\Delta\Delta^T$.

## Asymmetric nature of RPCA

Multivariate analysis methods can be grouped into methods handling either one state or multiple states. The methods which study one state include methods handling one multivariate variable, such as principal component analysis, factor analysis and independent components analysis, and methods handling two multivariate variables with concurrent measurement (a joint distribution) such as canonical correlation, regression and partial least squares. Discriminant analysis (classification) methods, on the other hand, aim at finding the variation between several states (classes). Although RPCA is similar to feature extraction for classification[2], there are fundamental differences between pattern classification and the physical definition of the change, which is introduced above. For example, the discriminant features in Fisher discriminant analysis (FDA) are related to the change of the averages of distributions and the information from the covariance matrices is whitened using a unified pooled (average) covariance matrix[2]. Therefore, the usage of FDA for dimensionality reduction[30] is known to be restricted by a lower limit on the number of dimensions which is equal to the number of classes minus one. While the objective functions in discriminant analysis are required to be symmetric[2], the objective function (KL divergence) for RPCA is asymmetric, which reflects the directed character of changes[31]. Let us consider the situation of the backward change from state *b* to state *a*. Combining the transformation matrix $\boldsymbol{G}$ in equations (3.4) with another nonsingular diagonal transformation matrix $\boldsymbol{\Lambda}^{-1/2}$, containing the inverse of the square roots of the generalized eigenvalues, yields a new transformation matrix $\boldsymbol{G_R} = \boldsymbol{G\Lambda}^{-1/2}$ which simultaneously diagonalizes $\boldsymbol{S}_b$ and $\boldsymbol{S}_a$:

$$\begin{aligned} \boldsymbol{G_R^T S}_b \boldsymbol{G}_R &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{G}^T\boldsymbol{S}_b\boldsymbol{G\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda\Lambda}^{-1/2} = \boldsymbol{I} \\ \boldsymbol{G_R^T S}_a \boldsymbol{G}_R &= \boldsymbol{\Lambda}^{-1/2}\boldsymbol{G}^T\boldsymbol{S}_a\boldsymbol{G\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{I\Lambda}^{-1/2} = \boldsymbol{\Lambda}^{-1} \end{aligned} \tag{3.20}$$

Clearly, the directions of the generalized eigenvectors are not changed and the eigenvectors $\boldsymbol{G_R}$ pertaining to the reverse direction are the scaled eigenvectors pertaining to the forward direction ($\boldsymbol{G}$):

$$\boldsymbol{G_R} = \boldsymbol{G\Lambda}^{-1/2} = \left[\frac{\boldsymbol{1}}{\sqrt{\boldsymbol{\lambda_1}}}\boldsymbol{g_1} \cdots \frac{\boldsymbol{1}}{\sqrt{\boldsymbol{\lambda_k}}}\boldsymbol{g_k}\right] \tag{3.21}$$

When studying the backward change from state *b* to state *a,* the generalized eigenvalues, which present the change of the variance, are inverted ($1/\lambda_i$) and, interestingly, can be interpreted in the following way: taking into account that the KL divergence due to the change of the variance $1/2\,(-\ln\lambda_i + \lambda_i - 1)$ in equation (3.5) is a convex function with a minimum of 0 at $\lambda_i = 1$, we can divide the relative principal

components into two groups. The first group includes the components with $\lambda_i > 1$ where the variance (quantified by $\lambda_i$) along the direction $\boldsymbol{g}_i$ increases as we transit from state *a* to *b*. The second group includes the components with $\lambda_i < 1$ since the respective variance decreases as we transit from state *a* to *b*. When the backward change is taking place from *b* to *a,* the roles of the components of these groups (quantified by $1/\lambda_i$) are exchanged. The same role of these groups is also observed when taking the KL divergence due to the change of the average where the directions $\boldsymbol{g}_i$ with $\lambda_i > 1$ have diminished contributions ($\frac{1}{2\lambda_i}\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i$) for the backward direction from *b* to *a* in comparison with their contribution ($\frac{1}{2}\boldsymbol{g}_i^T\Delta\Delta^T\boldsymbol{g}_i$) to the transition from state *a* to *b* and vice versa.

## RPCA and the distance metric

Even though the task of RPCA (analyzing the change between states) differs from the task of PCA (finding the variation within one state; see Figure 1), a similarity exists regarding the applied distance metric. However, it is first important to notice that the generalized eigenvectors are orthonormal with respect to the matrix[29] $\boldsymbol{S}_a$ ($\boldsymbol{g}_i^T\boldsymbol{S}_a\boldsymbol{g}_j = \delta_{ij}$) , orthogonal with respect to $\boldsymbol{S}_b$ ($\boldsymbol{g}_i^T\boldsymbol{S}_b\boldsymbol{g}_i = \lambda_i$ ; $\boldsymbol{g}_i^T\boldsymbol{S}_b\boldsymbol{g}_{j\neq i} = 0$) and not necessarily orthonormal to each other ($\boldsymbol{g}_i^T\boldsymbol{g}_j \neq \delta_{ij}$). The orthogonality between two vectors $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ with respect to a matrix $\boldsymbol{S}_a$ is defined in terms of the quadratic product $\boldsymbol{g}_i^T\boldsymbol{S}_a\boldsymbol{g}_j$ instead of the traditional dot product. RPCA analysis can be interpreted as using the distance metric[32] of state *a* to analyze state *b*. Indeed, applying the whitening transformation[2] in equation (3.7) removes the "information" within state *a* ($\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_a\boldsymbol{W} = \boldsymbol{I}$). The eigendecomposition of $\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_b\boldsymbol{W}$ in (3.10) is performed in the whitened space in which the squared Euclidean distance between two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ equals their Mahalanobis distance in the original space and the projections on the generalized eigenvectors factorizes the Mahalanobis distance (see Supplementary Material S4):

$$\begin{aligned}(\boldsymbol{x}_1 - \boldsymbol{x}_2)^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{a}}^{-}(\boldsymbol{x}_1 - \boldsymbol{x}_2) &= \left(\boldsymbol{G}^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right)^{\boldsymbol{T}}\left(\boldsymbol{G}^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right) = \sum_{i=1}^{k}\left[\boldsymbol{g}_{\boldsymbol{i}}^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right]^{\boldsymbol{2}} \\ &= \left(\boldsymbol{W}^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right)^{\boldsymbol{T}}\left(\boldsymbol{W}^{\boldsymbol{T}}(\boldsymbol{x}_1 - \boldsymbol{x}_2)\right)\end{aligned} \tag{3.22}$$

