DROPP: Structure-aware PCA for Ordered Data

A General Method and its Applications in Climate Research and Molecular Dynamics

Anna Beer
University of Vienna, Aarhus University
Vienna, Austria; Aarhus, Denmark
anna.beer@univie.ac.at

Olivér Palotás
Aarhus University, LMU Munich
Aarhus, Denmark; Munich, Germany
oliver.palotas@gmail.com

Andrea Maldonado
LMU Munich, MCML
Munich, Germany
maldonado@dbs.ifi.lmu.de

Andrew Draganov
Aarhus University
Aarhus, Denmark
draganovandrew@cs.au.dk

Ira Assent
Aarhus University
Aarhus, Denmark
ira@cs.au.dk

2024 IEEE 40th International Conference on Data Engineering (ICDE) | 979-8-3503-1715-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDE60146.2024.00093

***Abstract*—Ordered data arises in many areas, e.g., in molecular dynamics and other spatial-temporal trajectories. While data points that are close in this order are related, common dimensionality reduction techniques cannot capture this relation or order. Thus, the information is lost in the low-dimensional representations. We introduce DROPP, which incorporates order into dimensionality reduction by adapting a Gaussian kernel function across the ordered covariances between data points. We find underlying principal components that are characteristic of the process that generated the data. In extensive experiments, we show DROPP's advantages over other dimensionality reduction techniques on synthetic as well as real-world data sets from molecular dynamics and climate research: The principal components of different data sets that were generated by the same underlying mechanism are very similar to each other. They can, thus, be used for dimensionality reduction with low reconstruction errors along a set of data sets, allowing an explainable visual comparison of different data sets as well as good compression even for unseen data.**

***Index Terms*—dimensionality reduction, PCA, ordered data, random walks, correlation, molecular dynamics, climate data**

I. Introduction

Dimensionality reduction (DR) is a fundamental task that helps inspect and analyze data sets, e.g., by using it for meaningful visualizations or reduction onto the axes that contain the most information. Criteria for successful dimensionality reduction methods focus on maintaining information in the data: Objectives such as preserving pairwise distances or minimizing reconstruction error are effectively solved by standard tools such as PCA [1]. However, non-trivial data formats often require more sophisticated objectives in order to be properly analyzed without losing important information.

Many real-world data sets are manifestations of a single underlying mechanism. For example, in molecular dynamics (MD), all trajectories of a certain protein are governed by the physical constraints imposed by the specific order of its atoms. The same can be said of any set of chronological events that repeat at a regular interval, e.g., hourly temperature patterns over a year. These data sets have in common that they are inherently ordered and that points that are close in this order are more similar than those that are far. Given such manifestations, one key objective is to uncover information regarding the underlying process rather than extract the particulars of any specific manifestation. Furthermore, it can be important to regard several manifestations in comparison with each other, i.e., in a common, low-dimensional space. For this, the ordered nature of such processes needs to be modeled in a way that is robust to noisy differences between manifestations while easily generalizing to unseen data.

We propose DROPP (**D**imensionality **R**eduction for **O**rdered **P**oints with **P**CA), a DR method that fulfills these requirements by exploiting the order of data. DROPP relies on the premise that, for ordered data, the covariance between two elements depends on their distance in that order. By fitting a Gaussian to these similarities, we eliminate the instance-specific noise and obtain a covariance matrix whose principal components describe the essence of the generating process. An overview of the main steps of DROPP is given in Figure 1. In extensive experiments, we show that the components found by DROPP are similar between different manifestations of the same process. In consequence, they are suitable for dimensionality reduction of yet unseen data generated by the same mechanism. Low reconstruction errors confirm this on synthetic as well as real-world datasets from climate research. Furthermore, we extend DROPP to tensor data and apply it to protein trajectories from molecular dynamics (MD), showing similarly good results.

Our main contributions include:

1) We introduce DROPP, a dimensionality reduction method for ordered data, whose principal components characterize the generating mechanism of the data.
2) DROPP allows a better comparison of data sets generated by an underlying process as well as generalizability to yet unseen data generated by the same process.
3) Experiments show that DROPP is effective for both time series data like climate data, and tensor data from MD.

The full code is publicly available on github.

2375-026X/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDE60146.2024.00093

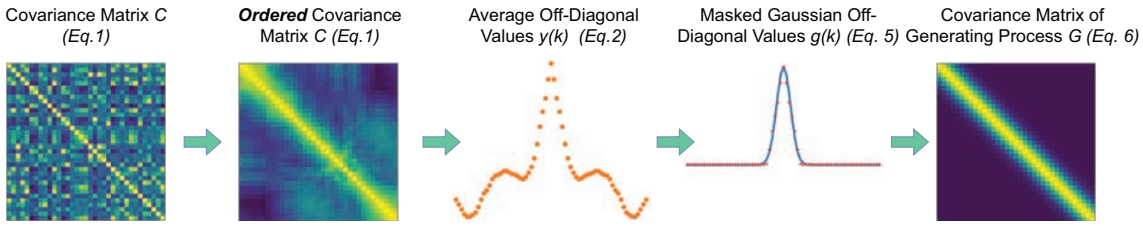


Fig. 1. Main idea of DROPP. The order is crucial and allows us to remove instance-specific noise, emphasizing the generating process' relevant information.

II. Ordered data

While most common data mining and dimensionality reduction methods are aimed at (unordered) data sets, data from more and more research areas is *ordered*: e.g., data from molecular dynamics (MD) [2], gene sequencing [3], or sensor data. With ordered data, we describe any data set $\mathcal{X}$, for which there exists a strict total or linear order $\cdot < \cdot$.

A. Temporal data

The most common type of ordered or sequential data is temporal data, e.g., time series or trajectories, which have a chronological order. While there is a plethora of literature in these research areas, temporal data is only a (proper) subgroup of ordered data and, thus, the use-cases and methods working on temporal data are not necessarily suitable for general ordered data, as time has specific additional properties compared to general (totally) ordered data:

1) Chronological order is a partial order: things can happen at the same time, data objects can have the same time stamp.
2) Time has an unambiguous "direction". While each order can be inverted, many events typical for temporally ordered data "break" if we invert the order, i.e., there is no inverse. Other ordered data can be interpreted bi-directionally, here, using "$\cdot < \cdot$" or "$\cdot > \cdot$" are conventions. These general orders and their inverses are equivalent, which does not hold true for the chronological order.
3) Chronological order comes from a global, continuous distribution; the order exists independently of any sampling of the data. In contrast, a general order might only be defined for a specific, discrete set.

These properties cause that a large area of research on temporal data is not necessarily useful for ordered non-temporal data: e.g., predictions or frequent pattern mining are mostly relevant because of temporal aspects. While we can speak of "changes" of an object over *time*, this does not make sense for most other orders. In contrast to temporal data, use cases for ordered non-temporal data are similar to the use cases of non-ordered data, but should incorporate the ordered nature of the data. In this paper, we focus on the non-temporal order of data. Note that approaches for general ordered data are also suitable for temporal data, but not the other way around.

B. Protein data

A protein is a macromolecule and can roughly be approximated as a chain of atoms with chemical bonds between them [4], [5]. The atoms that are most relevant for the shape of the backbone of a protein are the C^α atoms [6], which are connected by peptide bonds of roughly fixed length [4], [7]. These C^α atoms are inherently ordered along the chain. E.g., the distance between any two atoms' spatial coordinates depends on this order, even if it is not exclusively given by the order.

C. Molecular Dynamics data

Data from molecular dynamics (MD) describes the trajectory of each atom of a molecule over time. These trajectories are calculated with complex mathematical and physical models based on Newton's law of motion and the underlying structure of the molecule [8]. The temporal resolution usually lies in the femtosecond range and calculation of trajectories of a useful length only became feasible during the last years because of advances in computation hardware. Each molecule trajectory is a manifestation of the underlying computation of the molecule's movements which simulates its real-world movements. Moving molecules can be approximated as ideal chains [9] or semiflexible chains [10], which already points towards the connection to random walks.

Various features besides the spatial coordinates in three-dimensional space can be regarded for analysis, e.g., the accessible surface area or dihedral angles [11]. In this paper, we focus on the spatial coordinates, as most other properties are a consequence of those. Furthermore, we focus on the coordinates of the C^α atoms, which are decisive for most data analysis tasks [12], [13]. MD data is usually big, as the tensors describing it contain the above features for several ten to several hundred atoms over (ten) thousands of time steps. Furthermore, MD data is especially interesting because two of the three dimensions of the tensor are ordered: time steps are ordered chronologically and atoms are ordered by their appearance in the backbone of the protein. While there are dimensionality reduction (DR) methods for MD data that regard the chronological order, e.g., tICA [14] and time-lagged tSNE [15], there are, to the best of our knowledge, no DR methods incorporating the order of the atoms, yet. This comes with two main problems: 1) Important information in the data is neglected: The list of atoms has a strict order, i.e., the atoms form a connected chain, which cannot be captured by traditional DR methods. 2) Projections of different trajectories of a protein cannot be compared among each other, as the PCs differ heavily. Low-dimensional projections

of MD data are used to analyze different trajectories of a protein, e.g., when studying its folding process. Thus, DR methods for protein trajectories that incorporate the order can have a large impact on important use cases of MD research, which includes understanding protein misfolding diseases like Parkinson's [16] or drug discovery [17].

