Reconstruction and Decomposition of High-Dimensional Landscapes via Unsupervised Learning

Jing Lei
Dept of Statistics
George Mason University

Nasrin Akhter
Dept of Computer Science
George Mason University

Wanli Qiao*
Dept of Statistics
George Mason University

Amarda Shehu†
Dept of Computer Science
George Mason University

ABSTRACT

Uncovering the organization of a landscape that encapsulates all states of a dynamic system is a central task in many domains, as it promises to reveal, in an unsupervised manner, a system's inner working. One domain where this task is crucial is in bioinformatics, where the energy landscape that organizes three-dimensional structures of a molecule by their energetics is a powerful construct. The landscape can be leveraged, among other things, to reveal macrostates where a molecule is biologically-active. This is a daunting task, as landscapes of complex actuated systems, such as molecules, are inherently high-dimensional. Nonetheless, our laboratories have made some progress via topological and statistical analysis of spatial data over the recent years. We have proposed what is essentially a dichotomy, methods that are more pertinent for visualization-driven discovery, and methods that are more pertinent for discovery of the biologically-active macrostates but not amenable to visualization. In this paper, we present a novel, hybrid method that combines strengths of these methods, allowing both visualization of the landscape and discovery of macrostates. We demonstrate what the method is capable of uncovering in comparison with existing methods over structure spaces sampled with conformational sampling algorithms. Though the direct evaluation in this paper is on protein energy landscapes, the proposed method is of broad interest in cross-cutting problems that necessitate characterization of fitness and optimization landscapes.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; **Computational biology**; **Molecular structural biology**;

KEYWORDS

high-dimensional landscape, landscape decomposition, topological analysis, statistical analysis, unsupervised learning, protein energy landscape, decoy selection

ACM Reference Format:
Jing Lei, Nasrin Akhter, Wanli Qiao, and Amarda Shehu. 2020. Reconstruction and Decomposition of High-Dimensional Landscapes via Unsupervised Learning. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining USB Stick (KDD '20), August 23–27, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394486.3403300

1 INTRODUCTION

In many domains, it is central to characterize and understand the behavior of a complex dynamic system. This is particularly important in bioinformatics. Specifically, in computational molecular biology, uncovering the organization of the structure space of an inherently plastic molecule has become a central component in structure-driven discoveries of biological activity [18]. This is particularly true in studies of proteins, which can employ significantly different structures with which to "stick to" and bind other partners in the cell [6]. In a statistical mechanics treatment, the structure space has an inherently hierarchical organization and can be described in terms of macrostates, which contain thermodynamically inter-converting structures that are also referred to as microstates [7].

Revealing this organization is important to obtain information on the mechanism(s) of action of a molecule and possibly reveal new states employed for molecular interactions [24]. This organization is inherent in the concept of the energy landscape [7], which organizes structures by proximity and energetics. As is often the case with complex, highly-actuated systems, such as molecules (and proteins), the landscapes are high-dimensional. The landscape essentially lifts the structure space with an additional dimension which evaluates structures based on their energies.

It is worth noting that the concept of the energy landscape, often also referred to as fitness landscape, originated in theoretical biology [36] and has become a useful construct in many domains and disciplines, from describing the physics of disordered systems, to molecular biology and bioinformatics, to more general search and optimization problems in AI, the study of complex systems, etc. In particular, the energy landscape has improved our understanding of the relationship between protein structure, structure plasticity, and function [7, 23].

[*] Corresponding Author: wqiao@gmu.edu
[†] Corresponding Author: amarda@gmu.edu

KDD '20, August 23–27, 2020, Virtual Event, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403300

The inherent organization of a molecular energy landscape exposes structural states as recognizable components. The landscape dimensionality and multimodality make automated detection of such components exceptionally challenging. Moreover, the protein energy landscape is a continuous surface. Yet, whether through wet- or dry-laboratory techniques, we only have access to a set of structure-energy pairs that are sampled from the landscape. A central question is how to go from this discrete representation of the landscape to an information-rich representation that reveals the organization of structures in the landscape and do so in an unsupervised manner.

Many studies that utilize the energy landscape define 2 features with which to summarize structures or obtain such features via compression models. The featurized structures are visualized and color-coded by energies to detect the possible emergence of thermodynamically-stable and semi-stable states that present themselves as local neighborhoods of low-energy structures. Interesting discoveries linking structure to function are made via visualization [9, 19].

Other studies employ barrier or disconnectivity trees or graphs [35] to visualize landscapes. These build over informal notions of landscape components and assume critical points are known *a priori.* What is needed is a tool to reconstruct the energy landscape and then decompose it into (formally-defined) components of interest with no prior knowledge on the location of critical points in the landscape. This task can be formulated under the umbrella of unsupervised learning.

Our laboratories have made some progress in this direction via topological and statistical analysis of spatial data. Over the past few years, we have proposed methods for a more reliable visualization and methods that forego the ability to visualize for a more accurate identification of stable and semi-stable structural states via the concept of basins. In this paper, we address this methodological dichotomy and present a novel, hybrid method that combines the strengths of visualization- and discovery-driven methods for reconstructing and decomposing energy landscapes.

We demonstrate that the proposed method provides both useful visualization of high-dimensional energy landscapes and identification of homogeneous structural states. The method achieves these dual tasks in an unsupervised manner and leverages concepts and ideas from topological and statistical analysis and unsupervised learning. Though the direct evaluation is on protein energy landscapes, due to the importance of landscapes in many problems and domains, the proposed method is of broad interest in problems that necessitate characterization of and learning over fitness and optimization landscapes that summarize high-dimensional states of complex systems.

The rest of this paper proceeds as follows. After summarizing related work in Section 1.1, we relate methodological details in Section 2. Evaluation is provided in Section 3. The paper concludes in Section 4.

1.1 Related Work and Background

We first formalize useful concepts before describing how they are leveraged in the proposed method.

1.2 The Energy Landscape

The energy landscape is an example of the more general *fitness* landscape (referred to as the height landscape or the lifted space in other disciplines). The fitness landscape consists of a set of points X, a neighborhood $\mathcal{N}(X)$ defined on X, a distance metric on X, and a fitness function $f : X \rightarrow \mathbb{R}_{\geq 0}$. The neighborhood $\mathcal{N}(X)$ assigns a neighborhood to every point in X. The fitness function f assigns a fitness to every point $x \in X$. In a structure space probed by a computational method, the points X correspond to computed structures, and the fitness function is an energy function scoring the structures.