The Mahalanobis distance is defined via the matrix $\boldsymbol{S}_a$ as a distance metric instead of the Euclidean distance, which we obtain by setting $\boldsymbol{S}_a = \boldsymbol{I}$. Put differently, we can show that the average Mahalanobis distance of points in state *b* to the average of state *a* can be written as the sum of the generalized eigenvalues and the Mahalanobis distance between the averages of the states (see supplementary material S4):

$$\left\langle(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{a}})^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{a}}^{-}(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{a}})\right\rangle_b = \underbrace{\mathrm{trace}\left(\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{b}}\boldsymbol{W}\right)}_{\sum\lambda_i} + \underbrace{\Delta^{\boldsymbol{T}}\boldsymbol{S}_{\boldsymbol{a}}^{-}\Delta}_{\substack{\text{Mahalanobis}\\ \text{distance}}} \tag{3.23}$$

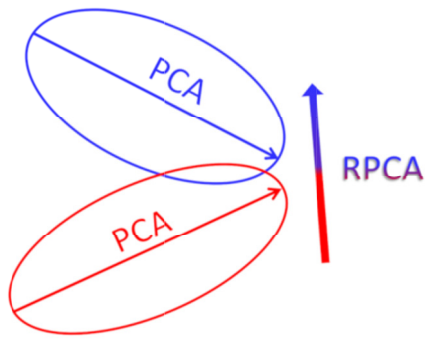


**Figure 1.** Schematic representation of the difference between PCA which finds the largest variation within each state and RPCA which finds the change from the initial to the final state.

# 4 Application of RPCA to reveal the thermodynamically important molecular conformational changes

Analyzing the conformational changes taking place during biomolecular reactions is one of the most important tasks in structural biology. Unfortunately, analyzing and mechanistically understanding biochemical interactions is quite tricky due to the complex conformational dynamics in the high-dimensional space where the interactions take place. For example, the association between two biomolecules entails (microscopic) conformational changes in a large number of internal (bonds, angles …etc.) and external (rotational and translational) degrees of freedom. The macroscopic changes in biochemical systems, on the other hand, are quantified using scalar (one-dimensional) quantities such as the change of free energy ($\Delta G^{\circ}$). Molecular simulations are becoming a more and more attractive tool for analyzing conformational changes of biomolecules. Current interest in analyzing molecular conformational changes (e.g. using Markov state models[33]) focusses on characterizing the kinetic changes between representative conformations within one macroscopic state. The probabilistic framework for analyzing the transitions between microscopic states is performed in a data space where the points are the (clustered) conformations. Methods such as PCA[34,35], normal mode analysis[36] and partial least squares[37] were used to analyze the conformational changes in the feature space (coordinates). However, most of these methods are only suited for studying the dynamic changes within one macroscopic state and not for studying transitions and changes between two states. On the other hand, methods adapted from multivariate analysis often lack thermodynamic insight (see the example in Figure 2).

Although our RPCA method is not limited to biomolecular data, the starting point for its development was based on the new insight stemming from our recently introduced perturbation-divergence formalism (PDF)[6,7]. PDF reveals that the KL divergence ($D_{kl}$) equals the work which is spent to change the conformational ensemble and hereby is the thermodynamically relevant objective function for analyzing conformational changes[7]. Indeed, PDF provides a flexible framework for studying conformational changes involving either all atoms of the structure, a part of them, or aspects of the structure based on phenotypical features (e.g. open/closed, pocket volume, domain movement, etc.); see the discussion in[6]. An important benefit of using the KL divergence for quantifying the energetics related to conformational

changes is avoiding the misleading entropic terms which are subject to enthalpy-entropy compensation when using the changes in conformational entropy to estimate the importance of conformational changes, as was previously shown[38,39,6].

In this work, we generalize PDF to all sorts of dynamic systems using the exponential family framework. This is beneficial for two general tasks that we face in the field of molecular simulations: (i) Molecular conformational changes take place in a high-dimensional space, which complicates their analysis. Here, RPCA recognizes the conformational changes, which are relevant to the macroscopic thermodynamic change. (ii) In "alchemical" free energy calculations[40–43], the microscopic factors underlying the change between two states are already known and applied by changing the parameters of the potential functions. $\Delta F$ is computed via sampling[15,16] the sufficient statistic(the perturbation) which is computed from the change of the potential energy $U_p(\boldsymbol{x}) = U_b(\boldsymbol{x}) - U_a(\boldsymbol{x})$. The estimation of $\Delta F$ using the sufficient one-dimensional variable (perturbations) has the advantage of avoiding the curse of dimensionality. Indeed, the exponential family framework of statistical estimation theory provides an improved understanding of the error bound in the estimation of free energy[18]. Moreover, the influence (the gradient and the Hessian) of the parameters $\boldsymbol{\lambda}$ in equation (2.7) is given by the mean and the covariance of their corresponding sufficient statistics.

In the following, we present use cases for applying RPCA to analyze the molecular conformational changes of the protein HIV-1 protease upon binding to several inhibitor molecules. The conformations of the protein are sampled via molecular dynamics simulations for both the initial (free, unbound) state and the final (bound) state. Technical details on these simulations are provided in the Supplementary Material S6. Figure 2 shows an assessment of the relationship between the relevant components of the dynamic change within one state, which are analyzed using traditional PCA of the data points of the final state, and the thermodynamic importance of the corresponding conformational changes. The thermodynamic importance of the conformational changes along a principal component to the association process is quantified by the KL divergence of the distributions of projections of both the free and the bound state conformations on the component[38]. Figure 2a shows that the eigenvalues of the principal components are not related to their thermodynamic importance for the association process. To illustrate this more clearly, Figure 2b shows that the principal component 3654 has more thermodynamic importance (KL divergence) than the first principal component, which is mostly irrelevant for the association process.

Figure 3 shows the RPCA of the conformational changes of wild-type HIV-1 protease upon binding to a drug molecule that has high affinity (Tipranavir). Presented are both the non-optimal RPCA (Figure 3a) and the optimal RPCA calculated with the sub-spacing method (Figure 3b). The scores (KL divergence) of the importance of the components show that the first few components account for most of the KL divergence. The data points (conformations) of both the free, unbound state and the bound state are projected on selected components to illustrate the correspondence between the score (model based) and the real divergence which can be extracted from the difference between the distributions of the projections of the states. There are clear benefits of the optimal sub-spacing RPCA (Figure 3b). Namely, the first component has a significant divergence because it collects the total contribution to the KL

divergence arising from the change of the average. In contrast, one obtains identical averages when projecting the conformations sampled in the two states onto the remaining components. Thus, the remaining components do not contribute to the KL divergence that is due to the change of the averages. Figure 4 shows a comparison of the KL divergence (discriminant information) of the top-scoring component versus the lowest-scoring and the principal component with largest eigenvalue from PCA of the bound state.