III. Related Work and Background

A. Dimensionality Reduction

Dimensionality reduction can have diverse goals, e.g., visualization, data compression, or accelerating downstream tasks that are not suitable for high dimensionality. These goals are intertwined: e.g., DR for visualization or downstream tasks should keep as much information as possible, i.e., it also needs to be a good compression of the data and have a low reconstruction error. DR techniques can be based on linear transformations of the data or non-linear embeddings: while, e.g., PCA [1], ICA [18], and tICA [19], [20] reduce dimensionality by linear projections onto subspaces of the data, modern techniques like tSNE [21] and UMAP [22] transform the data non-linearly.

1) Linear Dimensionality Reduction: The probably most often used linear DR method is PCA [1] (principal component analysis). It calculates the vectors along which the variance of the data is maximized. Projecting the data onto those 'principal' or 'main' components minimizes the reconstruction error for a given data set. There are numerous extensions of PCA that improve diverse aspects like its robustness [23]–[25], runtime [26], [27], or applicability to non-numeric data [28]. ICA [18] (independent component analysis) finds statistically independent components that describe the data. Here, the statistical independence between the components is maximized, e.g., by minimizing the mutual information between them. Note that the components returned by ICA are not orthogonal to each other like those returned by PCA. There are several adaptions and versions of ICA, e.g., FastICA [18] (introduced in the same paper), noise-robust versions [29] or those listed in [30]. ICA is also the basis for time-based ICA (tICA) [19], which is commonly used for MD data [14]. In our experiments in Sections VI, VII, and VIII, we compare our novel linear DR technique DROPP to PCA, FastICA, and tICA. .

2) Non-linear embeddings: Locally Linear Embeddings (LLEs) [31] preserve the local structure of each point but are not globally linear. Modern data analysis in biology and other life sciences has begun to strongly prefer tSNE [21] or UMAP [22] for visualizing high-dimensional data sets. These methods learn non-linear transformations that aim to preserve non-linear similarities in the high-dimensional space, often done by emphasizing the distances from a point to its k nearest neighbors. Due to their non-linearity and reliance on gradient descent, tSNE and UMAP do not produce main components that can be used to project the space globally consistently. While these methods yield visualizations that often reveal cluster structures from the high-dimensional space, they are not easily interpretable or transparent for the user. Even more, they can produce structures in the data that are not existent in high-dimensional space and can, thus, be misleading for downstream tasks like clustering or outlier detection [32].

Note that all of the described methods, linear as well as non-linear, are designed for unordered data sets – their results do not change when the data is shuffled. While this makes them robust to permutations in the data, it also prevents them from finding mappings that have a dependency on the order.

B. Dimensionality Reduction on Temporal Data

To the best of our knowledge, the only DR methods that take any order into account are explicitly focusing on the temporal aspect. Among these, the most pertinent are time-series factoring methods such as, e.g., the Generalized Dynamic Factor Model [33] [34], ForeCA [35], and Dynamic Orthogonal Components [36]. They intend to separate a multivariate time series into a set of independent underlying factors and white noise and are closely related to both PCA and ICA. However, we note that these differ significantly from our method for two reasons. First, our method is suitable for data with a more general order than time-series data (see Section II). Second (and more importantly), we do not assume that there are multiple processes at play that must be separated – we instead assume that there is a single process that has been sampled multiple times. We emphasize this difference in experimental Sections VII and VIII, where PCA, ICA, and also tICA (that takes the time aspect into account) perform consistently worse for our use-cases than DROPP.

C. Random Walks

A frequently encountered origin of ordered data are random walks. A random walk is "a random process which describes a path including a succession of random steps in the mathematical space" [37]. These steps in the path are ordered and random walks are used to approximate ordered data like climate data [38], [39] or molecular dynamics trajectories [20]. For our purposes, the direction of the steps is Gaussian distributed, so that the position of a random walker then depends on the step at which the random walker is regarded. This is a suitable simulation as it makes no assumption about the dynamics corresponding to this order. The covariance matrices and principal components of random walks have been studied intensively for various use-cases, e.g., genetic variation on stream networks [40], neural network planning [41], and protein analysis [20].

D. Dimensionality Reduction for MD data

DR for MD data has several use cases:

- Visualization: In MD research, visualization of protein trajectories takes an important place. Many investigations are still done by visual analysis of experts, e.g. [42], [43].
- Descriptors: Descriptors can be found by DR and are important for distinguishing different conformations of a protein. A conformation describes the arrangement of a protein's atoms in space [11].

- Preprocessing for downstream tasks: As MD data is usually very high-dimensional, DR methods are automatically used within most frameworks, e.g., for subsequent clustering or calculation of Markov State Models [44].

Even though PCA is a decades-old, elementary method and MD is a research area that only recently developed and produces complex structured data, PCA is still often the standard method for research in the area [45]. Among others, it is a valuable tool for investigating the energy landscape of a molecule, which is crucial for analyzing protein folding. Many further DR methods for MD data are PCA-based and the original PCA is frequently used for state-of-the-art research in MD, e.g., [45], [46]. [45], [47] and [48] explain how PCA can be used for MD data: Even though PCA is developed for *matrix* data, it is often 'blindly' applied to MD *tensor* data. This can lead to suboptimal results, as information on the structure of the data gets lost. E.g., applying PCA on the flattened data mistakenly assumes dependency between all coordinates, even though covariances between one atom's x-coordinate and another atom's y-coordinate could be relicts from preprocessing. Furthermore, using PCA prevents an explainable and fair comparison between different trajectories of the same protein: the axes onto which the data is projected differ heavily between different trajectories. Because of the complexity of molecules, a simple, linear PCA on flattened MD data cannot capture the correlations between atoms adequately [49]. Besides these widely used DR methods, there are also other DR methods regarding, e.g., the correlations between atoms based on their spatial distances [50] or as collectives [51]. While MD data is known to have properties similar to a random walk over time [20], this concerned so far only the temporal axis. In contrast, in this paper, we regard the physical order of the atoms themselves.

IV. Structure-aware PCA for Ordered Data

Ordered data often follows an underlying pattern due to its generating process. E.g., in time series data, such as hourly temperature measurements, neighboring time steps often have similar values. Likewise, in molecular dynamics (MD), consecutively listed atoms of a protein have similar coordinates in space, as they are connected by chemical bonds. These inherent connections between data points provide valuable insights for identifying components that are representative of the generating process. They can be seen in the ordered correlation matrix, where the high values around the diagonal indicate meaningful connections rather than chance occurrences. To see this, we plot an example matrix for an ordered dataset showing the peak that occurs along the diagonal in the first part of Figure 1. In order to capture these connections, we regard the correlation between parameters as a function of their distance in the order. We mask the covariance matrix by a set of 1D Gaussian distributions (described in Section IV-A1) to reduce the influence of noise. We then perform the PCA on this masked matrix instead of the original covariance matrix. This emphasizes the intrinsic connections within the ordered data and allows us to extract information on the underlying process using only a few samples from potentially very large data sets. Especially for real-world use cases, the ground truth is usually not known. Thus, explorative and unsupervised data mining methods are necessary. PCA does not require any external information and preserves the global structure of the data while minimizing reconstruction loss, which makes it a fundamental and powerful tool. We show in Section V that the principal components computed by DROPP are representative of the generating process and remain similar for different manifestations generated by the same process.

A. General approach: DROPP

In this section, we describe our new dimensionality reduction method DROPP in detail. It captures the order of the data by connecting the correlations of objects to their distances in the order. By adapting a Gaussian kernel to this function, we account for the strong connections between close elements and eliminate noise related to individual manifestations instead of the underlying process. DROPP consists of three steps: (1) calculating the covariance matrix $\mathcal{C}$ (2) calculating $\mathcal{G}$, the underlying Gaussian distribution of $\mathcal{C}$, and (3) calculate eigenvectors of $\mathcal{G}$ that give us the underlying principal components.

1) Calculating the Covariance-Matrix: Given a data set $\mathcal{X} \in \mathbb{R}^{n \times d}$ with n data points in d ordered dimensions, the covariance matrix $\mathcal{C} \in \mathbb{R}^{d \times d}$ of $\mathcal{X}$ is defined as

$$\mathcal{C} = Cov(\mathcal{X}) = \mathcal{X}^\top \cdot \mathcal{X} \tag{1}$$

We normalize the rows of $\mathcal{X}$ before calculating the covariances.