A point of the landscape is a structure-and-energy pair. An energy landscape may contain many components, such as basins and basin-separating barriers. The concept of a basin is tied to a local/focal minimum. A focal minimum is surrounded by a basin of attraction, which is the set of points on the landscape from which steepest descent/ascent converges to that focal optimum. Barriers regulate transitions of a system between different structural states corresponding to basins in the landscape. Algorithms that compute structures of a molecule effectively sample points from an unknown, underlying landscape and so obtain a discrete, sample-based view of the landscape as a set or collection of points on the landscape.

Under the energy landscape view, one can in principle identify the biologically-active states by identifying the corresponding basins in the landscape. This presents several challenges, because protein energy landscapes are high-dimensional, overly rugged, and probed by (conformational sampling) algorithms that obtain a limited, biased view of it. Below we describe a concept in Morse theory called Morse-Smale complex, which provides a mathematical model and tool to decompose, reconstruct, and visualize the energy landscape.

1.3 The Morse-Smale Complex

Suppose we have a smooth function $f : \mathbb{M} \mapsto \mathbb{R}$, where we think of $\mathbb{M}$ as the (possibly featurized) structure space; f as an energy function that maps a point in M to an energy/fitness value. Given a starting point $x \in \mathbb{M}$, an integral curve $\eta_x : \mathbb{R} \mapsto \mathbb{M}$ is a curve satisfying $\frac{d\eta_x(t)}{dt} = \nabla f(\eta_x)$, $\eta_x(0) = x$. A particle following this integral moves towards a destination defined by $\text{dest}(x) = \lim_{t\to\infty} \eta_x(t)$. The path η_x can also be traced back to its origin defined by $\text{org}(x) = \lim_{t\to-\infty} \eta_x(t)$. Both $\text{dest}(x)$ and $\text{org}(x)$ are *critical points* of f. The paths of the integral lines and the critical points cover the entire domain $\mathbb{M}$, and we may find partitions of $\mathbb{M}$ by aggregating the integral lines depending their origins and destinations.

Specifically, we define the stable (or ascending) manifold of a critical point y as $A(y) = \{x : \text{dest}(x) = y\}$, and the unstable (or descending) manifold of a critical point y as $D(y) = \{x : \text{org}(x) = y\}$. We assume that f is a Morse function, meaning that its stable and unstable manifolds intersect transversally. The Morse-Smale complex is a set consisting of the intersections $A(y_i) \cap D(y_j)$ for all the critical points y_i and y_j, which forms a partition of $\mathbb{M}$. In this paper, we focus on the unstable manifolds of the energy function with local minima as the origins, because those sets are naturally related to the concepts of energy basins.

1.4 Persistence

The concept of persistence measures how a critical point is vertically distinct from its neighboring critical points. Given a critical point y of f, let $N(y)$ be the set of critical points z of f such that they are directly connected with y through the integral curves (meaning that there are no other critical points in the path between y and z). We define $\text{pers}(y) = \min_{z \in N(y)} |f(y) - f(z)|$ as the persistence of the critical point y. In particular, when y is a local minimum, $\text{pers}(y)$ is the minimum difference between the energy values at y and its directly connected saddle points, which reflects the minimum amount of energy required to move out of the basin $\text{org}(y)$ from its bottom.

Using persistence, the Morse-Smale complex of f can be simplified in the following way. Given a threshold t, if the difference of the energy values between a critical point y and one of its directly connected critical points is below this threshold, then these paired critical points can be merged or canceled, which results in a simplified partition of the domain. This simplification procedure is useful when the small persistence of critical points is believed to be insignificant. For example, noise in the data can increase the complexity of the Morse-Smale partition by bringing artificial critical points of small persistence; simplification tries to recover the true Morse-Smale complex.

1.5 Nearest-neighbor Graph

One can embed structures in a connectivity data structure and utilize energies to identify basins. Specifically, consider an Ω set of structures (these can be uncomplexed structures of a protein, ligand, or even complexed, protein-ligand or protein-protein structures obtained via conformational sampling). The ensemble Ω can be embedded in a nearest-neighbor graph (nngraph) $G = (V, E)$. The vertex set V is populated with the structures, and edge set E is populated by inferring the neighborhood structure of the landscape. The distance between two structures is measured via root-mean-squared-deviation (RMSD) [20] after each of the structures is superimposed over some reference structure (arbitrarily, chosen to be the first in the ensemble; the superimposition minimizes differences due to rigid-body motions.

Each vertex $u \in V$ is connected to vertices $v \in V$ if $d(u, v) \leq \epsilon$, where ϵ is a user-defined parameter. A small ϵ may result in a disconnected graph, which is the result of a sparse, non-uniform sampling of the landscape. This can be remedied by increasing ϵ while controlling the density of the resulting nngraph via the number of nearest neighbors of u. It is worth noting that the concept of the nngraph is pervasive in computer science. It allows one to capture the connectivity of a space and has been leveraged, among other applications, in constructing paths in the configuration space of a robot for robot motion planning [14] or in computing structural transitions in proteins [19, 21, 22, 30, 32].

2 METHOD

We describe a novel method that utilizes all the concepts laid out in Section 1.1. The method hybridizes two recent methods, BD(nngraph) and MS(PCs). BD(nngraph) focuses only on the task of basin detection (BD) over the nngraph embedding tertiary structures sampled for a protein via a conformational sampling algorithm. MS(PCs) addresses on the broader task of energy landscape reconstruction and decomposition by constructing the Morse-Smale (MS) complex over a two-dimensional embedding (via PCA) of the sampled tertiary structures. BD(nngraph) has been published in [2]. Though originally not named, the naming BD(nngraph) better conveys the core concepts leveraged. MS(PCs) has been published in [1], originally referred to as LRD for landscape reconstruction and decomposition. We now proceed to relate details, starting with BD(nngraph) and MS(PCs) over which MS(PCs+nngraph) builds over.