RPCA facilitates studying the conformational changes from both the collective and the local point of view. The relevant (with respect to their KL divergence) collective conformational changes can be presented via reconstructing the conformations in the subspace, which is spanned by the relevant generalized eigenvectors; see equation (3.17). Collective conformational changes can also be represented via reconstructing the conformations from the (normally distributed) latent variable $\boldsymbol{y}_m$using equation (3.18). Interestingly, RPCA provides a clear distinction between the directions (generalized eigenvectors) in which the fluctuation of the conformations increases upon the change (corresponding to $\lambda_i > 1$) and the directions in which the fluctuation of the conformations decreases (corresponding to $\lambda_i < 1$). Figure 5a shows the collective conformational changes around the average conformation of the bound state along the 33 eigenvectors with the highest generalized eigenvalues ($\lambda_i > 10$) which represent the enhanced motions of the wild-type HIV-1 protease upon binding Tipranavir. Figure 5b, on the other hand, shows the collective conformational changes around the average conformation of the free state along the 33 eigenvectors with the smallest generalized eigenvalues ($\lambda_i < 0.009$) which, in turn, represent the most strongly restricted motions of the wild-type HIV-1 protease upon binding Tipranavir. However, it is important to notice that the importance of the local conformational changes (e.g. at residues) cannot be inferred from the collective conformational changes. In other words, a large fluctuation in a site, when presenting the collective conformational changes, does not imply higher thermodynamic importance than the associated less strongly fluctuating sites. Alternatively, hotspot analysis from RPCA in equation (3.19) can be used to rank the thermodynamic contribution to the association process of the local conformational changes (conformational hot spots) of the biological building blocks (residues) and the importance of the cooperative (correlation) interactions between them. Figure 6a shows a matrix plot representation (interaction map) of the contributions of the residues to the divergence. In Figure 5b, the relative local residue contributions to the divergence are mapped on the structure of the protein and the importance of the conformational changes is indicated by radii of varying size in the cartoon representation, in addition to using the color code. Interestingly, most of the marked hotspots are residues known to affect the binding affinity upon mutating them. Examples of the defined conformational hot spots are the residues in the active pocket (e.g. D25, V82) and the residues of the flap region (e.g. I50, I54). Figure 7a shows an analysis of the conformational hotspots from RPCA of the binding of a ligand (Saquinavir) to the HIV-1 protease mutant with a resistance-related mutation (I50V) which is located on the flap region outside the binding pocket. The conformational hotspots of the mutant are located at the flap region around the mutation V50. The same flap region does not show important conformational changes when applying the same analysis to the association of Saquinavir to the wild-type (I50) in Figure 7b.