2) Underlying distribution of Covariance Matrix: Figure 1 shows the unordered and ordered covariance matrices of some real-world data used in Section VIII. While both unordered and ordered data, have only values of 1 on the diagonal, the type of ordered data we are regarding here also has very high values (i.e., close to 1) on the band of the covariance matrix, i.e., the first few off-diagonals. We note that the covariance depends primarily on the distance in order between two elements. Thus, we regard the covariance as a function based on the order by taking the average value along each off-diagonal, where $k \in [-(d-1), d-1]$:

$$y(k) = mean(c_{ij} \in \mathcal{C}), \text{ where } i - j = k \tag{2}$$

The high covariance values that are caused by the order occur up to a certain k_0. Before that, the value of $y(k)$ decreases with increasing distance between the elements following a Gaussian function. We determine this k_0 as the first value where $y(k)$ is lower than or equal to 0 or the minimal $y(k)$:

$$k_0 = min\{|k| : y(k) \leq \min(0, min_k(y(k)))\}$$

Objects described within the band of width k_0 are strongly related to their neighbors in the given order and give meaningful information about the underlying generating process. Covariance values outside of this band depend more on the particular manifestation than on the underlying process. Thus,

we nullify the values outside of the band, as they can differ heavily between different manifestations.

$$y_G(k) = \begin{cases} y(k) & |k| < k_0 \\ 0 & k_0 \leq |k| < d \end{cases} \tag{3}$$

We fit a Gaussian function $g(k)$ to $y_G(k)$ s.t. the root mean squared error (RMSE) between them is minimized:

$$RMSE\left(g(k), y_G(k)\right) = \left(\frac{1}{d}\sum_{k=0}^{d-1}(g(k) - y_G(k))^2\right)^{\frac{1}{2}} \tag{4}$$

$$g(k) = e^{-\left(\frac{k}{2\sigma}\right)^2} \tag{5}$$

The matrix $\mathcal{G} \in \mathbb{R}^{d\times d}$ describes the covariance of the underlying process, i.e., it is representative of all the manifestations:

$$\mathcal{G}_{ij} = g(|i-j|) \tag{6}$$

In this sense, we are learning the σ that describes the width of the band along the diagonal.

3) PCA: As a result, we return the principal components of $\mathcal{G}$. PCA is an established method for dimensionality reduction and suitable for capturing linear correlations in the data.

4) Complexity and Scalability: The complexity of DROPP is comparable to the complexity of PCA, as calculating y is in $O(dn^2)$, calculating k_0 is in $O(d^2)$, and the fitting of a Gaussian to y_G is $O(d \log d)$ by a binary search over parameters. Independently of the complexity, DROPP scales well to large data sets as it finds a high-quality projection with very few samples (discussed further in Section VIII-C). Furthermore, once DROPP is calculated on one manifestation of a data-generating process, it can be used for all other manifestations, too. Thus, e.g., in a streaming setting, we can compute the principal components from the first batch of instantiations and apply the components to all subsequent instantiations.

B. DROPP for tensor data

MD data is usually given as a tensor describing the movements of a molecule's atoms. It consists of three spatial coordinates for every atom of an ordered set $a \in \mathcal{A}$ for every time step $t \in [0, T[$. As explained in Section II-C, we only regard the carbon-alpha atoms of each molecule, as they are decisive for the shape and state of a molecule, and in the following refer to them as atoms.

We present the adaptions of our method to fit tensor data describing MD trajectories. While for PCA, the atoms' information for every time step is simply flattened in order to calculate the covariance matrix [45], we regard the covariances between the atoms as total. For the data along one spatial coordinate c we write: $\mathcal{X}_{:,:,c} \in \mathbb{R}^{T\times A}$. As molecules are intrinsically three-dimensional objects in space and the relevant aspects are the relative motions of atoms within the molecule, we can furthermore facilitate some parts of the covariance matrix by setting covariances between different spatial axes to 0. The adapted covariance matrix $\widehat{\mathcal{C}} \in \mathbb{R}^{3|\mathcal{A}|\times 3|\mathcal{A}|}$ is calculated as follows, where we write $\lfloor\cdot\rfloor$ for the floor function:

$$\widehat{\mathcal{C}}_{ij} = \begin{cases} 0 & \text{if } i \not\equiv j \mod 3 \\ \frac{1}{3}\sum_{c=1}^{3} Cov(\mathcal{X}_{:,:,c})_{\lfloor i/3\rfloor,\lfloor j/3\rfloor} & \text{else} \end{cases} \tag{7}$$

Correspondingly, $\widehat{y}(k) = mean(c_{ij} \in \widehat{\mathcal{C}})$, where $i - j = 3k$ and

$$\widehat{\mathcal{G}}_{ij} = \begin{cases} 0 & \text{if } i \not\equiv j \mod 3 \\ g(|\lfloor i/3\rfloor - \lfloor j/3\rfloor|) & \text{else} \end{cases} \tag{8}$$

Experiments in Figure 7 show the covariance matrices and the respective Gaussian fits for the state-of-the-art of calculating the covariance matrix vs our method.

C. Qualitative Analysis of DROPP

Compared to other DR methods, DROPP has several differences and advantages: 1) It is the first DR method that inherently relies on order and gives a novel basis for capturing information on the order of data. 2) Once fit to a manifestation of a data-generating process, the remaining manifestations can be projected through the pre-determined basis. 3) It is inherently noise robust as the covariance masking process automatically eliminates spurious similarities. 4) It is directly expandable to most variations of PCA that are based on the covariance matrix. This makes it amenable to improvements in, e.g., speed, robustness, or distributed settings. At the same time, it can replace PCA for a variety of methods that use PCA as an intermediate step. 5) While being fully unsupervised, it accomplishes results that are comparable to supervised DR – the gold standard of any unsupervised method.

We note that many of these differences are in contrast to other existing approaches. Non-linear DR methods like tSNE [21] and UMAP [22] do not provide a natural method for projecting future batches. Additionally, most linear and non-linear DR methods are not able to achieve noise robustness or order-dependent embeddings. We revisit these points in Section VIII-D, where we reference experimental evidence in support of the above claims.

V. Experimental Setup

We compare DROPP to other dimensionality reduction methods PCA, ICA, and tICA on three types of data: Synthetic datasets in Section VI, climate data in Section VII, and molecular dynamics data in Section VIII.

A. Setup and Implementation Details

We run real-world experiments on a machine with Inter(R) Core(TM) i7-6700 CPU @3.40Ghz using 16 GB RAM and a machine with Intel(R) Core(TM) i7-1165G7 @ 2.80GHz using 16 GB RAM. We used PCA and FastICA implementations from sklearn[1]. For TICA, we used the implementation from PyEmma[2]. Our code is publicly available at https://github.

[1]https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition, last accessed: July 26th, 2023

[2]http://www.emma-project.org/latest/api/generated/pyemma.coordinates.transform.TICA.html, last accessed: July 26th, 2023

com/poliver269/DROPP and compatible with the scikit-learn interface TransformerMixin.

B. Quality measures

We study two quality measures: similarity of principal components and reconstruction error on different manifestations.

1) Similarity of principal components: We measure the similarity of sets of components for different manifestations. A high similarity implies that the DR method represents the overall data-generating process rather than a specific data sample. Assume manifestations A and B of a single ordered data-generating mechanism. Then a DR method that captures the data-generating process itself should return similar components regardless of which sample from the process it is given. To measure this, we report the cosine similarity among the top-k components calculated for A and B. Note that A's eigenvalue order may not be equivalent to B's. Thus, when matching components for the cosine similarity, we iteratively find which component in B is most similar to the given component in A. We remove that component from B when looking at A's next components. This gives us a set of k cosine similarity values, from which we return the median.

2) Reconstruction error on different manifestations: Although samples A and B from a single generating process may have similar principal components, this does not necessarily mean that those components are effective at describing the data. To this end, we also report the mean squared reconstruction error when applying the projection. For this, we calculate the PCs of one manifestation and measure the RE of all other manifestations when projected onto these PCs. We do this for every available manifestation of the set and report the average. This gives an indication of how well the PCs calculated on any manifestation are suitable for other manifestations of the same generating process.

VI. Experiments I: Synthetic data

We begin our experiments by analyzing DROPP's performance on synthetic data sets. To this end, we generate two ordered synthetic data sets with significantly different underlying mechanisms. The first synthetic data set is generated by a *random walk*. We randomly sample an initial position on the d-dimensional unit hypersphere and proceed to take n steps of magnitude 0.1 in random directions along the hypersphere. For the second synthetic data set – *trajectory splines* – we sample landmarks and define a path between them. To do this, we randomly sample $l \ll n$ 'landmarks' from an d-dimensional standard normal distribution. We then fit a cubic spline to these landmarks and take n samples along the spline. In all synthetic experiments, we default to $d = 20$, $n = 300$, $l = 30$ and generate 10 manifestations in each setting. We ensure that the dimensions are the ordered elements by regarding $\mathcal{X}^\top$. We note that, whereas the random walk has consistent step sizes along a pre-determined surface, the trajectory splines are much more unruly – they may have unbounded second derivatives and large variance in the step sizes.

TABLE I
Characteristics of Climate Data

Feature	Countries	Years	H
temperature	all 28 countries	2009-2019	24
radiation direct	BE, DE, DK, GB, IE, LT, LU, LV, NL	2019	17
radiation diffuse	BE, CH, DE, FR, HU, IE, LU, NL, PL, RO, SK	2019	17

For ease of presentation, we begin by comparing DROPP to PCA and include the other methods in Sections VII and VIII when verifying our results on real-world data sets. The results on the synthetic data are visualized in Fig. 2. There, we see that DROPP obtains a smaller reconstruction error on both synthetic data sets while maintaining consistently high cosine similarities among the principal components of differing random generations. Furthermore, these effects are exacerbated in the presence of noise. Indeed, adding Gaussian noise to the synthetic data sets does not lessen the divide between DROPP and PCA along both metrics[3]. We emphasize that although these two synthetic data sets differ significantly in their nature, both exhibit the diagonal dominance within the covariance matrix and allow DROPP's kernel-fitting to effectively capture the underlying mechanics.