2.1 BD(nngraph): From a Graph Decoy Embedding to Basins

As described in Section 1.1, tertiary structures are first embedded in an nngraph $G = (V, E)$. BD(nngraph) groups $V(G)$ into distinct basins as follows. Tying the concept of a basin to its focal minimum, BD(nngraph) starts by first identifying local minima. A vertex $u \in V$ is a local minimum if $\forall v \in V\ f(u) \leq f(v)$, where $v \in N(u)$ ($N(u)$ denotes the neighborhood of u). The remaining vertices are assigned to basins as follows. Each vertex u is associated a negative gradient estimated by selecting the edge (u, v) that maximizes the ratio $[f(u) - f(v)]/d(u, v)$. From each vertex u that is not a local minimum, the negative gradient is followed (via the edge that maximizes the above ratio) until a local minimum is reached. Vertices that reach the same local minimum are assigned to the basin associated with that minimum. This approach of tying a basin to a local minimum and using traversals of the nngraph to assign vertices to basins was first proposed in [8]. BD(nngraph) can be considered a realization of unsupervised learning over a graph. The method extracts non-overlapping clusters/basins without actually reconstructing the underlying landscape, and so it does not support visualization of the landscape.

2.2 MS(PCs): From a 2D Embedding to Landscape Reconstruction and Decomposition

MS(PCs) exposes basins in the landscape by first *explicitly* reconstructing the landscape. Conceptually, the method "fills in" the landscape via kernel regression, which infers energies of points on the landscape based on energies of sampled points (computed structures) that are nearest neighbors. This filling is limited to points on a 2d grid, which allows explicitly reconstructing and decomposing low-dimensional embeddings of multi-dimensional and continuous landscapes. Principal Component Analysis (PCA) [17] is used to extract low-dimensional structural coordinates of the landscape. We note that PCA is shown effective for analysis of protein tertiary structures in various applications of interest in computational biology [9, 10, 18, 26]. MS(PCs) leverages an analysis of the cumulative variance to assess the effectiveness of dimensionality reduction in PCA and reduces the structures to two-dimensional points consisting of the first two principal components.

Given energy-evaluated (scored) two-dimensional points, MS(PCs) computes the alpha convex hull containing the points via the method described in [27]. On a 2d grid (over PC1 and PC2) within the hull, energies are estimated using kernel regression, which is a smoothing technique that calculates the weighted average of the

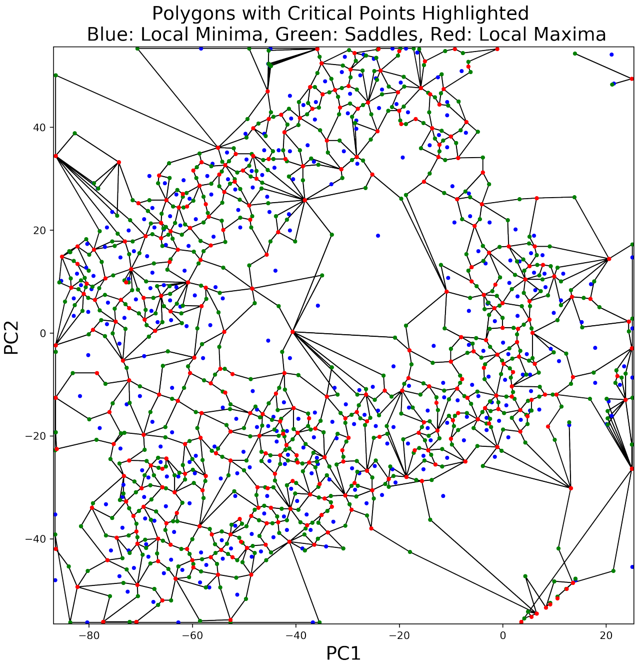


Figure 1: Tertiary structures are embedded via PCA, and the embedded landscape is reconstructed and decomposed into basins. Critical points are shown in different colors. The basin boundaries are drawn in black.

energies within a small neighborhood of a given point, where the weights are determined by the distance to the given point via a bounded-support Gaussian kernel. The MS complex described in Section 1.1 is then constructed over the grid points. In the case of $d = 2$, local minima are located in the bottom of the basins, and the basin boundaries are descending manifolds with saddle points as the origins, which include integral curves connecting local maxima and adjacent saddle points on the boundary (see Section 1.1).

Grid points with the same destinations in the path flows are grouped into the same basins. The grid points on the boundaries are connected using polygons for visualization purpose. The basins and their boundaries give hierarchical decomposition of the entire energy landscape. One such decomposition is shown in Fig. 1 for one of the test cases employed in this paper.

2.3 MS(nngraph): From a Multi-Dimensional Embedding to Landscape Reconstruction and Decomposition

MS(PCs) provides much richer information than BD(nngraph). It shows local minima, basins, saddles, and local maxima. However, this information is drawn from a 2d embedding of the structural data. Much detail is lost. It is possible, for instance, that what are depicted as separate basins may indeed overlap in higher dimensions. Extending this method to higher dimensions is not as straightfoward as increasing the dimensionality of the grid. Doing so brings a combinatorial explosion of the grid points and makes the method infeasible. Since the grid is the core connectivity structure that is the bottleneck of Morse-Smale complex-based approach, this bottleneck is what we address in MS(nngraph). The essential realization is that the nngraph in BD(nngraph) better captures the connectivity of the structure data, even though BD(nngraph) does not leverage it for more than detection of basins. Therefore, in MS(nngraph), we effectively construct a Morse-Smale complex over an nngraph. Let us describe this construction first before relating the rest of the details on the novel MS(nngraph) method that hybridizes BD(nngraph) and MS(PCs).

2.3.1 Morse-Smale complex over nngraph. The Morse-Smale complex constructed for a smooth function f can also be extended on a discrete set. We exploit the ideas laid out in [13] to do so. With a given data set $X = \{x_1, \cdots, x_n\} \subset \mathbb{R}^d$ and the associated energy values $Y = \{f(x_1), \cdots, f(x_n)\}$, one can first build an nngraph of X, denoted by $\mathit{nngraph}(X)$, and then use $\mathit{nngraph}(X)$ and Y to compute the Morse-Smale complex, which partitions the set X. The key is to estimate the gradient at each x_i and find the path following the gradient. For each x_i, define its adjacency as $\text{adj}(x_i) = \{x_j :\ x_i \in \mathit{nngraph}(x_j),\ x_j \in \text{nngraph}(x_i)\}$. Then from x_i, the steepest ascent move is $\arg\max_{x_j \in \text{adj}(x_i)}(f(x_j))$, and the steepest descent move is $\arg\min_{x_j \in \text{adj}(x_i)}(f(x_j))$. The trajectories tracking the steepest ascent and descent directions arrive at their destinations and origins, respectively. Each point in the set X is then clustered into one of the estimated Morse-Smale complex, depending on the pair of its destination and origin.