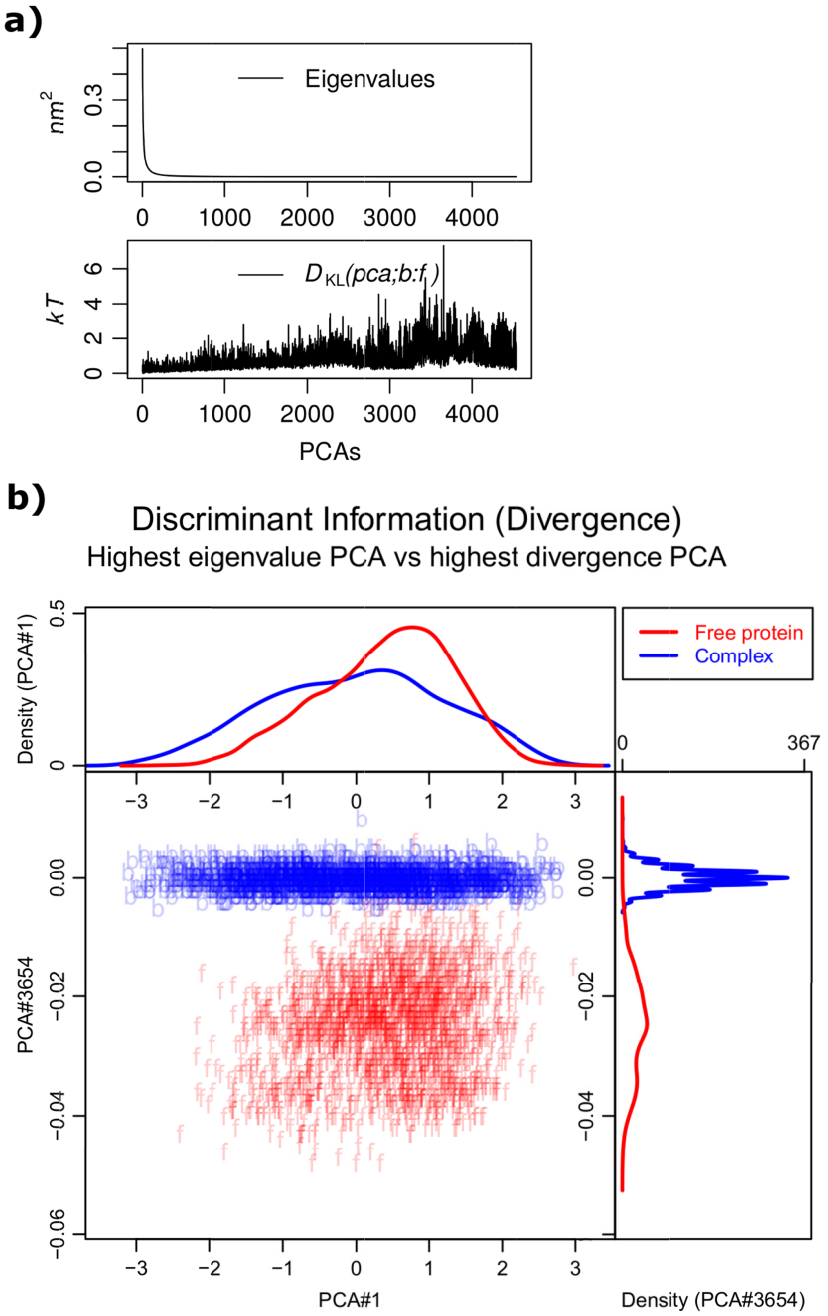


**Figure 2. Assessing the thermodynamic contribution of the conformational changes along the directions of the principal components from PCA to the association process. a)** Top panel: Shown are PCA eigenvalues derived from the covariance matrix of the conformational fluctuation (heavy atoms) in an MD simulation of HIV1 protease bound to its inhibitor Tipranavir. Bottom panel: Shown is the importance (KL divergence) along PCAs computed from the projections of the data points (conformations) of the final state (bound) and the initial state (free). **b)** The divergence of the largest eigenvalue PCA#1 ($D_{kl} \approx 0.1$) is compared to that of the PCA with highest divergence #3654 ($D_{kl} \approx 7.3$). 2000 data points are projected on the PCA vectors of the final (bound; blue) and the initial (free; red) states. The marginal probability distribution densities are shown on the side panels. The PCA analysis is performed on the conformations from the bound state using the heavy atoms of the protein. The superposition of the conformations of both the bound and the free state ensembles and the superposition of the two ensembles to each other are performed in a way similar to RPCA (see below). The KL divergences between the distributions of the projections of the two states are computed using the KLIEP method[44].

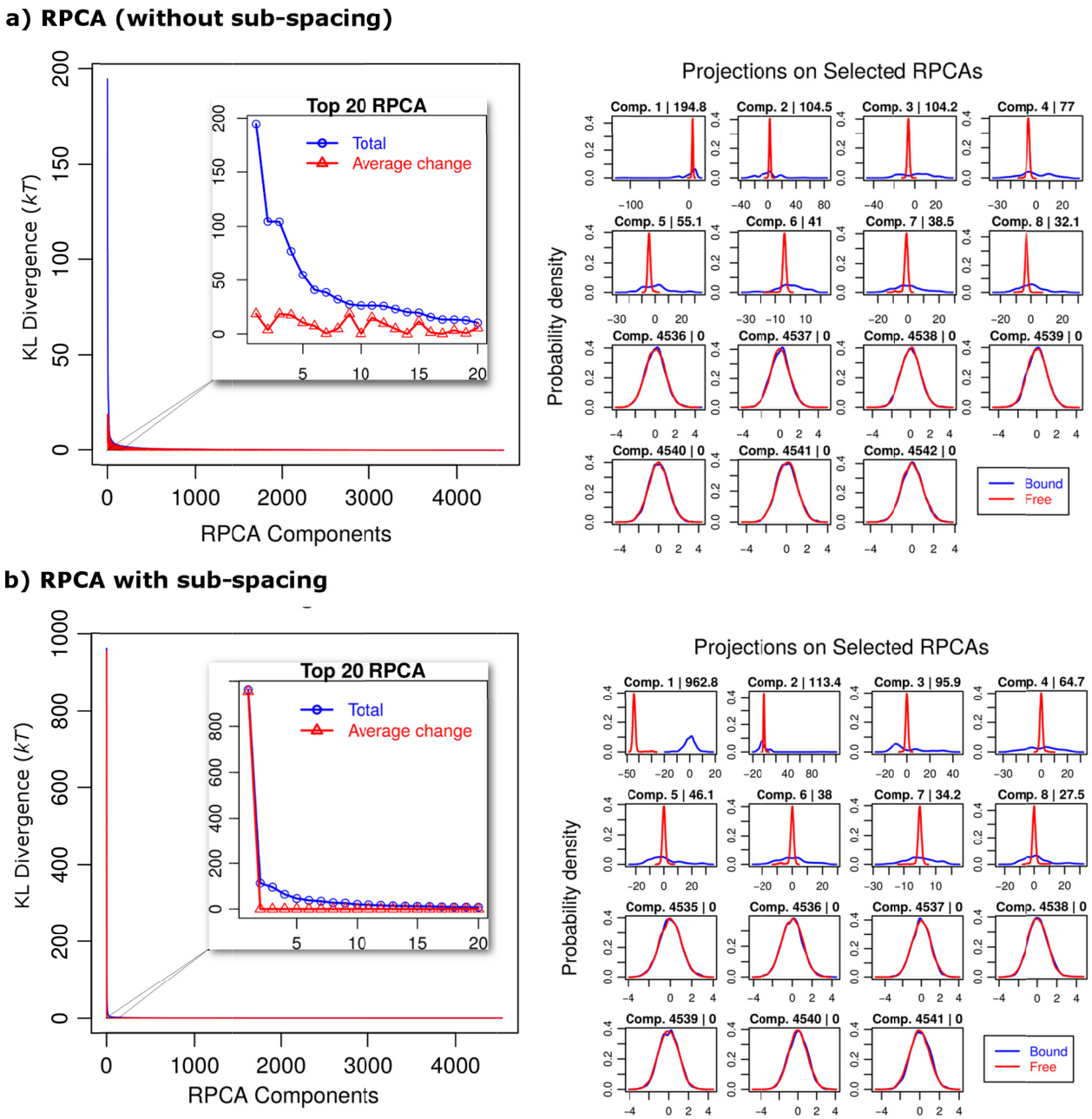


**Figure 3. RPCA of the conformational changes of HIV-1 protease upon its association with Tipranavir**. Shown are the non-optimal RPCA (a) and the optimal RPCA with sub-spacing (b). The right panels show the KL divergences of the components (blue colored) and their corresponding contributions due to the change of the average (red colored). Left panels show projections of the data points (conformations) of both the initial (free, unbound) and the final bound state on selected components. The scores (KL divergences) of the components are displayed after their corresponding number. The projections show that the components (components 1-8) with the highest rank (KL divergence) distinguish the change between the free and bound states while the components with the lowest rank do not distinguish the change (similar projections). The analysis is performed using the heavy atoms of the protein. The plots are generated using R[45] and the densities are smoothed using the kernel density estimation.