VII. Experiments II: Climate Data

For the climate data, we use geographically aggregated weather data for European countries [52]. It contains hourly values for temperature, diffuse sky radiation, and direct radiation over the course of a year. Table I gives an overview of the climate data used for experiments in this section. For every feature, we chose a group of countries that have equally many values per day. *Years* gives the range of years we regarded and H gives the number of values per day (hours). Regular climate measurements throughout the day, like temperature, typically follow a continuous underlying process. The measurements are ordered by hour of the day, and each day, resp. the measurements for each country, give a different manifestation of the same underlying mechanism.

Figure 3 shows the covariance matrices and their average off-diagonal values for temperature, direct radiation, and diffuse radiation. The temperature data can roughly be modeled by a random walk, i.e., the covariance matrix shows a Gaussian distribution of average off-diagonal values depending on the distance between measurements. This fit is weaker for radiation measurements, as we do not only have chronological order, here: The underlying mechanism for sun radiation measurements is circadian, and, thus, has cyclic properties. This shows in high covariance values around the anti-diagonal and in increased average off-diagonal values for high absolute indices in the plots on the right. The effect is especially strong for diffuse radiation (see bottom of Figure 3), as its values

[3]The noise levels have variances $[0, 0.005, 0.01, 0.02]$ and $[0.1, 0.2, 0.4, 0.8]$ for the random walk and trajectory spline synthetic data sets respectively. Note that the random walks are on the 20-dimensional unit sphere, so each element in the data set has very low absolute value.

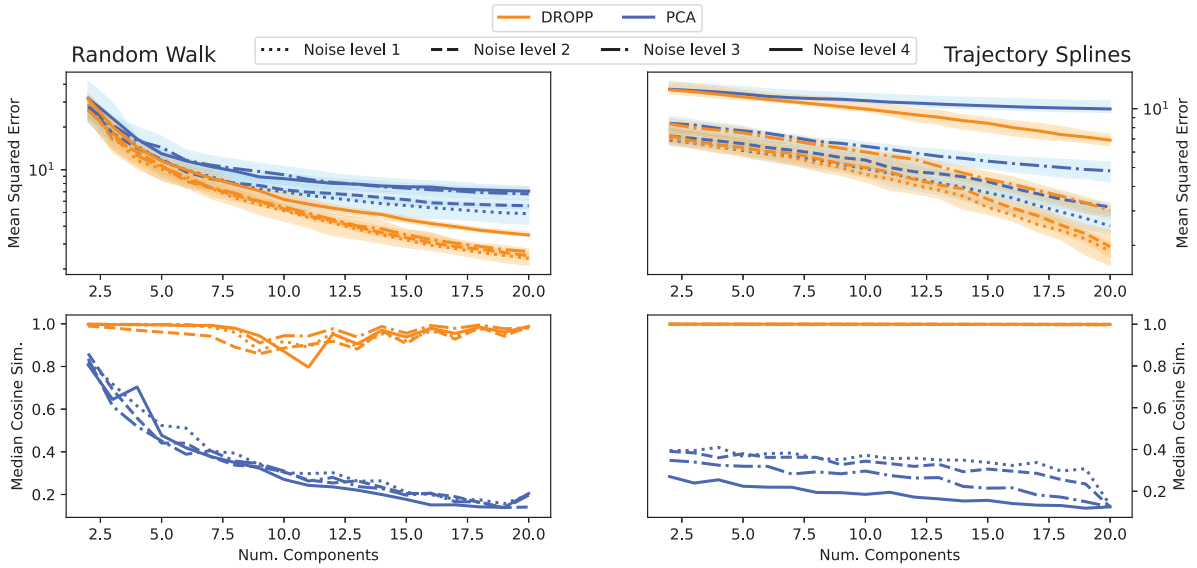


Fig. 2. Reconstruction error (top) and similarity (bottom) for different levels of noise in the random walk (left) and trajectory spline (right) synthetic data sets. Noise Levels 1-4 represent increasing amounts of random Gaussian noise added to the data positions. Bands around reconstruction error lines represent one standard deviation. We see that DROPP's mean squared error is consistently lower than PCA's while its cosine similarity is significantly higher.

are mainly influenced by the degree of latitude, and only marginally influenced by weather changes.

For all three properties in the climate data set, the course of measurements before noon is negatively correlated to the course after noon. However, temperature and direct radiation are dependent on, e.g., clouds in front of the sun, that can change over the course of a day, taking away the strong correlation seen for diffuse radiation. Thus, the high values on the antidiagonal of the covariance matrix are less pronounced for direct radiation. As temperature also depends on a multitude of other aspects (e.g., rain or wind), the negative correlation between prenoon and afternoon is almost negligible, here.

A. Similarity for Climate Data

We measure the similarity between increasingly large sets of components for temperature, direct radiation, and diffuse radiation for different countries[4] in Figure 4. In all three diagrams, DROPP consistently yields the highest similarity across principal components for different countries. Note that DROPP has a very low variance for radiation measurements, but an increased variance when applied to temperature data. This has two main reasons: 1) For temperature, we compare components between all 28 countries in the data set, while for radiation we only test on nine resp., eleven countries. 2) While the radiation values are consistent across all countries (they are all close geographically), the daily temperature progress also depends heavily on further aspects like closeness to the sea or mountains. Thus, the variance across different countries is higher for temperature than for radiation. [5]

B. Similarity of components for different countries

We investigate the similarity of the first ten PCs computed by PCA and DROPP for Great Britain (GB), Ireland (IE), Lithuania (LT) and Latvia (LV) in Figure 5. For PCA as well as for DROPP, components compressing temperature in Lithuania and Latvia are very similar across all 40 years. However, the PCs computed by PCA show a clear difference between the components for non-continental countries and the two baltic states. In contrast, DROPP returns very similar principal components uniformly across all years and countries. This allows comparing different countries' daily temperature progress without bias towards one of the countries (i.e., manifestations). Note that depending on the use case, PCA can be preferable, e.g., if only one country is regarded. However, our goal is to find common or similar principal components while simultaneously keeping the reconstruction error low.

C. Reconstruction Error for Climate Data

In Figure 6, we see that the principal components as computed by DROPP are suitable to compare temperature and direct radiation data of different countries: the reconstruction errors are comparable to those of PCA and significantly lower than for TICA and ICA. Even though the reconstruction errors for diffuse radiation data is significantly lower using DROPP components than when using components calculated by TICA or ICA, it is higher than for components calculated by PCA. This corresponds to our explanations at the beginning of this

[4]In this experiment, we exclude PL and FI for TICA for technical reasons

[5]Note that this is a difference between countries, while we previously described the difference between prenoon/afternoon daily behavior for each individual country.

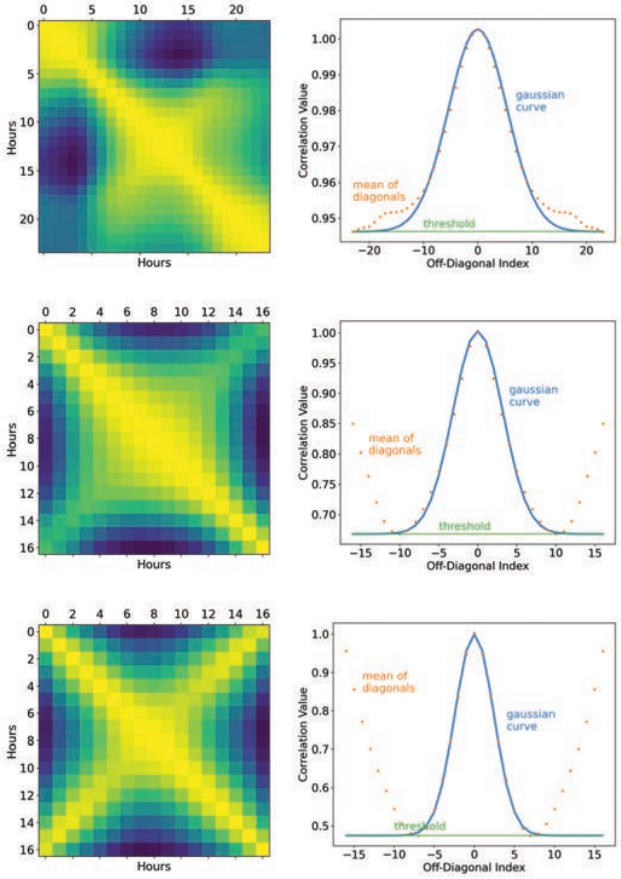
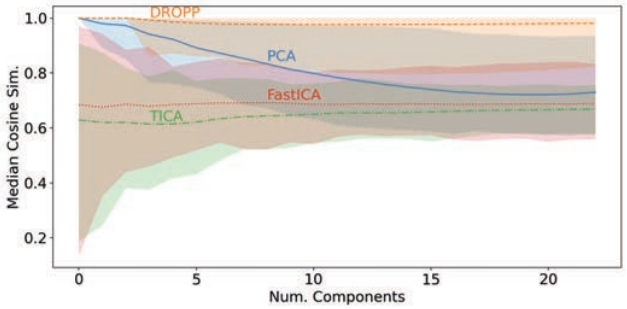
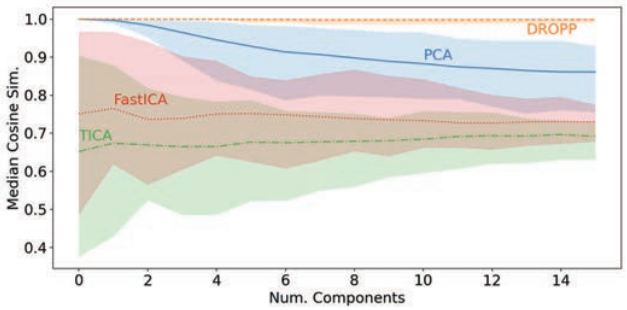