To retain the ability to visualize the Morse-Smale complex in lower dimensions, we construct the nngraph over the top K PCs obtained via preprocessing of the (structure) data set X with PCA. Note that the definition provided above makes no such demands. The main idea is that by utilizing the top K PCs, one can choose to visualize the constructed Morse-Smale complex over the top *PCs*, thus obtaining the benefit of visualization for interpretation of what the organization of the structure space may reveal. Yet, by choosing the K PCs that provide a desired cumulative variance, MS(nngraph) preserves a desired level of the structural variability, reconstructs the space spanned by the K principal components, and extracts the basins, saddles, local maxima, local minima (the entire Morse-Smale complex). The extracted K-dimensional basins can be projected onto the space spanned by the first two PCs for the purpose of visualization. It is worth noting that, since the original basins have dimensions higher than 2, the projected basins may have overlaps. We use the average energy (over structures assigned to a basin) to represent the overall energy of a basin, and plot the projected basins with the highest energy at the deepest bottom, as we are more interested in basins with low energy values.

2.3.2 Advantages of MS(nngraph) over other methods. Our new method MS(nngraph) provides a new flexible framework to handle the curse of dimensionality in the high-dimensional energy landscape data (and more generally in high-dimensional data) as well as to provide informative visualization of the decomposition of energy landscape at the same time. Compared with BD(nngraph), MS(nngraph) utilizes the low-dimension embeddings for which the user can directly choose the amount of the preserved variance. The dimension reduction technique significantly lowers the possibility of overfitting the Morse-Smale complex and the space complexity of the data used in the computation of the nearest graph. Similar to MS(PCs), the new method MS(nngraph) also uses PCs, but it is not based on kernel smoothing evaluated on grid points, which computationally restricts the MS(PCs) method from being applied to PCs of much more than 2 dimensions in practice. In addition, the visualization outcomes of the new method, MS(nngraph), are projections from the Morse-Smale complex of the K-dimensional PC space, which more closely reflects the structure of the original energy landscape. Only based on the first two PCs, the visualization using MS(PCs) runs a higher risk of oversimplifying the structure of energy landscapes, although the Morse-Smale complex decomposition using MS(PCs) is non-overlapping.

2.4 Implementation Details

BD(nngraph) and MS(PCs) are applied with default parameters as described in [1, 2]. Alignment of structures to remove rigid-body motions is implemented via BioPython [11]. PCA is carried out with the Python sklearn library. The construction of the nngraph over the K PCs is also implemented in Python. The R package msr is used to extract the Morse-Smale complex from the nngraph. In our experiments, we set $K = 10$ and $t = 3$, where t is the threshold for the persistent level. Different persistence values in the Morse-Smale complex are investigated, but they do not appreciably impact the results; results shown are obtained with a persistence value of 3. The construction of the nngraph in BD(nngraph) takes between 26 minutes and 2.5 hours depending on the size of the dataset (50 − 60K structures, 53 to 93 amino acids). This dominates the running time; the time to search for local minima and map vertices to these minima is insignificant. In MS(PCs), the running time is dominated by the computation of the alpha-convex hull and kernel smoothing of energy calculation on the grid, in addition to basin identification, resulting in a total of 6 to 21 minutes per dataset. The construction of the Morse-Scale complex in MS(nngraph) takes only about 5 minutes on each dataset. The additional visualization component in MS(nngraph), if desired, takes 40–60 minutes per dataset.

3 RESULTS

Results are presented on a dataset of 10 proteins of varying folds (α, β, $\alpha + \beta$, and *coil*) and lengths (53 to 93 number of amino acids). To generate tertiary structures, an amino-acid sequence is used as input to the Rosetta *ab-initio* protocol [15]. This is executed in an embarrasingly parallel manner to obtain an ensemble of $50,000 - 60,000$ structures per protein.

MS(PCs) and MS(nngraph) rely first on a PC-embedding of the probed structure space of a protein. Table 1 lists the cumulative variance captured by the top PCs in each protein/test case. In 5/10 of the test cases, the top two PCs capture 50% or more of the variance; this threshold is met by 7/10 of the test cases with three PCs. This suggests that visualizing reconstructed landscapes on two dimensions, PC1 and PC2, allows capturing a major portion of the structural diversity in the decoy ensemble data. More importantly, Table 1 shows that the top ten PCs capture over 70% of the cumulative variance on all the test cases, and over 80% on 6/10 of the test cases.

3.1 Visualization of Reconstructed and Decomposed Landscapes

We first provide in Fig. 2 a visual demonstration of the landscapes reconstructed by the proposed MS(nngraph) method on selected proteins. As described in Section 2, while the landscapes are reconstructed over the space spanned by the top ten PCs, the visualization utilizes an embedding over the top two PCs. Polygons, as described in Section 2, approximate basins for visualization. Color-coding is carried out via the Rosetta all-atom energy (score12). The blue-to-red color scheme indicates low-to-high energies.The energy associated with a basin is the average over decoys in the basin. Results that use the minimum energy instead are similar and not shown in the interest of space. The three largest basins are highlighted by drawing their boundaries in red.

Visualizing the landscape probed by a conformational sampling agorithm is informative. For instance, Fig. 2(a) shows that on LCI-CPA2 there are two distal regions in the landscape (and structure space) with deep basins. One region contains the larger basins (as annotated). Our quantitative evaluation below indicates that the larger basins contain structures similar to a known native structure and so possibly comprise the known native state. Could the low-energy basins in the other region be artifacts of the energy bias in Rosetta, or could these basins indicate alternative, long-lived states yet to be probed in the wet laboratory? Fig. 2(b) shows that on ArgR (DNA-binding D), an entire region of deep basins is probed by Rosetta, but (as the evaluation below shows), none of these are anywhere near the native state. On other proteins, such as L (B1), many regions (see Fig. 2(c)) are empty. These are not probed by Rosetta, indicating possible inherent bias in the algorithm or the energy function.