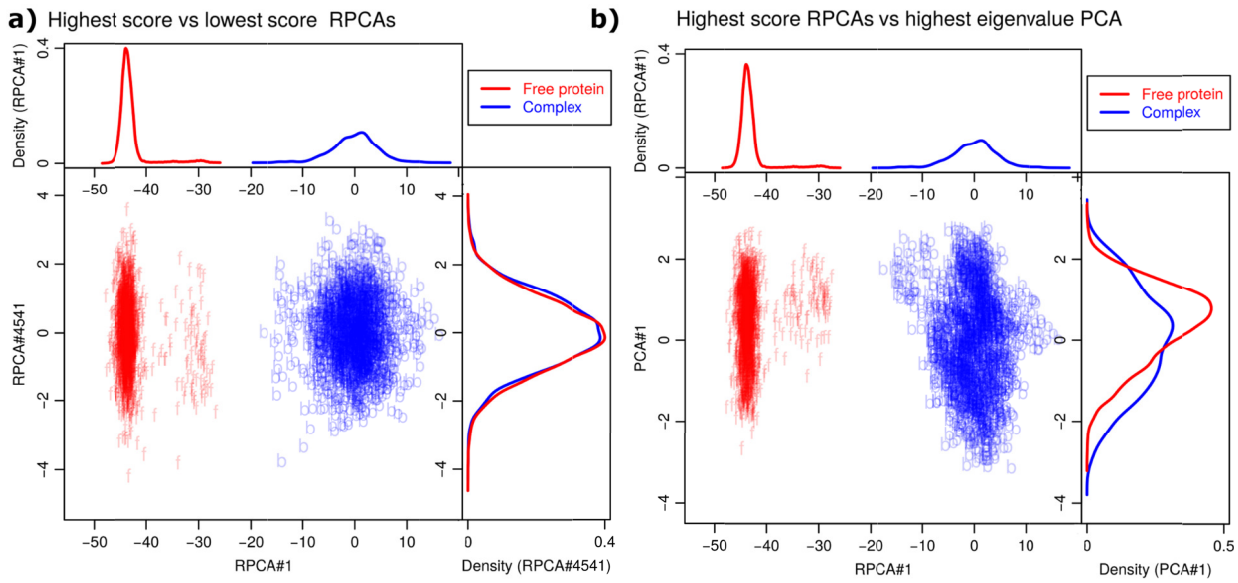


**Figure 4.** KL divergences of the highest scored relative principal component. The initial state (red) and the bound state (blue) are mapped on the top-scoring RPCA#1 against the lowest-scoring RPCA#4541 (a) and the top-scoring RPCA#1 against the top-scoring PCA#1 (highest eigenvalue) (b). RPCA is successful in extracting the components holding the most significant information on the change (binding process from free to bound state). The marginal probability distribution densities are shown on the side panels. Samples of 2000 data points where used for the projections.

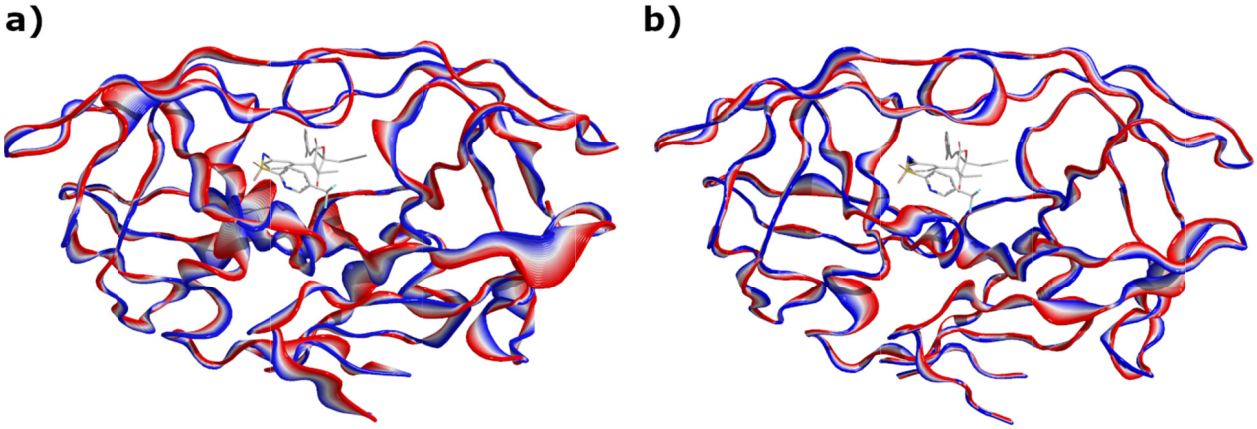


**Figure 5. Representation of the enhanced and restricted conformational fluctuations of HIV-1 protease upon binding the inhibitor Tipranavir.** Conformations around the average conformation are reconstructed from the latent variable after interpolation around its average along selected generalized eigenvectors; see equation (3.18). **a)** Enhanced conformational fluctuations around the average structure of the bound state along the 33 eigenvectors with the highest generalized eigenvalues ($\lambda_i > 10$). These conformational fluctuations increase the affinity of binding by optimizing the local conformations in the ligand-receptor interface. **b)** Conformational fluctuations around the average structure of the free state along the 33 eigenvectors with the smallest generalized eigenvalues. These latter fluctuations are highly restricted upon the association ($\lambda_i < 0.009$) because they decrease the affinity of binding via adverse local movements in the binding pocket and opening of the flap regions; see the supplementary movies. The cartoon representation is generated using Pymol[46]. The conformation of the ligand is taken from the experimental structure. Movies of these conformational changes are presented in the Supplementary Material S7.