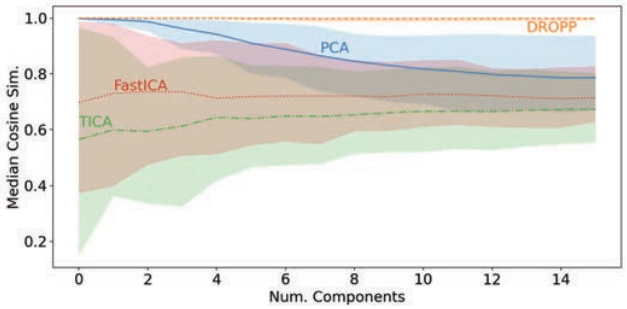
Fig. 3. Covariance matrices (left) and distributions of average off-diagonal values (right) for climate data of Germany in 2019. Top: temperature, middle: direct radiation, bottom: diffuse radiation.



(a) Temperature



(b) Direct radiation



(c) Diffuse radiation

Fig. 4. Similarities of components for increasingly large sets of components and different attributes of the climate data set.

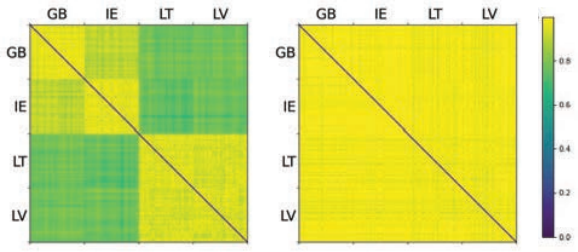


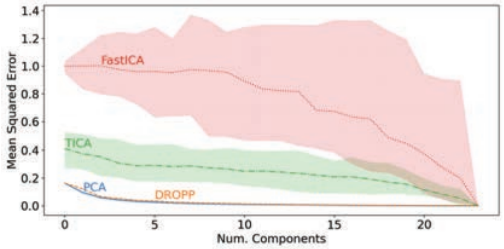
Fig. 5. Similarity matrix for temperature values over 40 years in Great Britain, Ireland, Lithuania and Latvia for the first ten principal components each. Lighter areas imply high similarity. Left: PCA, Right: DROPP.

section: the covariances for diffuse radiation are not distributed following a Gaussian curve as can be seen in Figure 3(f).

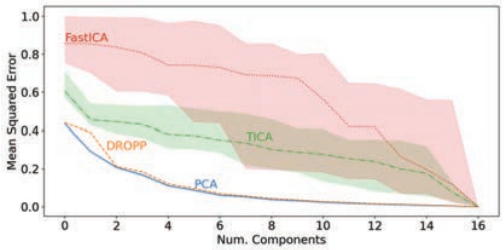
VIII. Experiments III: Tensor Data from MD

We describe MD data and its background in Section II-C in detail. For every time step, our MD trajectories describe the position of every atom in a protein with three spatial coordinates. As typical in the field, we only use coordinates of the "relevant" atoms, i.e., C^α atoms that describe the backbone of the protein, see Section II-C. A protein can be approximated by an ideal chain [9] of such C^α atoms. Thus, the atoms correspond to steps in a random walk, allowing a meaningful application of DROPP. We use trajectories describing the protein folding of three different proteins retrieved from [53] (FS-peptide) and [54](2F4K and 2WXC). Table II shows the characteristics of our data, where $\#$ is the number of available trajectories (i.e., manifestations of the generating MD simulation), A is the number of C^α atoms, and T is the number of time steps of each trajectory.

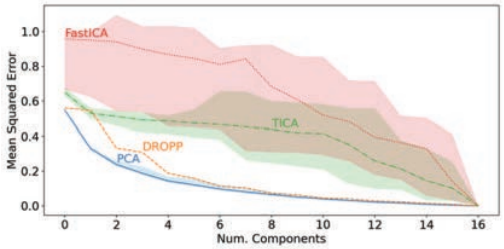
For MD research, the relative movements within a molecule are relevant (in contrast to the movement of the full molecule through space). Thus, we rotate and center the data as it is



(a) Reconstruction Error for Temperature



(b) Reconstruction Error for Direct Radiation



(c) Reconstruction Error for Diffuse Radiation

Fig. 6. Reconstruction errors on climate data depending on the dimensionality of the projected space

TABLE II
CHARACTERISTICS OF THE TRAJECTORY DATA SET

Protein name	#	A	T
2F4K	62	35	10000
FS-Peptide	28	21	10000
2WXC	136	47	40000

common in the field [55], [56], s.t. the RMSD to the molecule's atoms at a each time step is minimized as originally suggested by [57]. For our experiments, we choose this time step at random and center the molecules for every time step [58]. Finally, the data is scaled w.r.t. its standard deviation.

A. Adaptions for Tensor Data

While other DR methods in the field are simply applied to the matricified tensor data, we account for the underlying tensor shape of the data as explained in Section IV-B. Figure 7 shows why this is helpful: On the left, we see the sums over off-diagonals of the covariance matrix as it is used by state-of-the-art approaches incorporating PCA: vectorizing the coordinates leads to a mixed covariance matrix, where x, y, and z-coordinates are alternated. In contrast, we calculate the covariances between the *atoms*, instead of covariances between the atoms' individual coordinates. This is not only more meaningful, but also allows to detect the underlying covariance function – given by the blue line in Figure 7.

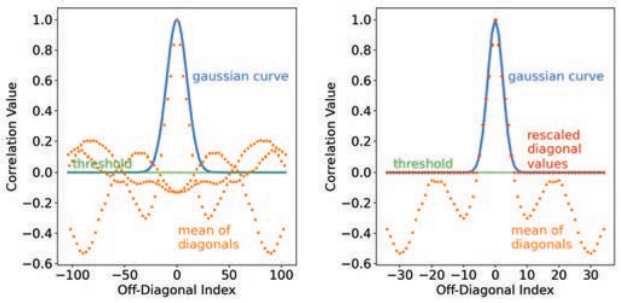


Fig. 7. Off-diagonal values of the covariance matrix. Left: calculated for flattened data. Right: According to our tensor-adaptions. There are three times as many off-diagonals in the left image because of the three spatial dimensions

B. Similarity and Reconstruction Error for MD

Figure 8 and Figure 9 show the similarity values and reconstruction errors as described in Section V for trajectories of the three proteins 2F4K, 2WXC, and FS-Peptide. In Figure 8, we see that DROPP returns consistently for all proteins and all dimensionalities of the reduced space very similar eigensystems. This implies that DROPP finds the same "important dimensions" independently of the respective manifestation. Figure 9 shows that these eigensystems are well suited for compressing the data, as the REs are consistently very low. Only for small dimensionalities of the reduced space, PCA can achieve lower REs. We note that the variance of the quality is by far the lowest for DROPP, making it a reliable tool for early phases of data exploration.

C. Different lengths of Trajectories for MD

As the covariances between different atoms can already be calculated when only a small part of the trajectory is known, DROPP returns constantly good principal components independently of the size of the known time window. This is especially valuable in molecular dynamics, where it is not known beforehand for which time frame a molecule has to be simulated until something "interesting" happens. Thus, the length of MD trajectories can be arbitrarily long. With DROPP, we do not need to know or load the full trajectory: instead of needing thousands of time steps, a few hundred are often sufficient for an equally low RE as Figure 10 shows: Here, we calculate the reconstruction error on other (full-length) trajectories of FS-Peptide when only a time window of size $x < 10000$ of one of the trajectories is known. The subfigures show the REs for different dimensionalities of the reduced space. For small known time windows with length < 1000,

[7]i.e., $x \in$ [105, 145, 201, 278, 385, 534, 740, 1024, 1418, 1964, 2720, 3766, 5215, 7221, 10000]

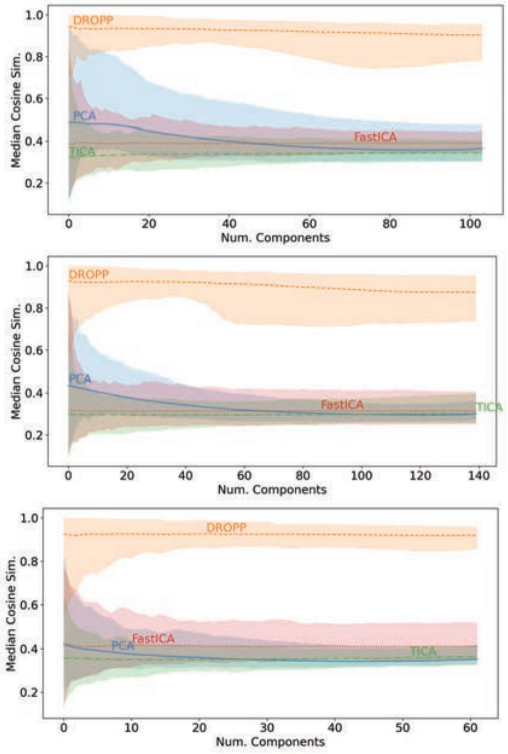


Fig. 8. Similarity between PCs of the proteins 2F4K, 2WXC, and FS-Peptide (top to bottom) depending on the dimensionality of the projection.