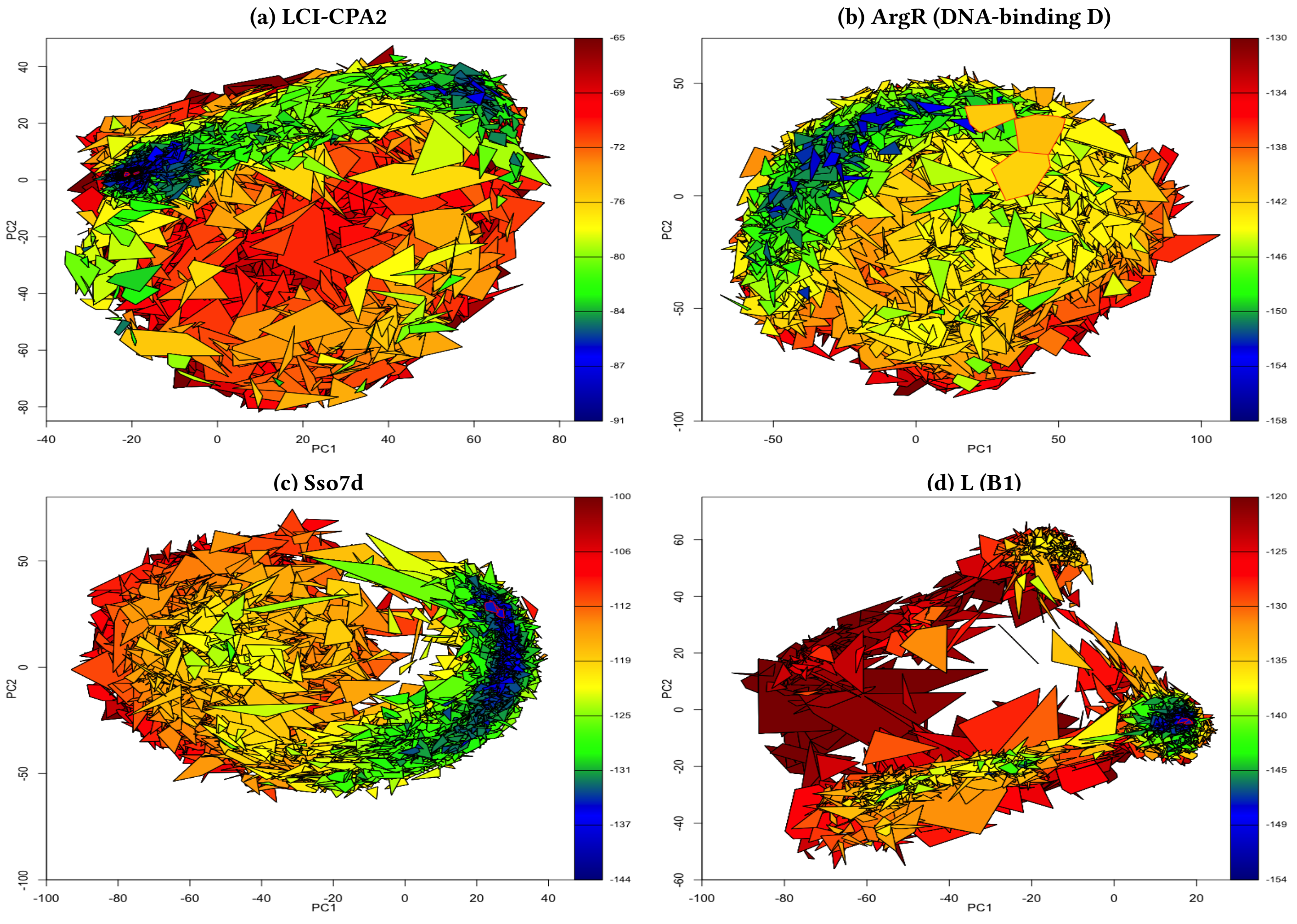
3.2 Relating Basins to Known Active States

We limit our analysis to known states we can find in the Protein Data Bank (PDB) [4]. For each of the proteins listed above, we can find one native structure, which we indicate by its 4-letter code (the PDB id), with the chain name in parentheses (for PDB entries that contain multiple chains). For instance, 1dtj(a) is the PDB id corresponding to a known native structure of the Nova-2 protein. We focus on a top basin that can be automatically selected via ranking on attributes, such as size (ordering basins from largest to smallest), size and energy (ranking first by size and then by energy, where the energy of a basin is measured as the average over the structures it contains), Pareto rank (PR; low to high) over the two criteria of size and energy, and by PR and Pareto count (PC; high to low) over these same two criteria (ordering basins with the same PR by PC). The last two rankings use Pareto-based metrics that employ the concept of dominance (related in the Appendix). A selected basin is then compared to a known native structure based on the fraction of structures in it that are within some proximity of the native structure. We refer to this metric as purity. Proximity is evaluated based on RMSD (over CA atoms), using a per-target threshold (we use default values as in BD(nngraph) [2]).

Fig. 3 relates the purity of the largest basin in (a) and the lowest-PR basin in (b) over basins detected by each of the three methods; the Appendix relates the top basin according to two more attributes. Fig. 3 (and the results related in the Appendix) shows that

Table 1: Testing dataset (protein names) are listed in Column 1. Domain names are in parentheses. Columns 2-5 list the cumulative variance captured by the top 2, 3, 5, and 10 PCs. Column 6 shows the number of PCs (over the total number of PCs) needed to reach 90% cumulative variance.