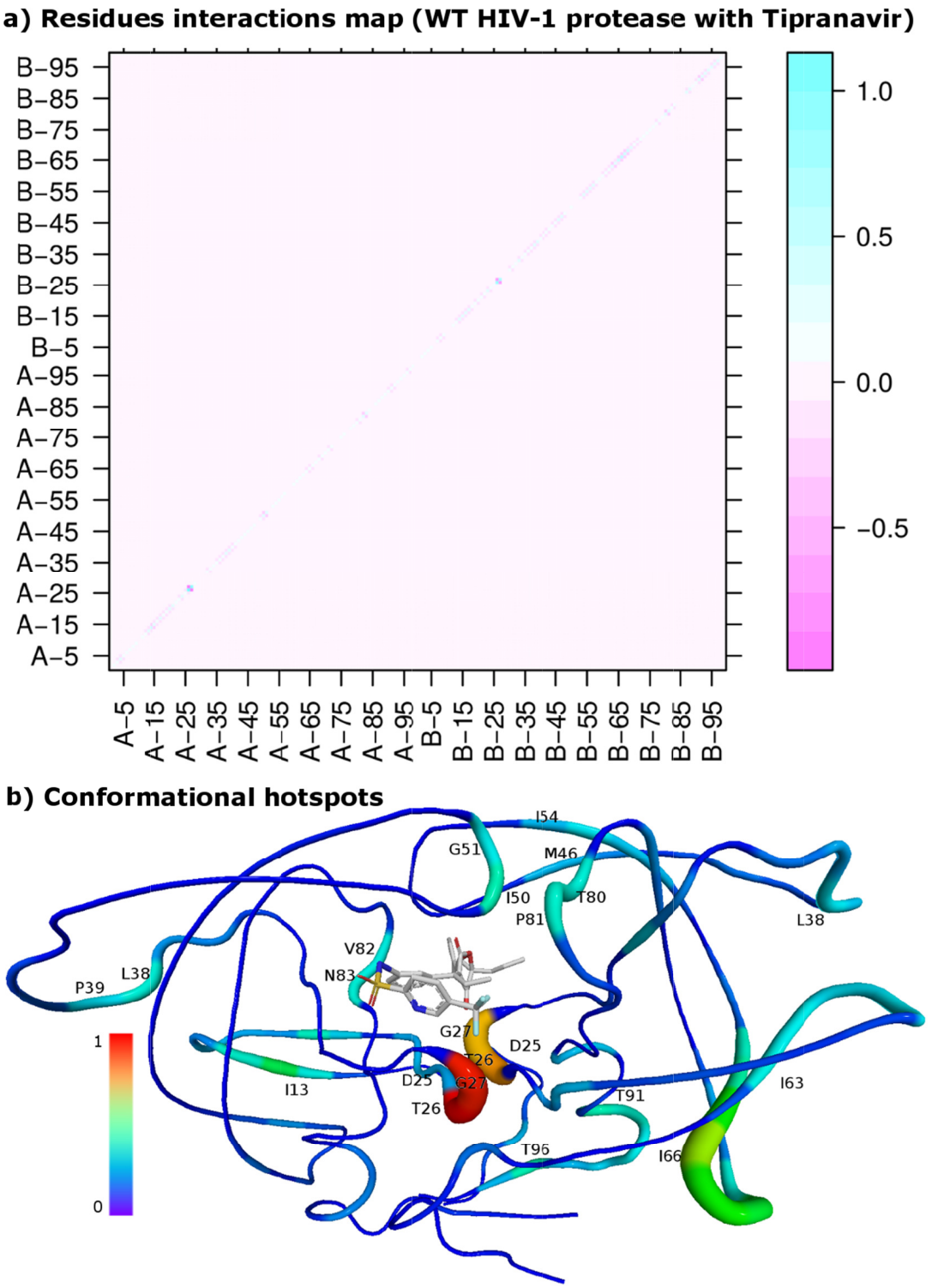


**Figure 6.** **a) Conformational interaction map** between the residues of HIV-1 protease upon binding its inhibitor Tipranavir. The contributions are concentrated around the diagonal elements indicating the important local conformational changes and the propagation of the conformational changes through the neighbor residues. **b) Conformational hotspots** mapped on the structure of the protein. The radius and the color of the cartoon indicate the importance of the conformational changes. The importance is normalized relative to the highest value that is set to 1.

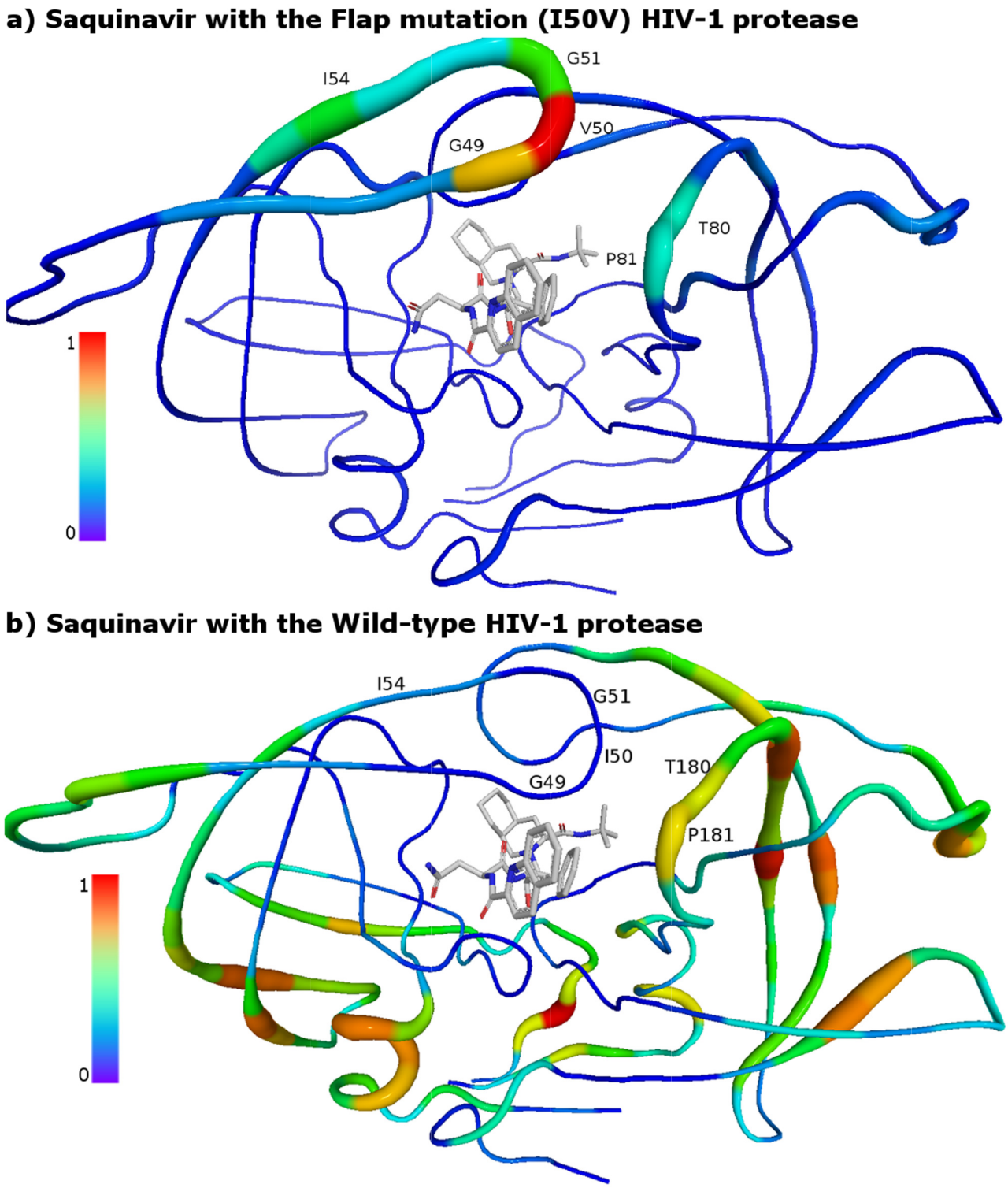


**Figure 7.** Conformational hotspots from RPCA analysis recognize the importance of the resistance-related mutation (I50V) of HIV1 protease when bound to Saquinavir. The conformational hotspots of the mutant (a) are located at the flap region around the mutation V50. The corresponding flap residues of the wild-type (I50) in figure (b) do not show energetically important (expensive) conformational changes. The radius and the color of the cartoon indicate the importance of the conformational changes. The importance is normalized relative to the highest value.