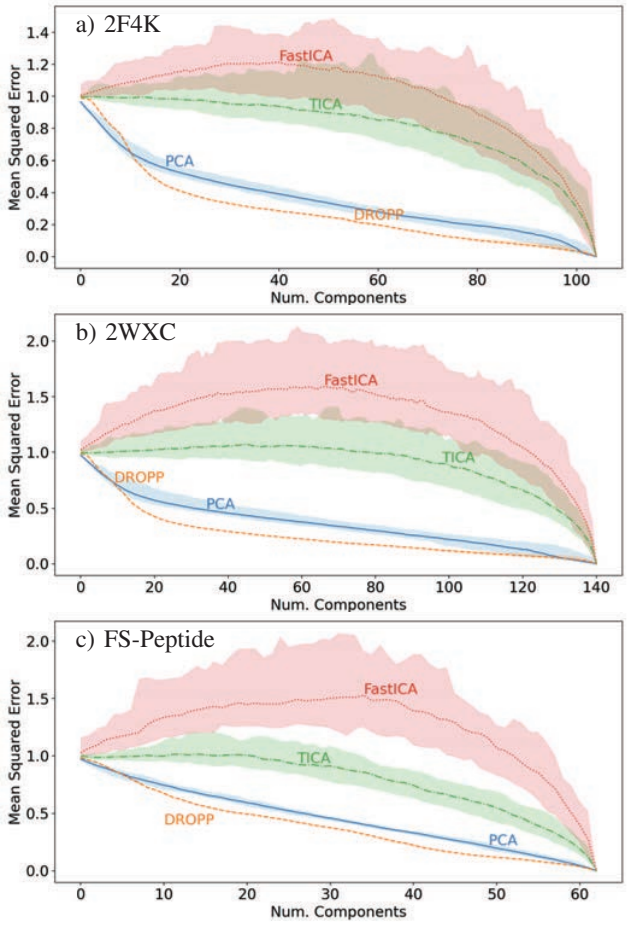


Fig. 9. Reconstruction errors on different trajectories depending on the dimensionality of reduced space for different proteins (top to bottom).

DROPP always achieves a lower RE than PCA. Only for very low-dimensional representations (less than ten), PCA can achieve lower RE than DROPP for a large enough window size. For higher dimensionalities of the reduced space, DROPP constantly results in lower RE.

D. Qualitative Analysis of DROPP for MD data

In MD research, often, several trajectories of a protein are regarded, which can be difficult considering their high dimensionality and tensor shape. With DR methods that are typically used, like PCA or tICA, each trajectory can be visualized in 2d independently. However, using different projections for every trajectory prevents a meaningful comparison between them: the principal components often have no common meaning for different trajectories of the same protein. Figure 11 visualizes this problem: It gives visualizations using linear transformations onto diverse principal components. On the left, these components were calculated with PCA, on the right, they were calculated with DROPP. Furthermore, the components were calculated for different trajectories as given by the rows and visualize different trajectories as given by the columns. Thus, the plots on the diagonal of the left part give the exact visualizations by PCA, i.e., each trajectory is transformed using the PCA fitted on itself. Each point represents a state of the protein at a certain time and the color of the point implies the time step within the trajectory. For a fair comparison of the trajectories, visualizations of a trajectory should be similar independent of the trajectory the DR method was fit to. I.e., the rows in Figure 11 should be similar, as all of them visualize the same five trajectories. However, we see heavy differences for the projections calculated by PCA, while visualizations computed by DROPP are similar across all rows. Thus, DROPP allows a meaningful comparison without strong bias to the trajectory which was used to calculate components.

Note that methods like tSNE or UMAP cannot be transferred between different data sets. In Figure 13, we show the visualizations by tSNE for the same five trajectories. These plots are frayed out and show some cluster-like structures that may or may not exist in the high-dimensional space. While tSNE visualizations can help identify structures, these cannot be taken for sure given the method's "black-box" approach (see also [32], [59]). Additionally, it is unclear how to compare the visualizations of different trajectories to one other – each fit is tailored to its individual trajectory.

In Figure 12 we show how adding external knowledge to the transformations given by PCA could also be used to

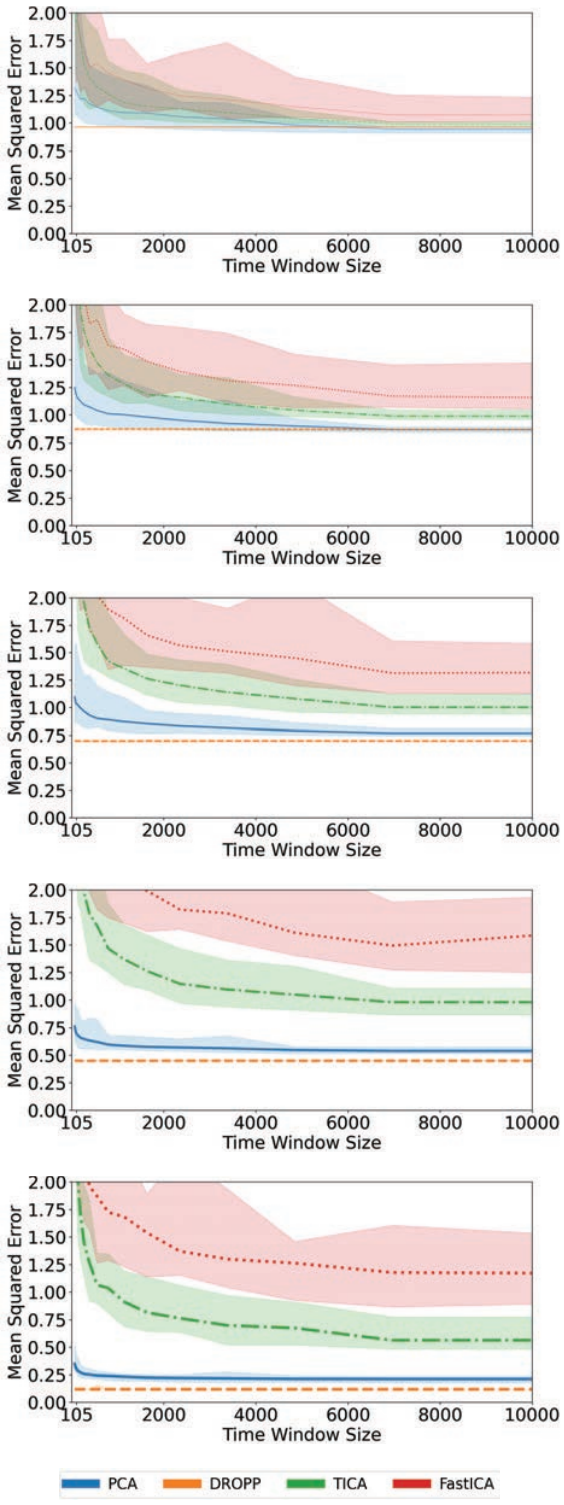


Fig. 10. Reconstruction Errors for different sizes of known time windows for the protein FS-Peptide. Line width corresponds to number of principal components $\in [2, 5, 10, 25, 50]$. Tested time window sizes were chosen following a geometric distribution[7]

TABLE III
NUMERICAL OVERVIEW OF REAL DATA EXPERIMENTS

Data		Climate			Molecular Dynamics		
Method		temp.	r. dir.	r. diff.	2f4k	2wxc	fs-pept.
2 comp.	DROPP	0.07	0.23	0.34	0.98	0.96	0.94
	PCA	**0.05**	**0.22**	**0.24**	**0.9**	**0.91**	**0.91**
	ICA	1.0	0.94	0.96	1.02	1.07	1.09
	TICA	0.34	0.49	0.5	1.0	0.99	0.99
5 comp.	DROPP	0.03	0.1	0.16	0.86	0.85	0.84
	PCA	**0.02**	**0.09**	**0.12**	**0.79**	**0.83**	**0.84**
	ICA	0.97	0.84	0.85	1.05	1.12	1.17
	TICA	0.29	0.42	0.46	1.0	1.0	0.99
10 comp.	DROPP	0.01	0.03	0.05	0.65	**0.68**	**0.67**
	PCA	**0.01**	**0.02**	**0.04**	**0.65**	0.71	0.75
	ICA	0.88	0.57	0.65	1.09	1.22	1.33
	TICA	0.26	0.29	0.35	0.99	1.01	1.0
15 comp.	DROPP	0.006	0.006	0.009	**0.488**	**0.521**	**0.561**
	PCA	**0.004**	**0.004**	**0.004**	0.576	0.620	0.662
	ICA	0.725	0.159	0.142	1.127	1.286	1.415
	TICA	0.209	0.095	0.111	0.986	1.012	1.009
30 comp.	DROPP	-	-	-	**0.332**	**0.336**	**0.376**
	PCA	-	-	-	0.448	0.504	0.458
	ICA	-	-	-	1.204	1.467	1.502
	TICA	-	-	-	0.953	1.049	0.911

achieve a similar effect. Specifically, we superpose each time step of the trajectory according to the structure of the fully folded protein (as given by the PDB database [60]). Note, however, that DROPP achieves comparable similarities. Thus, our unsupervised method obtains the same performance as supervised PCA. In some sense, this is the ultimate standard for an unsupervised learning method. Adding knowledge about the full structure of the folded state of a protein was not possible until a few years ago when alpha fold [61] was developed. For novel complex molecules, this folded structure is still not available with a 100% guarantee.