Name	Var_{1-2}	Var_{1-3}	Var_{1-5}	Var_{1-10}	$\#\text{PCs}_{\text{Var}=90\%}$ / #PCs
LCI-CPA2 [29]	33.4%	43%	55.6%	76.5%	18/183 (10%)
IF3 [5]	61.7%	67.6%	76.8%	87.3%	13/264 (5%)
Nova-2 [16]	59.7%	66.1%	75.9%	87.1%	13/222 (6%)
L (B1) [25]	62.4%	72.1%	80.5%	90.3%	10/192 (5%)
Sso7d [33]	53.2%	63.4%	75.7%	88%	12/192 (6%)
Rubredoxin [3]	36.5%	45%	58.4%	78%	19/159 (12%)
Sac7d [12]	49.9%	60.4%	74.9%	86.6%	13/198 (7%)
MuA (Ibetagamma D) [31]	39%	50.4%	65.2%	83.5%	15/279 (5%)
ArgR (DNA-binding D) [34]	42.9%	53.4%	68.2%	83.2%	15/234 (6%)
HiPIP [28]	24.9%	34.1%	49.3%	71.5%	21/186 (11%)


(a) LCI-CPA2

(b) ArgR (DNA-binding D)

(c) Sso7d

(d) L (B1)

Figure 2: MS(nngraph)-reconstructed and basin-decomposed landscapes are shown for selected proteins. The visualization is over PC1 and PC2. The contour lines show the basin boundaries. The largest three basins are highlighted with red boundaries.

BD(nngraph) yields basins of highest purity, followed by MS(nngraph) and MS(PCs), in this order. This is expected, as BD(nngraph) operates over the original dataset, whereas the other two methods operate over lower-dimensional embeddings. However, MS(nngraph) improves upon MS(PCs), as it considers more dimensions.

3.3 Quality of Detected Basins

Due to the concept of a local neighborhood, we tie the quality of a basin to the homogeneity of structures in it. A method that does not introduce deformations, unlike MS(PCs), should obtain structurally-homogeneous basins. Fig. 4(a) shows the width of the

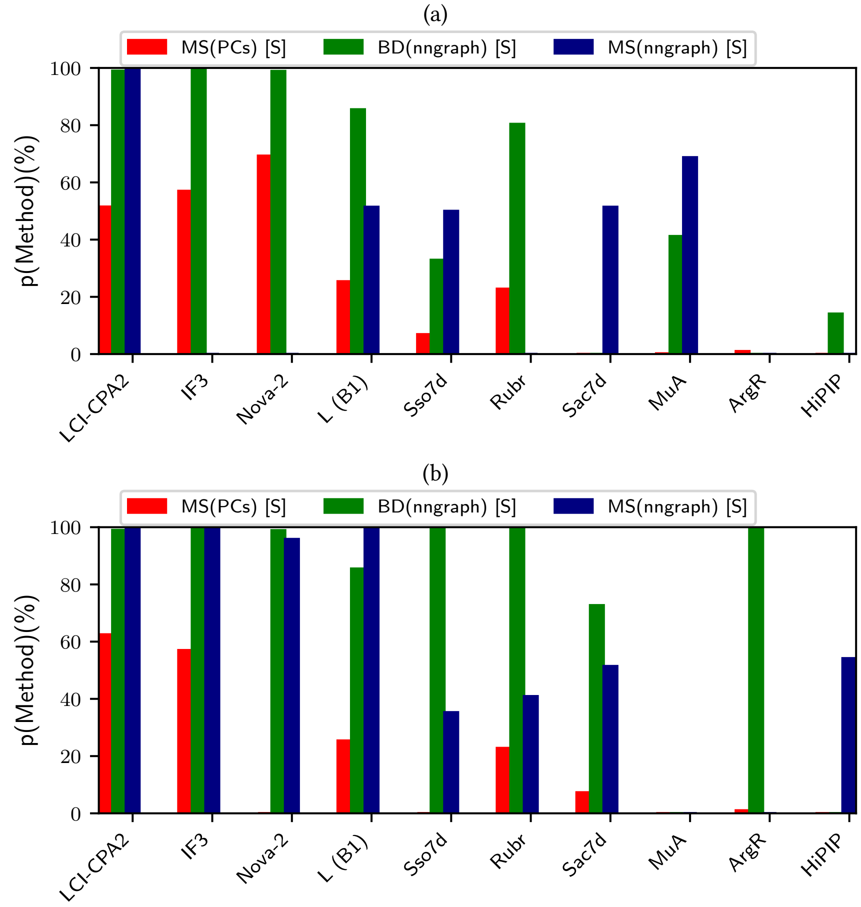


Figure 3: The basins detected by each of the three methods (color-coded in different colors) are ordered by different criteria; (a) size, (b) PR. The purity of the top basin is shown in each panel over the datasets. The names of some proteins are abbreviated in the interest of space.

largest basin detected by each method; width is measured as the average over the RMSDs between pairs of structures in a basin. Only the CA atoms of structures are used for the RMSD calculations. The "width" metric gives some insight into the homogeneity (conversely, the degeneracy) of a basin. Homogeneity is expected to suffer in MS(PCs) due to the assignment to a basin of structures based on a 2d embedding. Fig. 4(a) shows this to be the case. It shows that MS(nngraph) improves upon MS(PCs) due to the increased number of dimensions. As expected, by operating over the original dataset, BD(nngraph) has the more homoegeneous basin(s). On many of the datasets, however, the width of the largest basin obtained by BD(nngraph) is very similar to that obtained by MS(nngraph). These observations are reaffirmed in Fig. 4(b), which shows the width averaged over all basins detected by a method.

3.4 Runtime Comparison

Utilizing only top K PCs not only handles the curse of dimensionality while preserving informative structural information in the high-dimensional energy landscape (as shown in Section 3), it also saves us valuable CPU time. To elaborate on this topic, we show here a runtime comparison for nngraph construction (BD(nngraph)) using high-dimensional data and using only top 10 PCs. Note that, top 10 PCs retain 71.5% (*HiPIP*) to 90.3% (*L(B1)*) cumulative variance.

Table 2 compares the time required by BD(nngrph) on high-dimensional data with the time spent on 10 PCs for constructing nngraph. It is evident, as shown in columns 2 and 3, that the savings in time cost is significant. Runtime on high-dimensional data are at least more than 4 times (*Sso7d*) than runtime on 10 PCs. For instance, BD(nngrpah) requires more than 10 times (*LCI-CPA2*) CPU time than the time it requires for 10 PCs. The significant runtime savings using only 10-dimensional embedding and being able to retain important structural information show promise in uncovering structural space sampled with a conformational sampling algorithm by utilizing k-dimensional embedding.