## Technical aspects of using RPCA of molecular conformational changes

**Optimal superposition of conformations of the ensembles.** The first step in analyzing an ensemble of conformations is removing the six external rotational and translational degrees of freedom via superimposing the conformations. Although the internal coordinates are not altered due to these similarity transformations (rotations and translations), the average conformation and the covariance matrix are highly affected by these external degrees of freedom when using the Cartesian coordinates to present the conformations. Traditionally, the conformations are superimposed on an arbitrary reference structure (e.g. the starting structure of the MD simulation). However, the resulting ensemble and the average structure are dependent on the reference structure. The solution of this problem is well known in statistical shape analysis as the Generalized Procrustes Analysis[32,47] (GPA). GPA returns a compact ensemble of the conformations via minimizing the sum of the Euclidean distances between the conformations (which is equal to their Euclidean distances to the average conformation). GPA is performed via an iterative algorithm having two steps: (1) compute the average structure, (2) superimpose the conformations on the new average structure. Repeating these steps ensures the convergence of the average structure. Fortunately, few iterations are usually enough for convergence.

**Superposition of two ensembles.** The Mahalanobis distance term ($\Delta^T \boldsymbol{S}_{\boldsymbol{a}}^{-} \Delta$) of the KL divergence in equation (3.13) is affected by the fashion in which we superimpose the average conformations ($\boldsymbol{\mu}_b, \boldsymbol{\mu}_a$). However, superimposition via minimizing the Euclidean distance may overestimate the KL divergence via an artificial contribution to the Mahalanobis distance $\Delta^T \boldsymbol{S}_{\boldsymbol{a}}^{-} \Delta$; $\Delta = \boldsymbol{\mu}_b - \boldsymbol{\mu}_a$. Therefore, superimposition of the average conformations should aim at minimization of the Mahalanobis distance. In contrast to the minimization of the Euclidean distance, there is no analytical solution known for minimizing the Mahalanobis distance. This type of nonlinear minimization is known as the Covariance Weighted Procrustes Analysis[48] (CWPA) and numerical methods can be used to find the optimal superposition (rotations and translations) for minimizing the quadratic term $\Delta^T \boldsymbol{S}_{\boldsymbol{a}}^{-} \Delta$ .

**Steps for performing a RPCA analysis of two molecular states:**

(1) GPA fitting of the conformations sampled in the simulation of the first state. The covariance matrix of this state is also computed at this step. A successful superimposition of the ensemble will lead to a singular covariance matrix with at least six eigenvalues of zero value accounting for removing the external degrees of freedom.

(2) GPA fitting of the conformations sampled in the simulation of the second state to obtain the average conformation.

(3) Covariance weighted fitting of the average conformation of the second state on the average conformation of the first state via minimizing their Mahalanobis distance. This unconstrained nonlinear optimization is numerically performed using the line-search algorithm and the BFGS factored method to update the Hessian[49].

(4) The new average conformation of the second state is used as a reference to refit the conformations of the second state and to compute the covariance matrix of the second state.

(5) Simultaneous diagonalization of the covariance matrices is performed. Optionally, the sub-spacing optimal algorithm can be used.
(6) KL divergences of the relative principal components are computed and the components are reordered based on their scores (KL divergences).

We have developed efficient computational tools to perform these steps. The tools are written in the C programming language and the numerical linear algebra operations are performed using the BLAS and LAPACK routines[50]. The limit of the zero value of the eigenvalues is defined using the machine precision multiplied by the largest eigenvalue. The covariance-weighted superimposition is performed using the nonlinear minimization algorithm by Dennis and Schnabel[49].

# 5 Conclusion

Here, we introduced the RPCA method, which extracts the relevant principal components describing the change between two macroscopic states of a dynamic system represented by two data sets. The definition of the macroscopic change of a dynamic system and its quantification are based on previous work where we derived a generalized quantification of the change of a physical system based on statistical mechanics. We presented use cases for conformational changes taking place upon ligand binding to HIV-1 protease that clearly illustrate power of RPCA to characterize the relevant changes between two ensembles in a high-dimensional space. Moreover, software solutions were introduced to ensure the removal of the similarity transformations (rotations and translations) via superposing the conformations using GPA and CWPA. These procedures may also be beneficial for preparing the conformations for other analysis methods.

Although RPCA is currently limited to handling only continuous variables and two macroscopic states, the introduced framework for quantifying changes of dynamic systems using the exponential family of distributions is flexible regarding the nature of probability distributions and the nature of the microscopic variables (continuous and categorical). Therefore, the presented theoretical formalism opens the door for developing improved new methods for mining the factors underlying changes in dynamic systems in the directions of (i) handling both continuous and categorical data (e.g. the effect of sequence changes (mutations) on the binding affinity) and of (ii) handling multiple macroscopic states (e.g. study the binding of a series of ligands to a receptor).

## References


(1) Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.

(2) Fukunaga, K. *Introduction to Statistical Pattern Recognition, Second Edition*, 2 edition.; Academic Press: Boston, 1990.

(3) Murphy, K. P. *Machine Learning: A Probabilistic Perspective*; MIT Press, 2012.

(4) Theodoridis, S. *Machine Learning: A Bayesian and Optimization Perspective*; Academic Press, 2015.

(5) Bishop, C. M. Latent Variable Models. In *Learning in Graphical Models*; Jordan, M. I., Ed.; NATO ASI Series; Springer Netherlands, 1998; pp 371–403.

(6) Ahmad, M.; Helms, V.; Kalinina, O. V.; Lengauer, T. Elucidating the Energetic Contributions to the Binding Free Energy. *J. Chem. Phys.* **2017**, *146* (1), 014105.