IX. Summary of Experiments and Analysis

Table III gives a numerical overview of the reconstruction errors on real-world data sets. DROPP consistently achieves best (bold) or second-best (underlined) results, reaching comparable quality as PCA while returning similar components for different trajectories instead of independent projections. We summarize results that were consistent over all experiments:

1) High similarity: DROPP consistently delivers the most similar components for a wide range of manifestations over all experiments.
2) Low reconstruction error: reconstruction errors of DROPP when using components for other manifestations of the same process are comparable to PCA and significantly lower than for TICA or ICA.
3) Reliability: DROPP obtains consistently low variance.
4) The RE for different manifestations depends on how closely the covariances follow a Gaussian distribution (i.e., if the generating process is a random walk).

X. Conclusion

In this paper, we introduced DROPP, a DR method that effectively incorporates the order of data points by finding

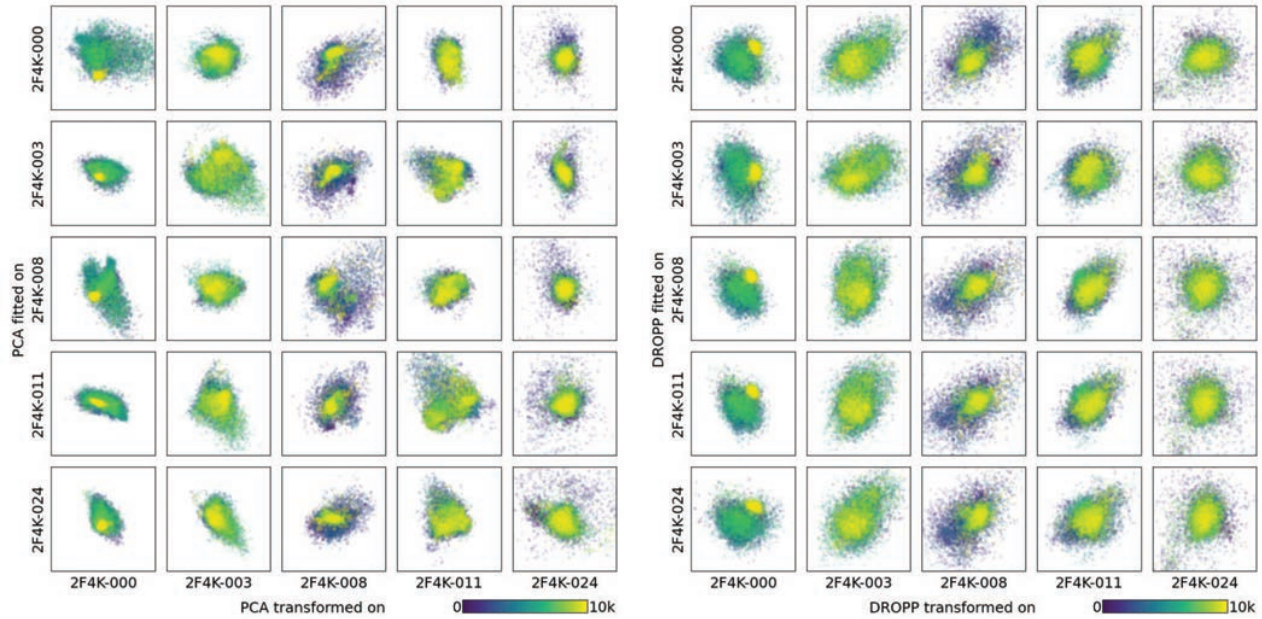


Fig. 11. Each row shows visualizations of five trajectories of the 2F4K protein with the traditional PCA (left) vs DROPP (right), superposed on a random frame. PCA computes each projection based on the individual trajectory, resulting in the visualizations on the diagonal. Using the same PCs for all trajectories leads to the visualizations shown in each row. A DR method that allows an equitable comparison and overview of the data should return similar projections independently of the fitted trajectory, i.e., each row should look similar.

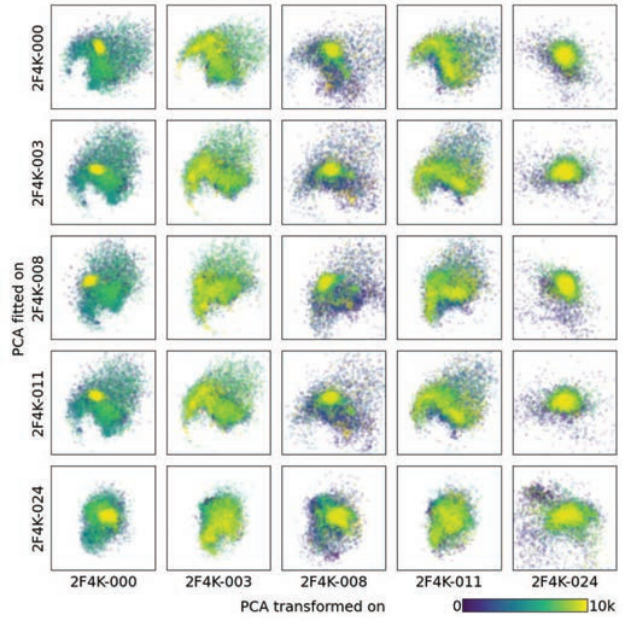


Fig. 12. Visualizations of 2F4K trajectories obtained by PCA after adding external knowledge by superposing the trajectories w.r.t. the structure of the folded state of the protein.

principal components that are common among different manifestations of a generating process. This allows for explainable visualizations of these manifestations as well as others that

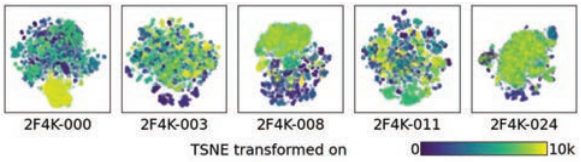


Fig. 13. Visualizations of 2F4K trajectories obtained by tSNE.

have not yet been seen. We extensively tested our method on synthetic as well as real-world data from climate research and molecular dynamics. Especially researchers in the field of molecular dynamics can profit from our methods, as it allows them to 1) compare trajectories in an explainable way, 2) apply the PCs as descriptors of proteins, which are usually hard to find. DROPP is suitable for early stages of data exploration, as it reliably returns similar relevant feature combinations for a range of related manifestations while maintaining reconstruction quality. In future work, we hope to extend DROPP by exploring more advanced kernels for the covariance matrix and to evaluate its performance in a streaming setting across other datasets. Additionally, we note that we merely learn the subspace of the data-generating process. If one wishes to expand on this, one could try to learn all the details of a data-generating process with, e.g., neural networks.

Acknowledgments

This project is partially funded by the Villum Foundation (project number 34326). We want to thank Ilias Patmanidis for his input about molecular dynamics.

REFERENCES

[1] G. H. Dunteman, *Principal components analysis*. Sage, 1989, vol. 69.

[2] N. Schneider, R. A. Sayle, and G. A. Landrum, "Get your atoms in order– an open-source implementation of a novel and robust molecular canonicalization algorithm," *Journal of chemical information and modeling*, vol. 55, no. 10, pp. 2111–2120, 2015.

[3] P. Desjardins and R. Morais, "Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates," *Journal of molecular biology*, vol. 212, no. 4, pp. 599–634, 1990.

[4] A. Marantan and L. Mahadevan, "Mechanics and statistics of the worm-like chain," *American Journal of Physics*, vol. 86, no. 2, pp. 86–94, 2018.

[5] A. R. Tejedor, J. R. Tejedor, and J. Ramírez, "Detailed dynamics of discrete gaussian semiflexible chains with arbitrary stiffness along the contour," *The Journal of Chemical Physics*, vol. 157, no. 16, 2022.

[6] T. Oldfield and R. Hubbard, "Analysis of cα geometry in protein structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 18, no. 4, pp. 324–337, 1994.

[7] A. D. Jenkins, P. Kratochvíl, R. F. T. Stepto, and U. W. Suter, "Glossary of basic terms in polymer science (iupac recommendations 1996)," *Pure and Applied Chemistry*, vol. 68, no. 12, pp. 2287–2311, 1996. [Online]. Available: https://doi.org/10.1351/pac199668122287

[8] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018.