4 CONCLUSION

The inspiration for the proposed MS(nngraph) method is the protein energy landscape. We anticipate the method may be of broad interest in cross-cutting problems that necessitate characterization

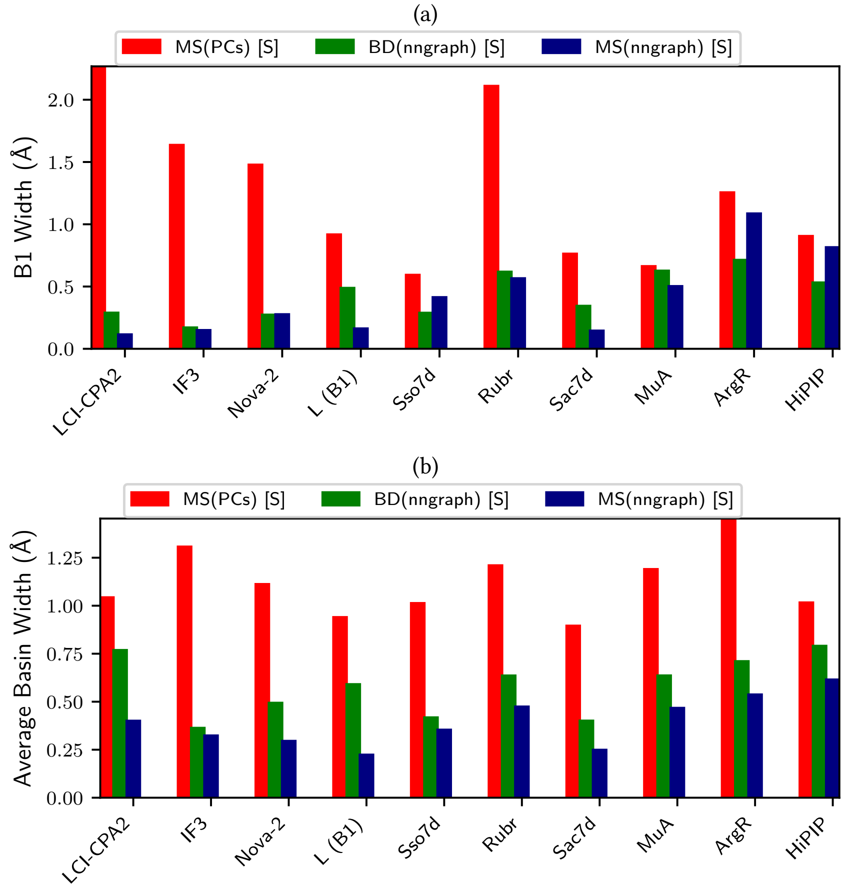


Figure 4: (a) The width of the largest basin (among basins detected by the different methods) is shown here, measured as the average RMSD over pairwise RMSDs between the structures in the basin. (b) shows the average width over all basins detected by a method. The names of some proteins are abbreviated in the interest of space.

Table 2: Runtime comparison. Protein names are listed in Column 1. Column 2 and 3 list runtimes incurred by BD(nngraph) on high-dimensional data and on 10 PCs, respectively. Runtimes are in minutes (m) and seconds (s).

Name	Runtime (High Dimension)	Runtime (10 PCs)
LCI-CPA2 [29]	81m 19s	8m 26s
IF3 [5]	80m 2s	7m 28s
Sso7d [33]	61m 10s	13m 24s
MuA (Ibetagamma D) [31]	74m 49s	14m 43s
ArgR (DNA-binding D) [34]	67m 32s	14m 59s

of fitness and optimization landscapes, so we make it available upon request.

In this paper, we choose to employ 10-dimensional embeddings for MS(nngraph) and, specifically, on embeddings obtained via PCA. Higher-dimensional embeddings may be desired in other applications to reach a desired cumulative variance in the context of PCA. The increase in dimensionality impacts the construction of the nngraph, as it affects the distance metric employed in the nearest-neighbor calculations. This impact becomes significant when approaching hundred and more dimensions. The embedding does not have to be linear or rely on PCA. The method can be applied to embeddings obtained via other compression models, including non-linear ones. We will investigate these directions in future work, motivated by discoveries on specific proteins of interest with possibly multiple stable and semi-stable states.

Finally, as some of the visualization-based analysis indicates, MS(nngraph) may be useful in comparing landscapes and structural states probed by different conformational sampling algorithms and different energy functions. In doing so, the method may assist researchers in highlighting inherent sampling biases and, more importantly, assist in the design of more powerful conformational sampling algorithms and more accurate energy functions for molecular modeling.

5 ACKNOWLEDGMENTS

The work is supported in part by the National Science Foundation Grant No. 1900061 and Grant No. 1821154. This material is additionally based upon work supported by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] N. Akhter, J. Lei, W. Qiao, and A. Shehu. 2018. Reconstructing and Decomposing Protein Energy Landscapes to Organize Structure Spaces and Reveal Biologically-active States. In *IEEE Intl Conf on Bioinf and Biomed (BIBM)*. IEEE, Madrid, Spain. accepted.

[2] N. Akhter and A. Shehu. 2018. From Extraction of Local Structures of Protein Energy Landscapes to Improved Decoy Selection in Template-free Protein Structure Prediction. *Molecules* 23, 1 (2018), 216.

[3] R. Bau, D. C. Rees, D. M. Kurtz, R. A. Scott, H. Huang, M. W. W. Adams, and M. K. Eidsness. 1998. Crystal Structure of Rubredoxin from Pyrococcus Furiosus at 0.95 Angstroms Resolution, and the structures of N-terminal methionine and formylmethionine variants of Pf Rd. Contributions of N-terminal interactions to thermostability. *J Biol Inorg Chem* 3 (1998), 484–493.

[4] H. M. Berman, K. Henrick, and H. Nakamura. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10, 12 (2003), 980–980.

[5] V. Biou, F. Shu, and V. Ramakrishnan. 1995. X-ray crystallography shows that translational initiation factor IF3 consists of two compact alpha/beta domains linked by an alpha-helix. *EMBO J* 14, 16 (1995), 4056–4064.

[6] D. D. Boehr, R. Nussinov, and P. E. Wright. 2009. The role of dynamic conformational ensembles in biomolecular recognition. *Nature Chem Biol* 5, 11 (2009), 789–796.

[7] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Genet.* 21, 3 (1995), 167–195.