(7) Ahmad, M.; Helms, V.; Lengauer, T.; Kalinina, O. V. How Molecular Conformational Changes Affect Changes in Free Energy. *J. Chem. Theory Comput.* **2015**, *11* (7), 2945–2957.

(8) Still, S.; Sivak, D. A.; Bell, A. J.; Crooks, G. E. Thermodynamics of Prediction. *Phys. Rev. Lett.* **2012**, *109* (12).

(9) Wu, D.; Kofke, D. A. Phase-Space Overlap Measures. I. Fail-Safe Bias Detection in Free Energies Calculated by Molecular Simulation. *J. Chem. Phys.* **2005**, *123* (5), 054103.

(10) Jarzynski, C. Rare Events and the Convergence of Exponentially Averaged Work Values. *Phys. Rev. E* **2006**, *73* (4), 046105.

(11) Crooks, G. E. Entropy Production Fluctuation Theorem and the Nonequilibrium Work Relation for Free Energy Differences. *Phys. Rev. E* **1999**, *60* (3), 2721.

(12) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22* (1), 79–86.

(13) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, N.J, 2006.

(14) Kullback, S. *Information Theory and Statistics*, Revised edition.; Dover Publications: Mineola, N.Y, 1997.

(15) Kirkwood, J. G. On the Theory of Strong Electrolyte Solutions. *J. Chem. Phys.* **1934**, *2* (11), 767–781.

(16) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426.

(17) Berger, R.; Casella, G. *Statistical Inference*, 2 ed edition.; Cengage Learning, Inc: Australia ; Pacific Grove, CA, 2001.

(18) Lehmann, E. L.; Casella, G. *Theory of Point Estimation*, 2nd ed. 1998. Corr. 4th printing 2003 edition.; Springer: New York, 2003.

(19) Barndorff-Nielsen, O. E. *Information and Exponential Families in Statistical Theory*; John Wiley & Sons Ltd: Chichester, 1978.

(20) Brown, L. D. Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lect. Notes-Monogr. Ser.* **1986**, *9*, i–279.

(21) Kullback, S. Inequalities of Information Theory. In *Information theory and statistics*; Dover Publications, 1997; pp 36–59.

(22) *Entropy*; Greven, A., Warnecke, G., Keller, G., Eds.; Princeton University Press: Princeton, 2003.

(23) Sugiyama, M.; Suzuki, T.; Kanamori, T. *Density Ratio Estimation in Machine Learning*; Cambridge University Press, 2012.

(24) Rockafellar, R. T. *Convex Analysis*; Princeton University Press, 1970.

(25) Direct Density-Ratio Estimation with Dimensionality Reduction. In *Density Ratio Estimation in Machine Learning*; Sugiyama, M., Suzuki, T., Kanamori, T., Eds.; Cambridge University Press: Cambridge, 2012; pp 89–116.

(26) Martin, R. S.; Wilkinson, J. H. Reduction of the Symmetric Eigenproblem Ax = ΛBx and Related Problems to Standard Form. In *Linear Algebra*; Bauer, P. D. F. L., Ed.; Handbook for Automatic Computation; Springer Berlin Heidelberg, 1971; pp 303–314.
(27) Stewart, G. W. *Matrix Algorithms Volume 2: Eigensystems*; SIAM, 2001.
(28) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press, 1979.
(29) Harville, D. A. *Matrix Algebra From a Statistician's Perspective*; Springer Science & Business Media, 2008.
(30) Sugiyama, M.; Kawanabe, M.; Chui, P. L. Dimensionality Reduction for Density Ratio Estimation in High-Dimensional Spaces. *Neural Netw.* **2010**, *23* (1), 44–59.
(31) Feng, E.; Crooks, G. Length of Time's Arrow. *Phys. Rev. Lett.* **2008**, *101* (9).
(32) Dryden, I. L.; Mardia, K. V. *Statistical Shape Analysis: With Applications in R*; John Wiley & Sons, 2016.
(33) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**.
(34) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential Dynamics of Proteins. *Proteins Struct. Funct. Bioinforma.* **1993**, *17* (4), 412–425.
(35) Kitao, A.; Go, N. Investigating Protein Dynamics in Collective Coordinate Space. *Curr. Opin. Struct. Biol.* **1999**, *9* (2), 164–169.
(36) Brooks, B.; Karplus, M. Harmonic Dynamics of Proteins: Normal Modes and Fluctuations in Bovine Pancreatic Trypsin Inhibitor. *Proc. Natl. Acad. Sci.* **1983**, *80* (21), 6571–6575.
(37) Krivobokova, T.; Briones, R.; Hub, J. S.; Munk, A.; de Groot, B. L. Partial Least-Squares Functional Mode Analysis: Application to the Membrane Proteins AQP1, Aqy1, and CLC-Ec1. *Biophys. J.* **2012**, *103* (4), 786–796.
(38) Ahmad, M.; Helms, V.; Kalinina, O. V.; Lengauer, T. The Role of Conformational Changes in Molecular Recognition. *J. Phys. Chem. B* **2016**, *120* (9), 2138–2144.
(39) Ahmad, M.; Helms, V.; Lengauer, T.; Kalinina, O. V. Enthalpy–Entropy Compensation upon Molecular Conformational Changes. *J. Chem. Theory Comput.* **2015**, *11* (4), 1410–1418.
(40) Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press, USA, 2010.
(41) *Free Energy Calculations Theory and Applications in Chemistry and Biology*; Chipot, C., Pohorille, A., Eds.; Springer Series in Chemical Physics.
(42) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
(43) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818.
(44) Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; Bünau, P. von; Kawanabe, M. Direct Importance Estimation for Covariate Shift Adaptation. *Ann. Inst. Stat. Math.* **2008**, *60* (4), 699–746.
(45) Team, R. C. R: A Language and Environment for Statistical Computing. **2013**.
(46) Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 2.1*; 2018.
(47) Gapsys, V.; de Groot, B. L. Optimal Superpositioning of Flexible Molecule Ensembles. *Biophys. J.* **2013**, *104* (1), 196–207.
(48) Brignell, C. J.; Dryden, I. L.; Browne, W. J. Covariance Weighted Procrustes Analysis. In *Riemannian Computing in Computer Vision*; Turaga, P. K., Srivastava, A., Eds.; Springer International Publishing, 2016; pp 189–209.
(49) Dennis, J.; Schnabel, R. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; Classics in Applied Mathematics; Society for Industrial and Applied Mathematics, 1996.
(50) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, L. S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A. *LAPACK Users' Guide*; SIAM, 1999.