[9] R. A. Jones, *Soft condensed matter*. Oxford University Press, 2002, vol. 6.

[10] A. R. Tejedor, J. R. Tejedor, and J. Ramírez, "Detailed dynamics of discrete gaussian semiflexible chains with arbitrary stiffness along the contour," *The Journal of Chemical Physics*, vol. 157, no. 16, 2022.

[11] A. Bender, J. L. Jenkins, J. Scheiber, S. C. K. Sukuru, M. Glick, and J. W. Davies, "How similar are similarity searching methods? a principal component analysis of molecular descriptor space," *Journal of chemical information and modeling*, vol. 49, no. 1, pp. 108–119, 2009.

[12] K. Hinsen, "Analysis of domain motions by approximate normal mode calculations," *Proteins: Structure, Function, and Bioinformatics*, vol. 33, no. 3, pp. 417–429, 1998.

[13] D. A. Case, "Normal mode analysis of protein dynamics," *Current Opinion in Structural Biology*, vol. 4, no. 2, pp. 285–290, 1994.

[14] Y. Naritomi and S. Fuchigami, "Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions," *The Journal of chemical physics*, vol. 134, no. 6, p. 02B617, 2011.

[15] V. Spiwok and P. Kříž, "Time-lagged t-distributed stochastic neighbor embedding (t-sne) of molecular simulation trajectories," *Frontiers in Molecular Biosciences*, vol. 7, p. 132, 2020.

[16] F. U. Hartl, "Protein misfolding diseases," *Annual review of biochemistry*, vol. 86, pp. 21–26, 2017.

[17] S. J. Teague, "Implications of protein flexibility for drug discovery," *Nature reviews Drug discovery*, vol. 2, no. 7, pp. 527–541, 2003.

[18] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[19] L. Molgedey and H. G. Schuster, "Separation of a mixture of independent signals using time delayed correlations," *Physical review letters*, vol. 72, no. 23, p. 3634, 1994.

[20] S. Schultze and H. Grubmüller, "Time-lagged independent component analysis of random walks and protein dynamics," *Journal of Chemical Theory and Computation*, vol. 17, no. 9, pp. 5766–5776, 2021.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[22] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[23] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery," *IEEE signal processing magazine*, vol. 35, no. 4, pp. 32–55, 2018.

[24] F. De la Torre and M. J. Black, "Robust principal component analysis for computer vision," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1. IEEE, 2001, pp. 362–369.

[25] M. Partridge and M. Jabri, "Robust principal component analysis," in *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No. 00TH8501)*, vol. 1. IEEE, 2000, pp. 289–298.

[26] G. Abraham and M. Inouye, "Fast principal component analysis of large-scale genome-wide data," *PloS one*, vol. 9, no. 4, p. e93766, 2014.

[27] A. Gang and W. U. Bajwa, "Fast-pca: A fast and exact algorithm for distributed principal component analysis," *IEEE Transactions on Signal Processing*, vol. 70, pp. 6080–6095, 2022.

[28] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[29] A. Hyvarinen, "Fast ica for noisy data using gaussian moments," in *1999 IEEE international symposium on circuits and systems (ISCAS)*, vol. 5. IEEE, 1999, pp. 57–61.

[30] G. R. Naik and D. K. Kumar, "An overview of independent component analysis and its applications," *Informatica*, vol. 35, no. 1, 2011.

[31] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[32] E. Schubert and M. Gertz, "Intrinsic t-stochastic neighbor embedding for visualization and outlier detection," in *Similarity Search and Applications*, C. Beecks, F. Borutta, P. Kröger, and T. Seidl, Eds. Cham: Springer International Publishing, 2017, pp. 188–203.

[33] M. Forni, M. Hallin, M. Lippi, and L. Reichlin, "The generalized dynamic-factor model: Identification and estimation," *Review of Economics and statistics*, vol. 82, no. 4, pp. 540–554, 2000.

[34] M. Hallin and M. Lippi, "Factor models in high-dimensional time series—a time-domain approach," *Stochastic processes and their applications*, vol. 123, no. 7, pp. 2678–2695, 2013.

[35] G. Goerg, "Forecastable component analysis," in *International conference on machine learning*. PMLR, 2013, pp. 64–72.

[36] D. S. Matteson and R. S. Tsay, "Dynamic orthogonal components for multivariate time series," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1450–1463, 2011.

[37] F. Xia, J. Liu, H. Nie, Y. Fu, L. Wan, and X. Kong, "Random walks: A review of algorithms and applications," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 2, pp. 95–107, 2019.

[38] J. Bye, K. Fraedrich, E. Kirk, S. Schubert, and X. Zhu, "Random walk lengths of about 30 years in global climate," *Geophysical research letters*, vol. 38, no. 5, 2011.

[39] A. Gordon, "Global warming as a manifestation of a random walk," *Journal of Climate*, vol. 4, no. 6, pp. 589–597, 1991.

[40] E. M. Hanks, "Modeling spatial covariance using the limiting distribution of spatio-temporal random walks," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 497–507, 2017.

[41] J. Antognini and J. Sohl-Dickstein, "Pca of high dimensional random walks with comparison to neural network training," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[42] M. Scheurer, P. Rodenkirch, M. Siggel, R. C. Bernardi, K. Schulten, E. Tajkhorshid, and T. Rudack, "Pycontact: rapid, customizable, and visual analysis of noncovalent interactions in md simulations," *Biophysical journal*, vol. 114, no. 3, pp. 577–583, 2018.

[43] A. Jurčík, K. Furmanová, J. Byška, V. Vonásek, O. Vávra, P. Ulbrich, H. Hauser, and B. Kozlíková, "Visual analysis of ligand trajectories in molecular dynamics," in *2019 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2019, pp. 212–221.

[44] H. Wu and F. Noé, "Variational approach for learning markov processes from time series data," *Journal of Nonlinear Science*, vol. 30, no. 1, pp. 23–66, 2020.

[45] J. Palma and G. Pierdominici-Sottile, "On the uses of pca to characterise molecular dynamics simulations of biological macromolecules: Basics and tips for an effective use," *ChemPhysChem*, vol. 24, no. 2, p. e202200491, 2023.

[46] S. Rajpoot, T. Ohishi, A. Kumar, Q. Pan, S. Banerjee, K. Y. Zhang, and M. S. Baig, "A novel therapeutic peptide blocks sars-cov-2 spike protein binding with host cell ace2 receptor," *Drugs in R&D*, vol. 21, pp. 273–283, 2021.

[47] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal component analysis for protein folding dynamics," *Journal of molecular biology*, vol. 385, no. 1, pp. 312–329, 2009.

[48] F. Sittel, A. Jain, and G. Stock, "Principal component analysis of molecular dynamics: On the use of cartesian vs. internal coordinates," *The Journal of Chemical Physics*, vol. 141, no. 1, 2014.

[49] O. F. Lange and H. Grubmüller, "Generalized correlation for biomolecular dynamics," *Proteins: Structure, Function, and Bioinformatics*, vol. 62, no. 4, pp. 1053–1061, 2006.

[50] H. Fataftah and W. Karain, "Detecting protein atom correlations using correlation of probability of recurrence," *Proteins: structure, function, and bioinformatics*, vol. 82, no. 9, pp. 2180–2189, 2014.
[51] T. Ichiye and M. Karplus, "Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations," *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 3, pp. 205–217, 1991.
[52] S. Pfenninger and I. Staffell, "Weather data," Open Power System Data, 2020. [Online]. Available: https://doi.org/10.25832/weather_data/2020-09-16
[53] R. T. McGibbon, "Fs MD Trajectories," 5 2014, 10.6084/m9.figshare.1030363.v1. [Online]. Available: https://figshare.com/articles/dataset/Fs_MD_Trajectories/1030363
[54] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw, "How fast-folding proteins fold," *Science*, vol. 334, no. 6055, pp. 517–520, 2011.
[55] G. R. Kneller, "Superposition of molecular structures using quaternions," *Molecular Simulation*, vol. 7, no. 1-2, pp. 113–119, 1991.
[56] F. Sittel, A. Jain, and G. Stock, "Principal component analysis of molecular dynamics: On the use of Cartesian vs. internal coordinates," *The Journal of Chemical Physics*, vol. 141, no. 1, p. 014111, Jul. 2014. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.4885338
[57] A. D. McLachlan, "A mathematical procedure for superimposing atomic coordinates of proteins," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 28, no. 6, pp. 656–657, 1972.
[58] N. I. of Standards and Technology, "Atomic weights of the elements 2013," *NIST*, 2013. [Online]. Available: https://physics.nist.gov/cgi-bin/Compositions/stand_alone.pl?ele=&ascii=html&isotype=some
[59] A. Draganov and S. Dohn, "Unexplainable explanations: Towards interpreting tsne and umap embeddings," *arXiv preprint arXiv:2306.11898*, 2023.
[60] J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, "Sub-microsecond protein folding," *Journal of molecular biology*, vol. 359, no. 3, pp. 546–553, 2006.
[61] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.