[8] F. Cazals and T. Dreyfus. 2017. The structural bioinformatics library: modeling in biomolecular science and beyond. *Bioinformatics* 33, 7 (2017), 997–1004.

[9] R. Clausen, B. Ma, R. Nussinov, and A. Shehu. 2015. Mapping the Conformation Space of Wildtype and Mutant H-Ras with a Memetic, Cellular, and Multiscale Evolutionary Algorithm. *PLoS Comput Biol* 11, 9 (2015), e1004470.

[10] R. Clausen and A. Shehu. 2015. A Data-driven Evolutionary Algorithm for Mapping Multi-basin Protein Energy Landscapes. *J Comp Biol* 22, 9 (2015), 844–860.

[11] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, and more. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 10, 11 (2009), 1422–1423.

[12] S. P. Edmondson, L. Qiu, and J. W. Shriver. 1995. Solution structure of the DNA-binding protein Sac7d from the hyperthermophile Sulfolobus acidocaldarius. *Biochemistry* 34, 41 (1995), 13289–13304.

[13] S. Gerber, P.-T. Bremer, V. Pascucci, and R. Whitaker. 2010. Visual exploration of high dimensional scalar functions. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS* 16, 6 (2010), 1271–1280.

[14] L. E. Kavraki, P. Svetska, J.-C. Latombe, and M. Overmars. 1996. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans Robotics Automation* 12, 4 (1996), 566–580.

[15] A. Leaver-Fay et al. 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol* 487 (2011), 545–574.

[16] H. A. Lewis, H. Chen, C. Edo, R. J. Buckanovich, Y. Y. Yang, K. Musunuru, R. Zhong, R. B. Darnell, and S. K. Burley. 1999. Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains. *Structure* 7, 2 (1999), 191–203.

[17] D. G. Luenberger. 1984. *Linear and Nonlinear Programming* (2nd ed.). Addison-Wesley.

[18] T. Maximova, R. Moffatt, B. Ma, R. Nussinov, and A. Shehu. 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comp. Biol.* 12, 4 (2016), e1004619.

[19] T. Maximova, E. Plaku, and A. Shehu. 2018. Structure-guided Protein Transition Modeling with a Probabilistic Roadmap Algorithm. *IEEE/ACM Trans Comput Biol and Bioinf* 15, 6 (2018), 1783–1796.

[20] A. D. McLachlan. 1972. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr. A.* 26, 6 (1972), 656–657.

[21] K. Molloy, R. Clausen, and A. Shehu. 2016. A Stochastic Roadmap Method to Model Protein Structural Transitions. *Robotica* 34, 8 (2016), 1705–1733.

[22] K. Molloy and A. Shehu. 2016. A General, Adaptive, Roadmap-based Algorithm for Protein Motion Computation. *IEEE Trans. NanoBioSci.* 2, 15 (2016), 158–165.

[23] R. Nussinov and P. G. Wolynes. 2014. A second molecular biology revolution? The energy landscapes of biomolecular function. *Phys Chem Chem Phys* 16, 14 (2014), 6321–6322.

[24] K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, and P. G. Wolynes. 2006. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. USA* 103, 32 (2006), 11844–11849.

[25] J. W. O'Neill, D. E. Kim, D. Baker, and K. Y. Zhang. 2001. Structures of the B1 domain of protein L from Peptostreptococcus magnus with a tyrosine to tryptophan substitution. *Acta Crystallogr D Biol Crystallogr* 57, Pt 4 (2001), 480–487.

[26] R. Pandit and A. Shehu. 2016. A Principled Comparative Analysis of Dimensionality Reduction Techniques on Protein Structure Decoy Data. In *Intl Conf on Bioinf and Comput Biol* (Las Vegas, NV), T. Ioerger and N. Haspel (Eds.). ISCA, 43–48.

[27] B. Pateiro-Lopez. 2008. *Set estimation under convexity type restrictions*. Ph.D. Dissertation. Universidad de Santiago de Compostela.

[28] G. Rayment, I.AND Wesenberg, T. E. Meyer, M. A. Cusanovich, and H. M. Holden. 1992. Three-dimensional structure of the high-potential iron-sulfur protein isolated from the purple phototrophic bacterium Rhodocyclus tenuis determined and refined at 1.5 resolution. *J Mol Biol* 228, 2 (1992), 672–686.

[29] D. Reverter, C. Fernández-Catalán, R. Baumgartner, R. Pfänder, R. Huber, W. Bode, J. Vendrell, T. A. Holak, and F. X. Avilés. 2000. Structure of a novel leech carboxypeptidase inhibitor determined free in solution and in complex with human carboxypeptidase A2. *Nat Struct Biol* 7, 4 (2000), 322–328.

[30] E. Sapin, D. B. Carr, K. A. De Jong, and A. Shehu. 2016. Computing energy landscape maps and structural excursions of proteins. *BMC Genomics* 17, Suppl 4 (2016), 456.

[31] S. Schumacher, R. T. Clubb, M. Cai, K. Mizuuchi, G. M. Clore, and A. M. Gronenborn. 1997. Solution structure of the Mu end DNA-binding ibeta subdomain of phage Mu transposase: modular DNA recognition by two tethered domains. *EMBO J* 16, 24 (1997), 7532–7541.

[32] A. Shehu and E. Plaku. 2016. A Survey of omputational Treatments of Biomolecules by Robotics-inspired Methods Modeling Equilibrium Structure and Dynamics. *J Artif Intel Res* 597 (2016), 509–572.

[33] S. Su, Y. G. Gao, H. Robinson, Y. C. Liaw, S. P. Edmondson, J. W. Shriver, and A. H. Wang. 2011. Crystal structures of the chromosomal proteins Sso7d/Sac7d bound to DNA containing T-G mismatched base-pairs. *J Mol Biol* 303, 3 (2011), 395–403.

[34] M. Sunnerhagen, M. Nilges, G. Otting, and J. Carey. 1997. Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA. *Nat Struct Biol* 4, 10 (1997), 819–826.

[35] D. J. Wales, M. A. Miller, and T. R. Walsh. 1998. Archetypal energy landscapes. *Nature* 394, 6695 (1998), 758–760.

[36] S. Wright. 1934. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. In *Intl Congress of Genetics*. 356–366.